Safety-Critical Scenarios for Autonomous Driving: A Survey of Methods, Benchmarks, and Verification Pipelines

ZIYU SONG, National Key Laboratory of Automotive Chassis Integration and Bionics, Jilin University, China
YUNFENG HU, College of Communication Engineering, Jilin University, China
YIQIN DENG, School of Data Science, Lingnan University, Hong Kong, China
ZHENG LIN, Department of Electrical and Computer Engineering, University of Hong Kong, China
JING YANG, Center of Research for Cyber Security and Network (CSNET), Faculty of Computer Science and Information Technology, Universiti Malaya, Malaysia
SUNIL PRAJAPAT, Department of Computer Engineering-AI Big Data, Marwadi University, India
ZIHAN FANG, Department of Computer Science, City University of Hong Kong, China
YANG ZHANG*, School of Mechanical Engineering, Southeast University, China
LIP YEE POR, Center of Research for Cyber Security and Network (CSNET), Faculty of Computer Science and Information Technology, Universiti Malaya, Malaysia
HAITAO DING, National Key Laboratory of Automotive Chassis Integration and Bionics, Jilin University, China
WEI NI, School of Engineering, Edith Cowan University, Australia
JUN LUO, College of Computing and Data Science, Nanyang Technological University, Singapore

The safe deployment of autonomous vehicles depends on their ability to perceive, handle, and validate safety-critical scenarios, namely rare but complex situations that are highly relevant to real world safety risks. Different from previous surveys on safety-critical scenarios in autonomous driving, this survey systematically reviews safety-critical scenarios from three perspectives: methods, benchmarks, and verification pipelines. We first introduce a five-domain taxonomy that organizes existing methods from geometry-based safety estimation and deep risk anticipation to risk grounding, uncertainty-aware control, and integrated control with world models and vision-language models. We then review traffic accident datasets and critical scenario benchmarks. Based on these datasets and

*Corresponding author.

Authors' Contact Information: Ziyu Song, National Key Laboratory of Automotive Chassis Integration and Bionics, Jilin University, Changchun, China, songziyu@jlu.edu.cn; Yunfeng Hu, College of Communication Engineering, Jilin University, Changchun, China, huyf@jlu.edu.cn; Yiqin Deng, School of Data Science, Lingnan University, Hong Kong, Hong Kong, China, yiqindeng@ln.edu.hk; Zheng Lin, Department of Electrical and Computer Engineering, University of Hong Kong, Hong Kong, China, linzheng@eee.hku.hk; Jing Yang, Center of Research for Cyber Security and Network (CSNET), Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur, Malaysia, s2147529@siswa.um.edu.my; Sunil Prajapat, Department of Computer Engineering-AI Big Data, Marwadi University, India, sunilprajapat645@gmail.com; Zihan Fang, Department of Computer Science, City University of Hong Kong, Hong Kong, China, zihanfang3-c@my.cityu.edu.hk; Yang Zhang, School of Mechanical Engineering, Southeast University, Nanjing, China, 15705608994@163.com; Lip Yee Por, Center of Research for Cyber Security and Network (CSNET), Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur, Malaysia, porlip@um.edu.my; Haitao Ding, National Key Laboratory of Automotive Chassis Integration and Bionics, Jilin University, Changchun, China, dinght@jlu.edu.cn; Wei Ni, School of Engineering, Edith Cowan University, Perth, Australia, wei.ni@ieee.org; Jun Luo, College of Computing and Data Science, Nanyang Technological University, Singapore, Singapore, junluo@ntu.edu.sg.

Manuscript submitted to ACM

benchmarks, we further derive thirty representative critical scenarios for evaluation and benchmark design. Finally, we summarize verification pipelines and outline future directions. The review shows that current studies are moving toward closed-loop safety assurance, but standardized benchmarks, uncertainty-aware control, and traceable validation evidence remain insufficient.

Additional Key Words and Phrases: Autonomous driving, safety-critical scenarios, vision-language models, world models

ACM Reference Format:
Ziyu Song, Yunfeng Hu, Yiqin Deng, Zheng Lin, Jing Yang, Sunil Prajapat, Zihan Fang, Yang Zhang, Lip Yee Por, Haitao Ding, Wei Ni, and Jun Luo. 2018. Safety-Critical Scenarios for Autonomous Driving: A Survey of Methods, Benchmarks, and Verification Pipelines. *ACM Comput. Surv.* 1, 1 (May 2018), 35 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

As technological and industrial advancements continue to accelerate, autonomous driving systems (ADS) have emerged as an important application area for integrating advances in artificial intelligence [1–8], the Internet of Things [9–17], cloud computing, information and communication technologies [18–22], and big data [23–29]. The development and deployment of ADS are expected to contribute to changes in the automotive industry, enhance traffic safety and efficiency, reduce energy consumption and emissions, and reshape individual mobility patterns.

Despite rapid progress from advanced driver-assistance systems [30–33] to higher-level autonomous driving systems (ADS) [34–36], and from standalone to connected vehicles, current technology remains far from achieving reliable unmanned driving. Public reports [37–39] indicate that vehicles equipped with advanced automated functions continue to be involved in high-profile crashes, raising concerns about their real world safety. For instance, Google's Waymo, often regarded as a frontrunner in autonomous driving, still requires approximately 0.18 manual interventions per 1,000 miles during testing; without these disengagements, incident statistics suggest that an accident would occur on average every 130,487 km [40]. These observations highlight that safety, rather than perception accuracy or comfort alone, has become the bottleneck for large-scale deployment.

An analysis of representative accident cases involving ADS in Table 1 reveals several key characteristics. First, the risk factors faced by ADS are diverse, encompassing adverse weather conditions [41], vulnerable road users [42], infrastructure defects [43], rare behaviors of surrounding traffic participants [44], and multi-source coupled conditions. Second, many failure modes differ fundamentally from conventional human errors or single-component faults and cannot be anticipated solely through traditional traffic experience. Third, such failures typically reside in the long-tail of the intended operation, including rare cases within the operational design domain (ODD) [45], near its boundary, or during ODD exit and violation conditions. They emerge only after hundreds of millions or even billions of kilometers of accumulated driving, making it impractical to rely on brute-force road testing to expose all relevant risks. In other words, ADS risk conditions are often complex, stochastic, rare, severe, and difficult to reproduce. These properties define critical scenarios and make them challenging for both methods design and safety validation.

Against this backdrop, safety assurance for ADS can no longer be framed as generic performance improvement, but must explicitly target the modeling, prevention, and testing of critical scenarios. Existing surveys have reviewed scenario-based safety assessment, critical scenario identification, and safety-critical scenario generation [53–55], or examined AI for safety-critical systems, deep learning for autonomous driving, and world models from broader methodological perspectives [56–60]. Most of them focus on either identifying or generating critical scenarios, or assessing ADS safety through scenario-based testing. The safety assurance process that connects open-loop risk perception, structured risk grounding, closed-loop control, benchmark evaluation, scenario construction, and verification and validation evidence

Table 1. An overview of typical autonomous driving accidents in recent years.

Date	Accident Description	Cause
Jun. 1st 2020	A Tesla Model 3, equipped with the Autopilot system, crashed directly into a rolled-over white truck on a highway in Taiwan, China [46].	The Autopilot failed to recognize the truck due to the daylight and the truck's white body.
Mar. 11th 2021	A Tesla Model Y, equipped with the Autopilot system, collided with the cargo container of a white truck in the suburbs of Detroit, USA [47].	The Autopilot system mistakenly classified the truck's cargo container as the sky, a misidentification attributed to its white hue.
May 5th 2021	A Tesla Model 3 crashed into an overturned truck in California, resulting in the death of its owner [48].	The Autopilot failed to recognize the truck due to high luminance during the daylight and the truck's white exterior.
Feb. 6th 2024	A Waymo robotaxi in San Francisco collided with a cyclist at a four-way stop after starting to move from a stop [49].	The ADS underestimated risk from a partially occluded cyclist, revealing limits in vulnerable road user (VRU) prediction and yielding.
Mar. 29th 2025	A Xiaomi SU7 with highway navigation on autopilot (NOA) crashed into work-zone barriers on the De-Shang Expressway, causing the death of its owner [50].	The NOA showed limited ability in recognizing and handling complex work-zone geometry and relied on rapid human takeover.
Apr. 8th 2025	A Zoox robotaxi in Las Vegas at >40 mph collided with a car slowly protruding from a side access lane and stopping [51].	The ADS made an over-confident wrong prediction of the side vehicle's path, leading to an unsafe evasive maneuver.
Aug. 6th 2025	A Baidu Apollo Go robotaxi in Chongqing drove through temporary barriers into a construction pit, injuring passengers [52].	Inconsistency between high-definition (HD) map, perception and the actual work-zone closure, plus weak ODD management for such scenes.

is still less systematically summarized. By contrast, this survey adopts a closed-loop perspective to examine ADS safety in critical scenarios. Specifically, the survey (i) reviews prevention and control methods that handle critical scenarios in a closed-loop manner, (ii) summarizes traffic accident datasets and critical scenario benchmarks that support the modeling and evaluation of such situations, and (iii) discusses scenario-based verification and validation techniques that integrate simulation, scenario generation, X-in-the-Loop (XiL), and real-vehicle experiments into a safety assurance framework. Table 2 positions the survey relative to representative prior surveys.

The main contributions of this paper are summarized as follows:

(1) We introduce a taxonomy that categorizes critical scenario methods into five domains. This taxonomy shows that research on safety-critical scenarios is shifting from open-loop risk perception toward closed-loop risk handling. Building on this taxonomy, we analyze the strengths and limitations of existing methods in terms of risk modeling fidelity, interpretability, and their integration with downstream planning and testing pipelines.
(2) We integrate heterogeneous traffic accident datasets and derive a catalog of thirty representative critical scenarios for controller evaluation. This catalog connects dataset annotations, scenario semantics, and closed-loop benchmark requirements, helping bridge the gap between data collection and executable testing of safety-critical scenarios. The analysis shows that current datasets cover common highway and urban conflicts relatively well, but still underrepresent rural hazards and compound environmental conditions.
(3) We summarize a scenario-based verification and validation architecture that connects scenario construction, simulation, XiL testing, virtual-real fusion, and real-vehicle evaluation. This survey highlights the need to transform generated critical scenarios into traceable validation processes, so that scenario design, test execution, and safety evidence can better support ADS deployment.

Table 2. Comparison with existing surveys on safety-critical scenarios and safety assurance for autonomous driving.

Survey	Main focus	Critical scenarios	Closed-Loop control	Benchmarks	XiL	Safety evidence
Riedmaier et al. [53]	Scenario-based safety assessment	Partial	×	Partial	Partial	✓
Zhang et al. [54]	Critical scenario identification	✓	×	Partial	×	Partial
Ding et al. [55]	Safety-critical scenario generation	✓	×	Partial	Partial	Partial
Gao et al. [61]	Foundation models for scenario generation	Partial	Partial	Partial	×	×
Perez-Cerrolaza et al. [56]	AI for safety-critical systems	Partial	Partial	×	×	✓
Zhao et al. [57]	Deep learning for ADS	Partial	Partial	Partial	×	×
Ding et al. [58]	World models and future prediction	Partial	Partial	Partial	×	Partial
This survey	Critical scenario centered ADS safety assurance	✓	✓	✓	✓	✓

A summary of this survey's structure is presented in Fig. 1. The remainder of the survey is organized as follows. Section 2 presents a closed-loop taxonomy of methods for preventing and handling critical scenarios and summarizes representative approaches and related benchmarks. Section 3 reviews scenario construction techniques, verification and validation methods for ADS. Section 4 discusses open challenges and future research directions. Section 5 concludes the paper.

2 Risk Perception, Closed-Loop Control Strategies, and Benchmarking Methods

2.1 Critical Scenario Definition

While the general concept of a scenario is well established in standards such as ISO 21448 [62] and foundational studies [63, 64], this survey focuses on critical scenarios. A scenario describes the temporal evolution of road layout, dynamic actors, and environmental conditions. In contrast to nominal driving, a critical scenario contains risk factors that increase the likelihood or severity of hazards [65, 66]. From a system verification perspective, recent studies further define such scenarios as specific parameter combinations that can induce unsafe states [67], or as dangerous near misses without collision that challenge the control boundary of an ADS [54].

Synthesizing these perspectives, we define a critical scenario as a traffic situation that is relevant to the intended operation or ODD management of an autonomous driving system and, due to adverse combinations of actors, environment, infrastructure, and system state, exhibits an elevated risk of collision, loss of control, or unsafe fallback. Such scenarios may occur within the declared ODD, near its boundary, or during ODD exit and violation conditions.

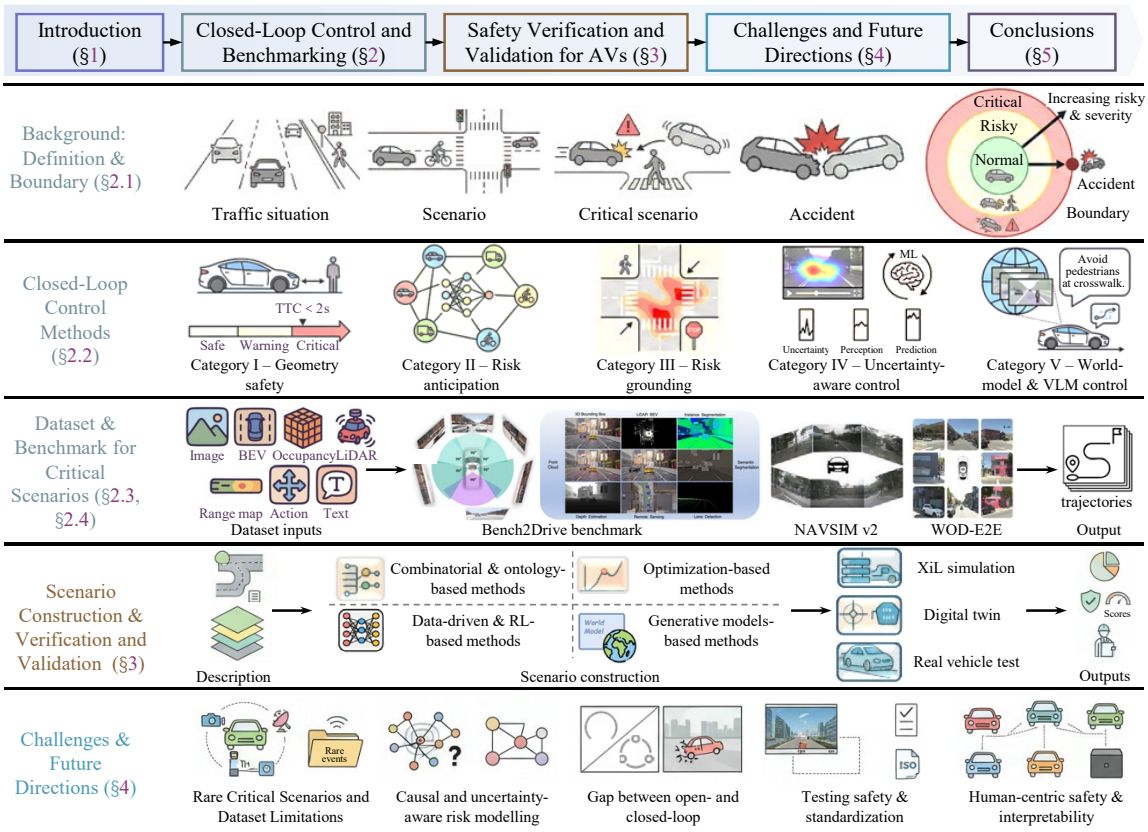


Fig. 1. Overview of the survey framework for safety-critical scenarios. The figure provides a section-aligned map of the survey, where the top axis shows the paper organization and the lower rows summarize five levels of content: scenario definition and risk boundary, closed-loop control methods, datasets and benchmarks, scenario construction and validation pipelines, and open challenges. Together, these levels show how safety-critical scenarios are defined, modeled, evaluated, instantiated, and validated in autonomous driving safety assurance.

This definition covers both near-miss and pre-crash situations, as well as ODD-boundary cases that challenge perception, decision-making, planning, control, fallback, and safety assurance. Subsequent sections focus on the modeling, prevention, benchmarking, and verification and validation of such critical scenarios.

2.2 Methods for Preventing Critical Scenarios

From a methodological perspective, existing work on critical scenario prevention and control in autonomous driving can be grouped into five categories. Category I comprises geometry-based safety estimation from trajectory prediction. These methods follow the classical "detection–tracking–trajectory prediction–risk evaluation" pipeline, in which autoregressive integrated moving average (ARIMA) [68–70], support vector machine (SVM) [71, 72], hidden Markov model (HMM) [73–75] or kinematic model–based [76–78] predictors extrapolate future motions and then compute surrogate safety indices such as minimum inter-vehicle distance, time-to-collision (TTC), post-encroachment time, or trajectory variance. Violations of analytically defined safety envelopes are mapped to collision risk and used to

trigger automatic emergency braking (AEB) or lane keeping interventions. Such approaches are computationally lightweight and provide clear physical interpretability, but they typically assume simple interaction models and struggle in interactive urban scenes with complex multi-agent negotiations or occlusions.

Category II covers deep accident anticipation models that predict scalar risk signals from short video clips, typically including frame-wise accident probabilities and time-to-accident (TTA) curves. A common instantiation is relation-aware modeling via spatiotemporal interaction graphs, where objects are treated as nodes and relative geometry as edges, and multi-frame evidence is aggregated by graph convolutional network (GCN) [79, 80] or graph neural network (GNN) [81, 82] modules coupled with recurrent neural network (RNN) [83], gated recurrent unit (GRU) [84] or Transformer backbones [85–87]. In parallel, many works adopt object-centric temporal encoders without an explicit graph, using detected objects as tokens and learning a compact latent state whose evolution supports risk and TTA prediction. Early-warning losses further encourage earlier anticipation by emphasizing frames closer to the accident. As illustrated in Fig. 2(a), chain-of-thought (CoT)-guided multimodal accident anticipation [88] instantiates this category by fusing object detections, motion cues, language-style scene descriptions, and driver-attention maps into a unified temporal predictor that outputs frame-wise accident probabilities and TTA curves, enriching purely geometric or relational cues with higher-level semantics. AccNet [89] extends object interaction modeling into three dimensions by combining monocular depth, 3D collision graphs, and the BA-LEA loss, achieving larger mean TTA in urban scenes while preserving a physically meaningful notion of risk. EQ-TAA [90] augments scalar risk learning with generative, causality-aware self-supervision: attentive video diffusion synthesizes pseudo-accident/pseudo-normal clips, and an equivariant triple loss enforces representation consistency across real and counterfactual samples, improving robustness under limited labels. W3AL [91] can also be viewed as a lightweight instance in this category, where MobileNet-style encoders, temporal attention, and GRUs jointly model scene-level and object-level features to produce frame-wise accident probabilities that feed subsequent modules. Compared with purely geometric methods, Category II models better capture nonlinear interactions and semantic cues, but their scalar outputs provide limited guidance on which actor, region, or interaction should be constrained during planning.

Category III targets structured risk grounding for control, extending scalar risk estimation to answer not only "how risky" but also "where" and "who" is risky. Instead of outputting only a frame-level probability or TTA curve, Category III methods predict spatially grounded cues such as risk heatmaps, high-risk agent sets, localized accident participants, or pixel-/box-level hazard regions, often augmented with textual rationales. For closed-loop planning, these grounded outputs are more actionable than scalar scores because they can be converted into agent-specific yielding rules, spatial no-go regions, or shielding constraints around localized hazards. Fig. 2(b) highlights the type of supervision and evaluation that Category III relies on. In particular, the intent-aware annotation schema of ImagiDrive [92] assigns each key agent a bounding box, relative position, and motion intent, together with natural language descriptions and binary "be cautious" recommendations, providing exactly the spatial and semantic control cues that Category III models aim to predict. Bridge methods such as W3AL [91] illustrate this transition by complementing accident likelihood and TTA prediction with accident-involved object localization. SafePLUG [93] pushes grounding granularity further by introducing region-level and pixel-level visual question answering for accident videos with dense segmentation masks and multimodal queries about hazardous agents and zones. Although SafePLUG is primarily a benchmark suite for fine grained accident understanding rather than a closed-loop controller, its pixel-accurate hazard regions and textual rationales can serve as high resolution control primitives, enabling planners to reason about "who is dangerous, where they are, and why" instead of acting on a single scalar score. Compared with Categories I–II, Category III offers

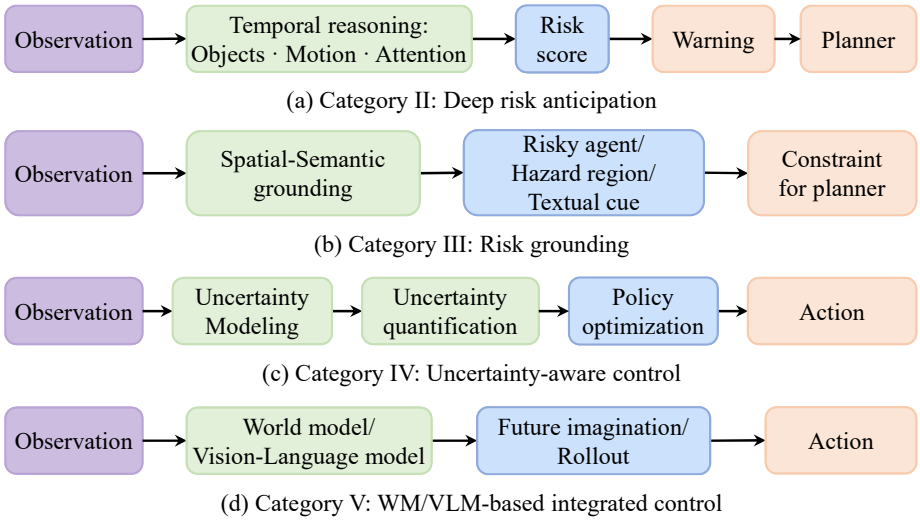


Fig. 2. Overview of representative methods for preventing critical scenarios. (a) **Category II (Deep Risk Anticipation):** temporal reasoning over objects, motion, and attention cues is used to estimate risk scores and provide warnings or planning triggers. (b) **Category III (Risk Grounding):** spatial-semantic grounding localizes risky agents, hazard regions, and textual cues, which can be converted into constraints for downstream planning. (c) **Category IV (Uncertainty-aware Control):** uncertainty modeling and quantification are incorporated into policy optimization to support robust action selection under uncertain observations. (d) **Category V (WM/VLM-based Integrated Control):** WMs and VLMs support future imagination or rollout, enabling action generation.

improved interpretability and potential for rule-based shielding but typically incurs higher annotation and computation costs and remains underexplored in closed-loop autonomy stacks.

Category IV moves beyond passive risk estimation by embedding predicted risk and uncertainty into decision-making and control. As shown in Fig. 2(c), this category typically follows a risk–uncertainty–policy pipeline, where uncertainty estimates adjust warning thresholds, enlarge safety margins, or trigger more conservative actions when the model is less confident. For example, robustness-aware accident anticipation [94] formulates warning issuance as a long-horizon decision problem. In this framework, generative modeling improves robustness under adverse visual conditions, while an actor–critic policy optimizes when to raise an alert under decaying rewards and fixed penalties. This example shows that uncertainty-aware control is not limited to estimating whether a scene is risky; it further determines how the system should respond under uncertain observations. Diffusion models provide a representative generative tool for extending this idea from warning-level decision-making to trajectory-level planning. By sampling multi-hypothesis futures, diverse trajectory candidates, or risk-conditioned motion distributions, diffusion models can represent uncertainty in future interactions and provide richer inputs for downstream policy optimization. They can also be used to refine candidate motions, as summarized in Fig. 3. At the policy-learning level, decision-constrained accident-aware reinforcement learning (DCARL)-style systems [95] embed anticipated risk and uncertainty into a Markov decision process (MDP) or reinforcement learning (RL) loop, enabling the driving policy to adapt online to evolving risk patterns. Overall, Category IV methods are closer to closed-loop autonomy than scalar accident anticipation because they directly connect risk assessment, uncertainty modeling, and action selection. However, they are also more demanding in terms of data, online computation, and safety validation, and existing methods still lack formal guarantees under distribution shift.

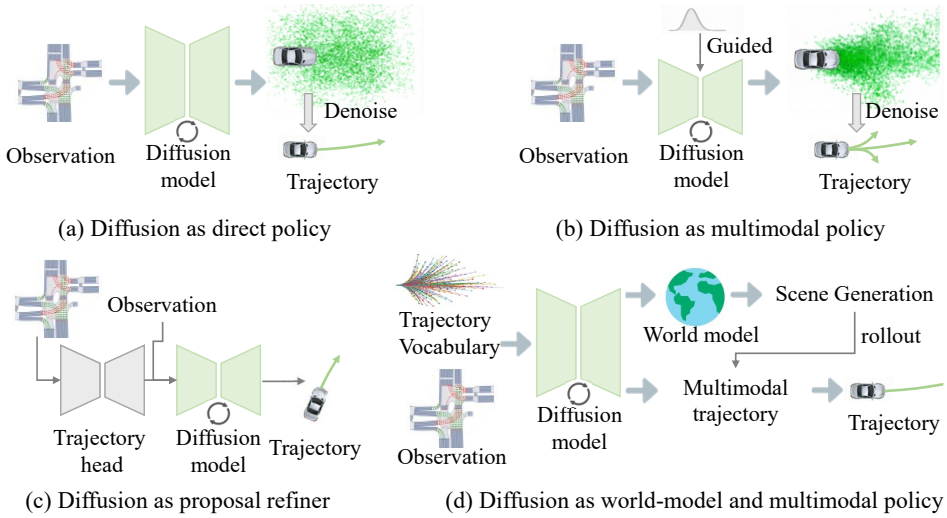


Fig. 3. Representative diffusion-based paradigms. (a) **Diffusion as direct policy:** diffusion models directly denoise trajectory samples conditioned on observations. (b) **Diffusion as multimodal policy:** guided diffusion generates diverse trajectory hypotheses under route, goal, or interaction conditions. (c) **Diffusion as proposal refiner:** diffusion models refine candidate trajectories produced by an upstream trajectory head. (d) **Diffusion as world model and multimodal policy:** diffusion-based planning is coupled with future scene generation or rollout to generate actions under uncertainty.

Category V represents an integrated control paradigm based on world model (WM) and vision-language model (VLM). Different from scalar accident anticipation or localized risk grounding, this category aims to use semantic reasoning and future imagination to support action generation in closed-loop driving. As summarized in Fig. 2(d), the key idea is to transform current observations into future scene representations or rollouts, which are then used to guide decision-making, planning, or direct action prediction. Within this category, WM/VLM modules can be integrated into the driving stack in several ways, as illustrated in Fig. 4. First, a WM/VLM can serve as a scene encoder [96, 97], where high-level semantic or predictive representations are extracted from observations and passed to a downstream planner. Second, it can operate as an asynchronous secondary system [97–100], where a slow WM/VLM reasoning module provides conditions or guidance to a fast planner that runs at the control frequency. Third, it can act as a critic or verifier [101, 102], evaluating candidate trajectories according to semantic risks, traffic rules, or future interaction outcomes before the final trajectory is selected or revised. Fourth, it can be incorporated into an end-to-end model [100, 101, 103–106], where reasoning and action tokens are generated within a unified architecture.

FSDrive [107] is a representative example of this category. Rather than predicting only a risk score or a time-to-accident (TTA) curve, FSDrive uses a large VLM to parameterize a visual world model that imagines future driving contexts. Given current observations, it generates structured future perception representations, such as lane markings and 3D bounding boxes, as intermediate spatiotemporal tokens. Conditioned on these predicted tokens, the model directly outputs control commands or trajectories. In this way, accident anticipation, risk interpretation, and motion planning are integrated within a unified framework. Recent works further explore different forms of WM/VLM-based control. ImagiDrive [92] integrates a VLM-based driving agent with a scene imaginer in a recurrent planning process: the agent first proposes a trajectory, the imaginer generates future frames conditioned on this trajectory, and the imagined frames are then used to refine the plan. Policy World Model (PWM) [108] unifies world modeling and planning

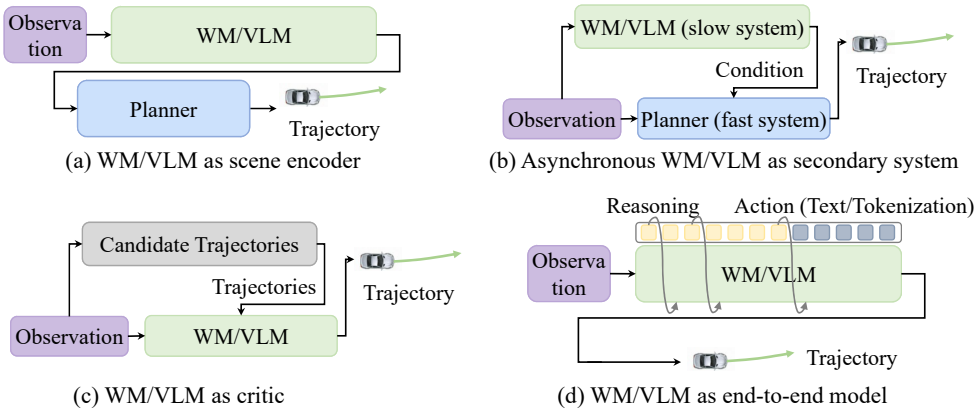


Fig. 4. Representative WM/VLM paradigms. (a) **WM/VLM as scene encoder:** the WM/VLM extracts semantic or predictive representations from observations and provides them to a downstream planner. (b) **WM/VLM as asynchronous secondary system:** a slow WM/VLM reasoning module provides conditions or guidance to a fast planner running at the control frequency. (c) **WM/VLM as critic:** the WM/VLM evaluates candidate trajectories according to semantic risks, traffic rules, or future interaction outcomes. (d) **WM/VLM as end-to-end model:** reasoning and action generation are integrated within a unified model for direct trajectory or action prediction.

in an end-to-end Transformer, using predicted future states as rationales for action prediction. DriveVLA-W0 [109] emphasizes deployability by using future frame prediction as dense self-supervision while introducing a lightweight action expert for real-time control. Controllable video world models, such as VideoGPT [110], can further support multiple future hypotheses and counterfactual evaluation, which is useful for planning under occlusion and long-horizon uncertainty. Despite these advantages, Category V also introduces new safety challenges. First, future rollout and WM/VLM inference can impose substantial computational cost. Second, safety constraints are often learned implicitly from data rather than formally verified. Third, generated futures may be sensitive to prompts, training distributions, or visual ambiguity. For safety-critical scenarios, the key issue is not only whether the imagined future is visually plausible, but also whether its error can be detected and prevented from inducing unsafe planning decisions. Future progress should therefore move beyond visually plausible generation toward verifiable imagination, uncertainty-calibrated rollout evaluation, and constraint-aware action selection.

Table 3 summarizes representative pipelines from the perspectives of model family, input modality, and intermediate cues for risk prediction and prevention. Overall, the literature shows a clear progression from low dimensional geometric surrogates to deep temporal and relational models, and then to richer cues such as attention, 3D consistency, and future tokens generated by WMs and VLMs. Compared with scalar risk scores, these cues provide more structured information for downstream decision making and closed-loop control.

2.3 Traffic Accident Scenario Datasets

From the data perspective, traffic accident scenario datasets form the empirical backbone for the control methods. Early traffic accident datasets are primarily designed to support tasks such as accident detection, accident type classification, and identification of involved objects [129]. The A3D dataset [130] provides annotations for accident categories, bounding boxes of involved objects, and timestamps indicating when accidents are identified. DoTA [131] extends A3D by incorporating more videos and richer annotations, including anomaly types, related objects, and tracking IDs.

Table 3. Representative prior works on preventing safety-critical scenarios. DSA = dynamic soft attention; CNN = convolutional neural network; RNN = recurrent neural network; GCN = graph convolutional network; BNN = Bayesian neural network; GRU = gated recurrent unit; GNN = graph neural network; LSTM = long short-term memory; DSTA = dynamic spatiotemporal attention; RGB = red–green–blue image. "Closed-loop use" indicates how each method's outputs can be consumed by an autonomy stack, namely as trigger, cost, constraint, policy, or within an imagination loop.

Methods	Year	Models	Inputs	Key supervision	Closed-loop use
Hu et al. [111]	2003	3D model	Grayscale frames	trajectory interaction	Constraint
Shan et al. [112]	2014	Autoregression	vehicle position	object distance	Trigger
Chan et al. [113]	2016	DSA, RNN	RGB	object interaction	Trigger
W. Bao et al. [114]	2020	RNN, GCN, BNN	RGB	object interaction	Trigger
H. Kim et al. [115]	2021	Domain Adaptation	RGB, bounding boxes	frame clip feature	Trigger
Drive [116]	2021	CNN	RGB	frame-level hidden state	Trigger
Karim et al. [117]	2021	GRU	RGB	frame riskiness	Trigger
COLLIDE-PRED [118]	2021	Transformer	RGB	object distance	Trigger
Malawade et al. [119]	2022	GNN, LSTM	RGB	object feature relationships	Trigger
Karim et al. [120]	2022	DSTA, GRU	RGB	frame-level hidden state	Trigger
Karim et al. [121]	2023	GRU	RGB, optical flow	frame riskiness	Trigger
GSC [122]	2023	GCN	RGB	object interaction	Trigger
Fang et al. [123]	2023	Transformer, GCN, RNN	RGB, driver attention maps	frame-level hidden state	Trigger
AccNet [89]	2024	3D GCN, GRU	RGB, depth	3D object interaction	Trigger
W3AL [91]	2024	CNN, GRU, LLM	RGB, bounding boxes	frame riskiness, actor localization	Trigger
EQ-TAA [90]	2025	Transformer, Diffusion	RGB	causal consistency triple loss	Trigger
FSDrive [107]	2025	VLM, WM, policy head	RGB, HD map	future perception CoT	Policy
ImagiDrive [92]	2025	VLM agent, driving WM	RGB, map	future frame rollout feedback	Imagination loop
PWM [108]	2025	Transformer, policy WM	RGB	future rollouts as rationales	Policy
DriveVLA-W0 [109]	2025	VLA, WM, MoE action expert	RGB	future prediction self-supervision	Policy
DrivingGPT [124]	2025	AR Transformer	RGB, action tokens	image-action token prediction	Policy
Epona [125]	2025	AR diffusion WM	RGB, trajectory	future video and trajectory prediction	Policy
WorldDrive [126]	2026	DWM, planner	RGB, trajectory vocabulary	future scene and motion representation	Policy
UniDrive-WM [127]	2026	VLM, WM	multi-view RGB, instructions	trajectory-conditioned future image	Policy
DynVLA [128]	2026	VLA, dynamics tokenizer	RGB, BEV, actions	future dynamics tokens	Policy

The CCD dataset [114] further offers accident causes for each video sequence, while DADA [132] explores the role of driver attention in traffic accident prediction by collecting eye-gaze data. Building upon these foundational works, DADA-Seg [133] refines a subset of 313 video sequences with fine grained segmentation masks for semantic objects. Although these datasets have significantly advanced vision-based traffic accident analysis, they primarily support coarse-grained tasks and lack detailed language annotations. These early benchmarks are mainly tailored to vision-based traffic accident detection, where the focus is on spatiotemporal localization of short accident windows (typically 20–60 frames) in long-tailed, imbalanced data, making models vulnerable to background confounding and dataset bias. Most

of them are restricted to dashcam or surveillance RGB videos with limited coverage of adverse weather, nighttime conditions, and diverse road geometries. Only a few [134, 135] explore synthetic or Vehicle-to-Everything (V2X) settings, which constrains their scalability for studying rare, safety-critical scenarios in open-world traffic.

To move beyond coarse accident detection toward higher level understanding and causal reasoning, recent datasets increasingly couple video with language through question answering (QA). SUTD-TrafficQA [136] is the early large-scale benchmark and offers video-QA pairs, such as accident description, forecasting, and reasoning. MM-AU [137] provides textual annotations that cover three aspects of traffic accidents: causality, prevention strategies, and accident types. TAU-106K [138] advances this direction with questions requiring temporal localization and spatial grounding, where textual answers include timestamps and bounding box coordinates. The RoadSocial dataset [139] further broadens the task scope with diverse video QAs for general road events. Meanwhile, AV-TAU [140] enriches the multimodal context of traffic accident scenarios by incorporating audio signals. SafePLUG-Bench [93] further advances the field by uniquely supporting both region-level QA and pixel-level grounding QA. A comprehensive comparison detailing the specific features, annotation types, and sensor modalities of these datasets is summarized in Table 4. Despite this progress, current QA-style benchmarks remain predominantly limited to 2D sensing without dense 3D or bird's-eye view (BEV) data, indicating a need for future multimodal, 3D-aware benchmarks that capture long-tail near-miss dynamics for safety assessment. Recent surveys on LiDAR-based place recognition further highlight the importance of 3D sensing for robust autonomous driving perception and localization [141].

Beyond the task- and annotation-level comparisons, this section distills the accident and near-miss situations covered by existing datasets into a compact catalogue of thirty representative critical scenarios (Table 5) for controller-oriented testing and benchmark design. We first collected scenario descriptions, anomaly types, and author- or annotator-provided tags from the datasets in Table 4, and then normalized them according to the ISO 34502 road–traffic–environment layering. The normalized schema considers road context, key actors, conflict topology, and environmental stressors. Based on this schema, recurring situations were grouped into scenario families, and representative cases were selected to balance coverage and usability. The resulting catalogue is organized by road type, test function, scenario description, environmental condition, and compact tags. These tags encode the main conflict pattern and stressor profile of each scenario, linking high-level scenario semantics to control-relevant test cases. They also support coverage analysis, showing that existing datasets cover common highway and urban conflicts relatively well, while rural hazards and compound environmental conditions remain less represented.

2.4 Critical Scenario Benchmarks

Traditional benchmarks have mainly focused on nominal driving conditions, leaving many critical safety challenges insufficiently tested. For instance, nuScenes provides open-loop metrics and an imbalanced validation set in which 75% of scenarios involve straightforward driving. Such evaluations seldom reveal how models handle rare but high risk events. In real driving, critical scenarios occur in less than 0.03% of daily driving, yet they have disproportionate safety importance. To address this gap, several new benchmarks explicitly target critical scenarios beyond simple highway or straight-road driving. Benchmarks like Bench2Drive [155], NAVSIM v2 [156] and WOD-E2E [157] push evaluation into interactive, long-tail traffic situations where current models struggle.

As illustrated in Fig. 5(a), Bench2Drive is a closed-loop evaluation benchmark developed in CARLA [158]. It expands the coverage of interactive maneuvers, including lane changes, merging, and unprotected intersection traversal. Compared with CARLA Leaderboard v2, Bench2Drive provides broader and more fine grained evaluation through 44 scenario types and 220 short routes. Table 6 further reports closed-loop metrics and a multi ability test, which helps

Table 4. Existing traffic accident datasets. FPV = First-person View (egocentric dashcam videos); TPV = Third-person View (roadside or surveillance cameras); BEV = Bird's-eye View; Bbox = Bounding Box; Mask = pixel-wise segmentation mask; TG = Temporal Grounding; D/N = Day/Night; U/S/R/H = Urban/Suburban/Rural/Highway; Urban (sim) = urban scenes generated in simulation; Var. = varied or diverse conditions mentioned by the dataset but not provided as explicit structured annotations; Syn. = synthetic data generated in simulation environments.

Dataset	Year	Frames	View	Annotations			Traffic Condition		
				Bbox	Mask	TG	Weather	Time	Region
DAD [113]	2016	175K	FPV	✓	–	✓	–	–	–
CADP [142]	2018	518K	TPV	✓	–	✓	Var.	Var.	–
NIDB [143, 144]	2018	1.30M	FPV	–	–	✓	✓	✓(D/N)	U/S/R/H
A3D [130]	2019	208K	FPV	–	–	✓	Var.	–	Var.
GTACrash [134]	2019	228K	FPV	✓	–	✓	Var.	Var.	Var.
VIENA2 [145]	2019	2.25M	FPV	–	–	✓	✓	✓	U/S/R/H
Drive-Anomaly106 [146]	2019	11K	FPV	–	✓	✓	Var.	Var.	–
RetroTrucks [147]	2020	36K	FPV	–	–	–	Var.	Var.	Var.
ADV [148]	2020	10K	FPV	✓	–	✓	–	–	–
CTA [149]	2020	853K	FPV+TPV	✓	–	✓	Var.	Var.	Var.
DADA [132]	2021	658K	FPV	–	–	✓	✓	✓(D/N)	U/S/R/H
DADA-Seg [133]	2021	12K	FPV	–	✓	✓	✓	✓(D/N)	U/S/R/H
CCD [114]	2021	75K	FPV	✓	–	✓	✓	✓(D/N)	U/S/R/H
SUTD-TrafficQA [136]	2021	1.90M	FPV+TPV	–	–	–	Var.	Var.	Var.
DoTA [131]	2022	732K	FPV	✓	–	✓	✓	✓(D/N)	U/S/R/H
MP-RAD [150]	2022	366K	TPV	–	–	✓	✓	✓	–
TRA [151]	2022	43K	FPV	–	–	✓	–	–	–
TAD [152]	2022	298K	TPV	✓	–	✓	–	–	U/S/R/H
DeepAccident [135]	2023	57K	FPV+TPV	✓	✓	✓	✓	✓	Urban (sim)
CTAD [153]	2023	792K	TPV	–	–	✓	✓	–	Urban (sim)
ROL [121]	2023	100K	FPV	✓	–	✓	✓	✓(D/N)	U/S/R/H
MM-AU [137]	2024	2.19M	FPV	✓	–	✓	✓	✓(D/N)	U/S/R/H
TAU-106K [138]	2025	–	FPV+TPV+Syn.	✓	–	✓	Var.	Var.	Multi-region
Nexar [154]	2025	1.70M	FPV	–	–	✓	✓	✓	U/S/R/H
RoadSocial [139]	2025	14M	FPV+TPV	–	–	✓	Var.	Var.	100+ countries
AV-TAU [140]	2024	3.16M	FPV+TPV	–	–	✓	Var.	Var.	Multi-region
SafePLUG-Bench [93]	2025	2.26M	FPV+BEV	✓	✓	✓	✓	✓	U/S/R/H

identify whether performance bottlenecks come from route completion, efficiency and comfort tradeoffs, or specific driving skills such as merging, overtaking, emergency braking, yielding, and traffic sign compliance.

Despite recent progress, even the leading reported baselines remain limited in route success and interaction intensive scenarios. The detailed ability breakdown shows that merging, overtaking, emergency braking, and yielding are still difficult for current planners. This suggests that imitation learning dominated pipelines often struggle to learn robust negotiation behaviors from offline logs, where rare safety related interactions are sparsely represented. Bench2Drive therefore provides a useful controlled stress test by exposing weaknesses that aggregate scores may hide and by motivating methods that better handle interaction, uncertainty, and recovery in closed-loop execution.

NAVSIM v2 provides a complementary path to large-scale closed-loop evaluation by leveraging real nuPlan logs [159] and injecting targeted perturbations that emulate trajectory drift and compounding errors. As illustrated in Fig. 5(b), it adopts a two stage pseudo-simulation protocol: the first stage evaluates planners on recorded sensor observations, while the second stage perturbs the ego endpoint and uses neural 3D reconstruction to render synthetic follow-up frames. This design exposes planners to distribution shift without relying on a fully interactive simulator, and directly

Table 5. 30 representative critical scenarios spanning multiple road types. Environment (Env.) denotes the lighting condition (D = day, N = night, IL = indoor or tunnel lighting). Coverage tags summarize the main conflict types (C-*) and key stressors (S-*): C-LK (lane keeping/curve), C-MG (merge/topology/markings), C-CI (cut in/weaving), C-RE (rear-end/stationary lead), C-SO (static obstacle/debris), C-CR (crossing event), C-MX (mixed/unstructured interaction); S-CZ (construction/temporary control), S-CG (congestion), S-TN (tunnel/indoor lighting), S-GL (glare), S-RN (rain).

Road Type	Test Function	Scenario	Env.	Tag
Highway (including tunnels and ramps)	Curve speed-limit and lane-keeping test	Construction zone with a detour ahead	N	C-MG|S-CZ
		Escape lane at sharp curve ahead	D	C-LK
	Vehicle target detection and response test	Stationary or maintenance vehicle ahead	D	C-RE
		Oversized cargo vehicle ahead	D	C-RE
		Unexpected crashed vehicle on the highway	N	C-RE
		Continuous lane changes on the highway	D	C-CI
		Sudden cut in of a vehicle ahead during congestion	D	C-CI|S-CG
		Vanishing lead vehicle on highway	D	C-RE
		Aggressive cut in at the highway entrance	D	C-CI
		Abnormally stationary lead vehicle inside a tunnel	IL	C-RE|S-TN
	Static obstacle detection and response test	Fallen tires on highway	N	C-SO
	Irregular dynamic object detection and response test	Flattened plastic bag ahead	D	C-SO
		Laterally crossing wild boar	D	C-CR
	Traffic facility and road-marking response test	Three-lane to two-lane merge ahead	D	C-MG
		Road marking confusion zone	D	C-MG
		Temporary construction on the highway	D	C-MG|S-CZ
		Truck encountered in construction zone	N	C-MG|S-CZ
		Flexible bollards placed at the gore area ahead	D	C-MG
	Special Env. detection and response test	Obstacle avoidance during nighttime rainstorm	N	C-SO|S-RN
		Glare from oncoming headlights	N	C-MX|S-GL
Rural road	Irregular dynamic object detection and response test	Crossing sheep ahead	D	C-CR
	Traffic facility and road-marking response test	Driving through traffic lights	D	C-MG
Urban road (including intersections and parking lots)	VRU detection and response test	Child crossing under high-beam glare	N	C-CR|S-GL
		Urban village driving	D	C-MX
		Irregular tricycle movement in congestion	D	C-MX|S-CG
		Mixed pedestrian–vehicle interaction	D	C-MX
		Delivery rider intruding at intersection	D	C-CR
		Child darting out behind during reversing	IL	C-CR|S-TN
	Irregular dynamic object detection and response test	Dropped cargo ahead	D	C-SO
	Traffic facility and road-marking response test	Width restriction ahead	D	C-MG

stress tests robustness to small deviations that can accumulate into safety-critical outcomes. NAVSIM v2 reports an extended EPDMS metric that aggregates safety, rule compliance, efficiency, and comfort via multiplicative penalties, so that a single infraction can significantly reduce the overall score. In practice, results on the benchmark show that even state-of-the-art planners remain far from the maximum EPDMS, highlighting that robust, comprehensive driving proficiency under uncertainty is still a major bottleneck.

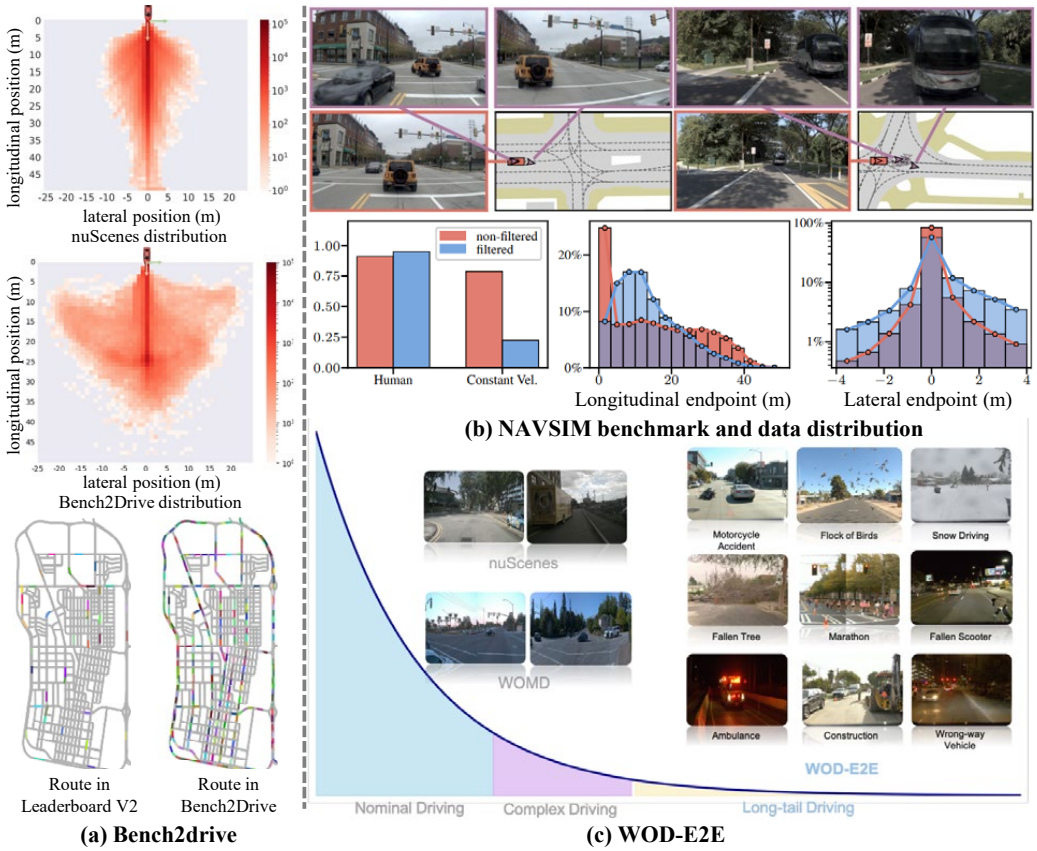


Fig. 5. Representative safety-critical scenario benchmarks. The figure shows Bench2Drive for closed-loop evaluation in CARLA, NAVSIM v2 for perturbation-based pseudo closed-loop evaluation on real driving logs, and WOD-E2E for rare long-tail real world scenarios. The sample images are from Bench2Drive [155], NAVSIM v2 [156], and WOD-E2E [157].

WOD-E2E focuses explicitly on rare real world long-tail scenarios (Fig. 5(c)), collecting 4,021 short segments with construction zones, unusual obstacles and atypical agent behaviors. Instead of relying on a single logged future, WOD-E2E adopts the rater feedback score (RFS), which compares predicted trajectories against human preference labels over multiple plausible futures, making it particularly suitable for safety–efficiency trade-offs and reasonable rule-breaking in emergencies.

Across Bench2Drive and NAVSIM v2, three consistent lessons can be drawn. First, aggregate scores may hide skill specific bottlenecks. The multi ability test in Table 6 shows that even strong methods still struggle with interactive skills such as merging, overtaking, and yielding, indicating that negotiation under uncertainty remains a major failure mode. Second, closed-loop performance depends on multiple objectives rather than a single accuracy measure. On Bench2Drive, higher efficiency does not necessarily lead to higher success or comfort, since aggressive progress may increase near miss risk while overly conservative policies may reduce mobility. Third, robustness and deployability have become key constraints. The multiplicative EPDMS penalty in Table 7 shows that weaknesses in safety, rule compliance, comfort, or progress can dominate the final score, while perturbation driven pseudo closed-loop evaluation exposes sensitivity to

Table 6. Closed-loop metrics and multi-ability test results on Bench2Drive [155]. Driving score (DS) is a composite closed-loop score in the range $[0, 100]$. Success denotes the percentage of routes completed without critical infractions. Efficiency reflects average progress and speed, and comfort measures ride smoothness based on acceleration and jerk. The multi-ability test reports success rates (%) on five driving skills and their mean. * denotes expert feature distillation. Best results are in **bold**.

Methods	Closed-loop Metric ↑				Multi-Ability Test (%) ↑					
	Efficiency	Comfort	Success	DS	Merging	Overtaking	Emergency Brake	Give Way	Traffic Sign	Mean
TCP* [160]	54.26	47.80	15.00	40.70	16.18	20.00	20.00	10.00	6.99	14.63
TCP-ctrl* [160]	55.97	**51.51**	7.27	30.47	10.29	4.44	10.00	10.00	6.45	8.23
TCP-traj* [160]	76.54	18.08	30.00	59.90	8.89	24.29	51.67	40.00	46.28	34.22
ThinkTwice [161]	76.93	16.22	3.13	62.44	27.38	18.42	35.82	50.00	54.23	37.17
DriveAdapter* [162]	70.22	16.01	33.08	64.22	28.82	26.38	48.76	50.00	56.43	42.08
AD-MLP [163]	48.45	22.63	0.00	18.05	0.00	0.00	0.00	0.00	4.35	0.87
UniAD-T. [164]	123.92	47.04	13.18	40.73	8.89	9.33	20.00	20.00	15.43	14.73
UniAD-B. [164]	129.21	43.58	16.36	45.81	14.10	17.78	21.67	10.00	14.21	15.55
VAD [165]	157.94	46.01	15.00	42.35	8.11	24.44	18.64	20.00	19.15	18.07
DriveTransformer [166]	100.64	46.01	35.01	63.46	17.57	35.00	48.36	40.00	52.10	38.60
Hydra-Next [167]	197.76	20.68	50.00	73.86	40.00	64.44	61.67	50.00	50.00	53.22
TF++ [168]	**245.10**	25.48	67.26	84.21	58.75	57.77	83.33	40.00	82.11	64.39
R2SE [169]	243.89	23.26	69.54	86.28	53.33	61.25	**90.00**	50.00	84.21	67.76
HiP-AD [170]	203.12	19.36	69.09	86.77	50.00	**84.44**	83.33	**50.40**	72.10	65.98
DiffRefiner [171]	-	-	**71.40**	**87.10**	**63.80**	60.00	85.00	50.00	**86.30**	**69.00**

Table 7. Comparison with state-of-the-art methods on the NAVSIM v2 [156] benchmark. Ego status denotes a state-only baseline that conditions the planner on the navigation goal and low-dimensional ego kinematic states without using perception inputs. NC = No at-fault Collision; DAC = Drivable Area Compliance; DDC = Driving Direction Compliance; TLC = Traffic Light Compliance; EP = Ego Progress; TTC = Time to Collision; LK = Lane Keeping; HC = History Comfort; EC = Extended Comfort; EPDMS = Extended Predictive Driver Model Score in the range $[0, 100]$. Best results are in **bold**.

Methods	NC ↑	DAC ↑	DDC ↑	TLC ↑	EP ↑	TTC ↑	LK ↑	HC ↑	EC ↑	EPDMS ↑
Ego Status	93.1	77.9	92.7	99.6	86.0	91.5	89.4	**98.3**	85.4	64.0
TransFuser [172]	96.9	89.9	97.8	99.7	87.1	95.4	92.7	**98.3**	87.2	76.7
HydraMDP++ [173]	97.2	97.5	**99.4**	99.6	83.1	96.5	94.4	98.2	70.9	81.4
DriveSuprem [174]	97.5	96.5	**99.4**	99.6	**88.4**	96.6	95.5	**98.3**	77.0	83.1
ARTEMIS [175]	98.3	95.1	98.6	**99.8**	81.5	97.4	96.5	**98.3**	-	83.1
DiffusionDrive [176]	98.2	95.9	**99.4**	**99.8**	87.5	97.3	**96.8**	**98.3**	**87.7**	84.5
DriveVLA-W0 [109]	**98.5**	**99.1**	98.0	99.7	86.4	**98.1**	93.2	97.9	58.9	**86.1**

compounding errors. Together, these benchmarks suggest that improving critical scenario handling requires explicit mechanisms for interaction, uncertainty, and recovery, as well as balanced objective design that accounts for safety, rule compliance, mobility, comfort, and real-time feasibility.

3 Verification and Validation for ADS

3.1 Scenario Construction and Description

Scenario-based validation serves as the bridge between unstructured real world traffic logs and reproducible, parameterized test cases. To operationalize this process, the industry has standardized scenario descriptions through hierarchical ontologies. While early frameworks focused on abstraction levels (functional, logical, concrete) [177], modern approaches have converged on a multi-layer architecture to capture the complexity of critical driving environments.

Scenario descriptions are commonly organized through layered architectures. Early studies separated road geometry, dynamic actors, and environmental factors [177, 178]. Bagschik et al. [179] further proposed a five-layer structure covering road, infrastructure, temporary modifications, objects, and environment, while Bock et al. [180] extended

it with a sixth layer of digital information. In this survey, we do not reproduce the well-known layered model as a standalone figure. Instead, we use it as a reference structure for organizing critical scenario parameters, including road topology, traffic facilities, temporary work-zone changes, dynamic actors, weather and lighting conditions, HD map errors, and V2X communication delays. Such layering helps transform high-level critical scenario descriptions into machine-executable parameters for simulation injection and scenario-based verification and validation.

3.1.1 Traditional Methods for Critical Scenario Construction. From the perspective of safety validation, it is not enough to describe nominal scenarios; one must systematically expose critical scenarios that stress the limits of an ADS. Classical analyses show that verifying advanced ADS solely by real world mileage would require on the order of billions of kilometers to reach an acceptable confidence level, which is practically infeasible. At the same time, the ODD of high level automation is high dimensional and strongly coupled across human-vehicle-environment factors, and many safety relevant situations are rare long-tail events rather than frequent patterns. To cope with this, projects such as PEGASUS [181] advocate reconstructing real scenarios in virtual environments to improve test efficiency, while ISO 21448 [62] introduces a four-quadrant matrix over {known, unknown} $\times$ {safe, hazardous} scenarios, emphasizing that unknown hazardous scenarios should be identified, analyzed and gradually converted into known and mitigated ones through scenario searching and testing.

Traditional critical scenario construction methods largely follow this philosophy and can be grouped into three broad categories. The first category comprises combinatorial and ontology-based approaches. Here, scenario elements (road segments, traffic participants, weather, etc.) are discretized into parameter sets, often structured by an ontology, and combinatorial test (CT) or test-matrix (TM) techniques are used to systematically enumerate combinations. Representative works employ backtracking algorithms, Monte Carlo sampling and CT+TM with Bayesian optimization to balance coverage and complexity in the generated scenarios [182–184]. These methods provide explicit, traceable models and can guarantee a certain degree of coverage, but they tend to produce a large number of redundant or low risk scenarios, making it difficult to focus on truly safety-critical conditions.

The second category consists of optimization-based search methods that treat the pursuit of critical scenarios as an optimization problem. Risk indicators such as collision speed, minimum distance or TTC are embedded into cost functions, and evolutionary algorithms, tree search or RL-based planners are used to search the parameter space for scenarios that maximize risk while remaining physically plausible [185–188]. These methods are effective at uncovering high-risk corner cases and can significantly improve test efficiency, but their performance depends heavily on the design of the cost function and parameter bounds, and they may miss moderately risky yet safety relevant scenarios that do not strongly excite the chosen objective.

The third category encompasses data-driven and adversarial approaches that learn scenario structure directly from traffic data. Clustering-based methods extract typical maneuver patterns from speed and yaw rate trajectories, while neural generators trained on highway datasets, e.g., HighD synthesize realistic motion trajectories [189, 190]. More recent work combines generative models with adversarial RL: high-risk initial scenes are enriched using tabular or conditional tabular generative adversarial network (CTGAN)-style generators, and an RL-based adversarial agent interacts with the ADS in simulation to search for dynamic critical scenarios, thereby jointly enhancing scenario coverage and risk severity while keeping behaviors realistic [191]. These approaches still face challenges in controllability, interpretability and quantifying what portion of the relevant risk space has been actually explored.

In summary, traditional CT-based, optimization-based and data-driven scenario construction techniques provide the basic toolbox for building critical scenario libraries, but no single class guarantees coverage, criticality and engineering

controllability. At the same time, these limitations motivate the exploration of more expressive generative models and foundation-model-based frameworks for scenario generation, which are discussed in the next subsection.

3.1.2 Foundation Models for Critical Scenario Construction. Beyond classical clustering and neural-network-based generators, recent work has started to explore foundation models [192–195], including variational autoencoder (VAE), generative adversarial network (GAN), diffusion models, and LLMs, for critical scenario generation.

As sketched in Fig. 6, VAEs and diffusion models follow an encoder–latent–decoder paradigm: an encoder or noising process $q_\phi(z \mid x, c)$ maps a real scene x and condition c to a latent variable z, while a decoder or denoising network $p_\theta(\bar{x} \mid z, c)$ reconstructs or synthesizes a new scene $\bar{x}$. GANs instead train a generator $G(z, c)$ and a discriminator $D(x, \bar{x}, c)$ in an adversarial game, where the discriminator distinguishes real scenes from synthetic ones (labeled 1/0). VLMs couple a vision encoder with an LLM via a connector that projects visual features into the LLM token space, enabling language-conditioned editing and generation of traffic scenarios from textual descriptions or instructions.

Building on these architectures, early work on critical scenario construction primarily adopted VAE- [196] and GAN-based [197] generators. A conditional VAE learns a low-dimensional latent space of trajectories or scenes, from which potentially critical cases can be sampled by importance weighting or by steering latent variables toward unsafe regions. GAN-based approaches train a generator to produce map-consistent traffic scenes or surrounding-agent trajectories that fool a discriminator, sometimes augmented with risk-aware losses to bias generation toward near-miss or collision-prone situations. Phenomena such as mode collapse, limited diversity in long-tail behaviors, and difficulties in conditioning on rich scene priors have motivated a shift toward diffusion- and VLM-based generators.

Diffusion-based generators [192, 198] learn joint distributions over multi-agent behaviors conditioned on maps and goals and can be guided by cost functions, game-theoretic solvers, or risk priors. This allows controllable sampling of both nominal and corner-case traffic, so that the same generative model supports routine regression testing, stress testing, and closed-loop robustness evaluation under a unified framework. RGB-D diffusion models [194] for long horizon, geometry-consistent scene generation further provide photorealistic camera and LiDAR streams, which are increasingly used to validate perception stacks and sensor fusion modules in a simulation–in-the-loop manner. At the data level, map- and trajectory-centric generators [191, 195, 199] enable large-scale augmentation of rare interactions such as cut ins, forced merges, or red-light violations, and have been shown to improve safety metrics of end-to-end driving models when evaluated on real world benchmarks. In parallel, VLM-driven frameworks [200–204] treat scenario design and test case selection as a language-guided reasoning problem: multi-agent VLM systems parse regulations and accident reports, propose critical interaction patterns, and iteratively edit the behaviors of selected "adversarial" participants to create progressively more challenging scenes that expose weaknesses in the ADS policy. Overall, these foundation model-based approaches provide rich semantic control, scalable coverage of long-tail events, and tighter coupling between generation and evaluation, and are beginning to turn generative models from mere data sources into active tools for scenario-based testing and safety assessment; however, ensuring physical consistency, enforcing hard safety constraints, and quantifying coverage and validity of the generated scenarios remain open research challenges. Representative VLM- and diffusion-based scenario generators are summarized in Tables 8 and 9, respectively.

3.2 Verification and Validation Methods Methods

3.2.1 XiL Simulation. Once constructed, representative scenarios should be executed through verification and validation methods that balance scalability with fidelity. To bridge the gap between pure software simulation and costly road tests, the industry employs a continuum of XiL methods, as illustrated in Fig. 7.

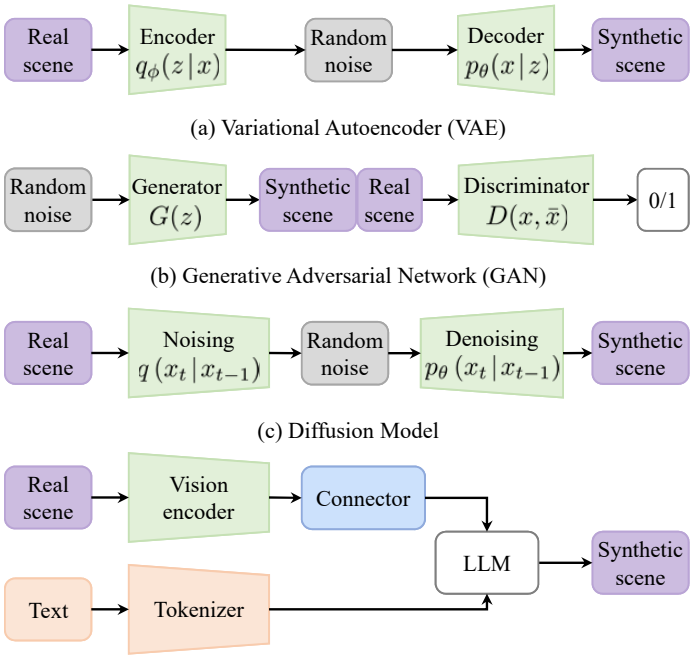


Fig. 6. Schematics of representative generative model architectures, where 0/1 indicates the real or synthetic label.

Table 8. Summary of scenario generation studies using VLMs. FPV = first-person view; BEV = bird's-eye view; RAG = retrieval-augmented generation; XML = extensible markup language.

Methods	Input					Model	Technique	Method Descriptions	Simulator
	Text	Image	Video	Dataset	Database				
CurricuVLM [205]	✓	✓	BEV	Waymo [206]	–	GPT-4o LLaVA	CoT	Curriculum-driven critical-scene synthesis and selection	Metadrive [207]
AutoScenario [208]	✓	✓	FPV	SUMO [209] CARLA	NHTSA GPS	GPT-4o	CoT	Text-grounded scene construction and editing	CARLA
LLM-Attacker [200]	✓	–	–	Waymo	Accident reports, traffic rules	LLaMA	Multi-agent adversarial search	Closed-loop adversary policy synthesis for stress testing	CARLA
Seeking to Collide [202]	✓	–	–	Waymo	Driving logs, accident cases	DeepSeek-R1, DeepSeek-V3	Online scene editing and RAG	Retrieval-guided iterative scenario hardening	CARLA
Generating OOD [210]	✓	–	–	nuScenes [211]	NHTSA crash reports	GPT-4o	CoT, tree generation and augmentation	OOD template mining and parameterized augmentation	CARLA
DriveGen [212]	✓	✓	BEV	Argoverse2 [213]	Vehicle asset DB	Claude-3, GPT-4-turbo	LLM+RAG init; VLM goal selection; diffusion planner	Goal-conditioned multi-stage generation for diverse traffic	SMARTS [214]
OmniTester [215]	✓	✓	–	–	Road network DB, crash reports	LLM + VLM	Prompt engineering + SUMO tools + RAG	Text-conditioned from-scratch generation (road + vehicles) with iterative evaluation	SUMO

Table 9. Summary of scenario generation studies using diffusion models. DDPM = Denoising Diffusion Probabilistic Model; LDM = Latent Diffusion Model; DiT = Diffusion Transformer; MDP = Markov Decision Process.

Methods	Output	Input				Technique	Base Model	Dataset
		Road Topology	Initial State	Text Prompt	Bounding Boxes			
CCDiff [198]	Traffic flow	✓	✓	-	-	MDP as guidance	DDPM	nuScenes
DiffScene [216]		✓	✓	-	-	Gradient-based guidance	DDPM	CARLA
Lu et al. [217]		✓	✓	-	-	Gradient-based guidance	DDPM	nuScenes
AdvDiffuser [218]		✓	✓	-	-	Gradient-based guidance	LDM	nuScenes
SafeSim [219]		✓	✓	-	-	Partial diffusion	DDPM	nuPlan nuScenes
VBD [220]		✓	✓	-	-	Gradient-based guidance	DiT	Waymo
Zhong et al. [221]		✓	✓	✓	-	LLM-generated loss function	DiT	nuScenes
LD-Scene [201]		-	-	✓		LLM-driven scene initialization	LDM	nuScenes
DrivingGen [222]	Driving video	-	-	✓	-	Temporal shift adapter	LDM	DoTA
AVD2 [223]		-	-	✓	-	Adapting existing methods	DiT	MM-AU
SynAD [195]		✓	✓	-	-	Diffusion-based synthetic data pipeline	LDM	SynAD
AutoScape [194]	RGB-D scenes	✓	✓	✓	-	Geometry-consistent 3D diffusion	DiT	nuScenes, KITTI [224]

Standard approaches such as Software-in-the-Loop (SiL) facilitate rapid, iterative verification of algorithms in purely virtual environments [225–228]. Moving toward hardware integration, Hardware-in-the-Loop (HiL) incorporates real electronic control units (ECUs) to assess real-time execution feasibility and signal processing robustness [229]. While effective for functional regression testing, these methods inherently rely on simplified vehicle dynamics models and synthetic sensor data. They often fail to capture the complex, nonlinear vehicle behaviors and rich perception uncertainties that characterize safety-critical near-miss scenarios in the real world.

To address these fidelity gaps, Vehicle-in-the-Loop (ViL) testing couples a complete physical vehicle with a virtual traffic environment. In typical ViL setups [230, 231], as illustrated in Fig. 8(a), Channel ① [232] bypasses perception and feeds simulator ground-truth states directly into planning and control modules, which is simple and suitable for early-stage functional testing but ignores real sensor artifacts. Channel ② [233] attaches camera, LiDAR or radar models to the virtual ego vehicle and streams synthetic sensor data to the ECUs or perception software, thereby exercising the perception stack but still keeping the sensors themseflves out of the loop. Channel ③ [234] goes one step further by converting virtual objects into physical excitation signals, so that the real sensors are stimulated by a virtual world while the vehicle drives on a test track. These architectures progressively increase fidelity, yet they often rely on dedicated indoor facilities or simplified environments, and they provide limited support for systematically validating long-tail, critical scenarios under realistic vehicle dynamics.

To bridge the gap between high fidelity real vehicle behavior and controllable critical scenario replay, we summarize a virtual–real fusion ViL reference architecture as a design pattern for future validation, as illustrated in Fig. 8(b). This architecture is not presented as a validated system, but as a synthesized blueprint distilled from recurring limitations of existing XiL pipelines, including limited realism of perception inputs, restricted exposure to rare interactive hazards, and the difficulty of scaling experiments beyond empty closed tracks.

In this reference design, a real test vehicle provides authentic ego dynamics, actuator limits, and timing delays, while a simulator supplies controllable surrounding agents and hazards. The vehicle is equipped with localization and navigation sensors and a lightweight map of the proving ground, enabling continuous synchronization between the physical ego pose and its virtual counterpart. Real perception sensors, especially front cameras, contribute background

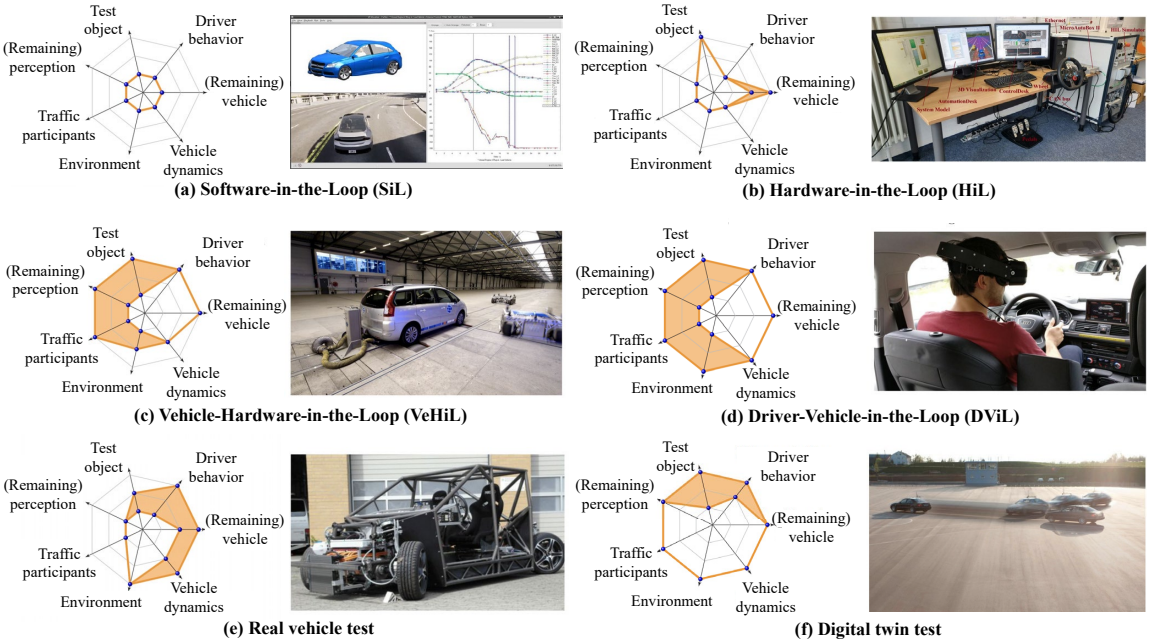
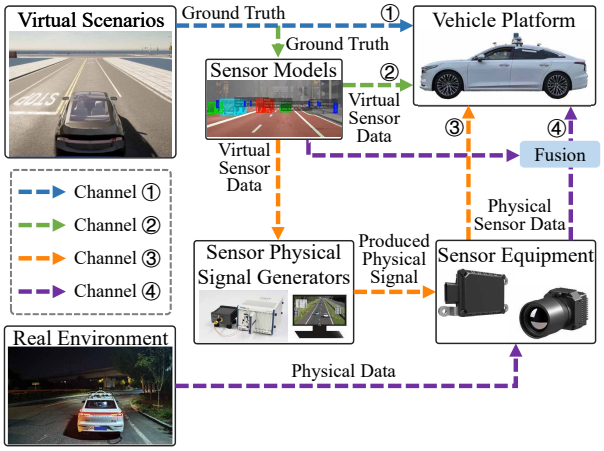


Fig. 7. Qualitative comparison of fidelity coverage across different validation configurations. The radar plots illustrate a progression from purely virtual simulation (center) to physical reality (outer edge). As validation evolves from SiL to real-vehicle testing, the inclusion of physical components increases, highlighting the trade-off between experimental controllability and environmental realism.

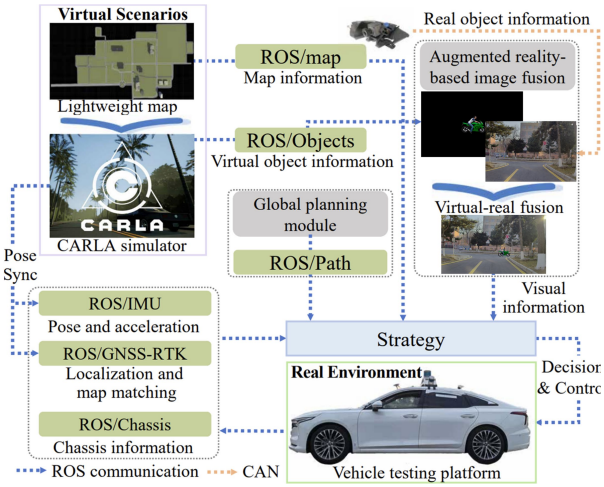
imagery that preserves real lighting and motion effects. Risk relevant virtual agents and objects are instantiated in simulation and rendered within the camera frustum, then fused with the real images through an augmented reality pipeline to produce mixed reality observations that combine real ego motion with repeatable and scriptable interactions.

Viewed as a validation oriented construct, this virtual–real fusion pattern highlights several capabilities that are difficult to obtain simultaneously with purely simulated or purely physical tests. First, it enables repeatable stress testing of rare events with controlled variations in timing, friction, and sensing conditions, supporting statistically meaningful comparisons of risk-aware planners. Second, because the ego platform is real, it naturally exposes hardware and latency related failure modes that pure simulation can mask, such as actuator saturation, tyre slip under low friction, and embedded scheduling delays. Third, the synchronized logs produced by such a setup can serve as reusable assets for evaluation and dataset building, pairing fused sensor inputs with ground truth scenario parameters for benchmarking anticipation, world modeling, and closed-loop control.

Finally, the comparative analysis in Fig. 7 visualizes the fundamental trade off in current validation pipelines. As physical fidelity increases, the ability to precisely control environmental variables and traffic participants typically diminishes. Pure simulation offers control but limited fidelity, while road testing offers high fidelity but limited repeatability. This dichotomy motivates the proposed virtual–real fusion ViL architecture, which aims to bridge this gap by combining the high fidelity physical ego dynamics of real world testing with the flexible and risk-free scenario controllability of simulation.



(a) Comparison of the ViL platform



(b) Information transfer process

Fig. 8. ViL architectures for critical scenario validation with augmented-reality virtual-real fusion. (a) compares existing ViL architectures with the proposed architecture, which injects virtual hazardous objects into real camera streams to combine real vehicle execution with controllable scenario generation. (b) shows the information flow among virtual scenario generation, sensor-level fusion, vehicle state feedback, and control execution. This architecture enables controllable hazardous events to be tested while preserving real vehicle dynamics and onboard execution.

3.2.2 Real-vehicle Testing and Naturalistic Driving. The traditional mileage-based validation approach relies on statistical demonstrations of reliability through extensive real world operation. However, this method is limited by the rarity of safety-critical events [235]. Winner et al. [236] statistically estimated that demonstrating an ADS to be twice as safe as a human driver would require more than 6 billion kilometers of driving, which is economically and temporally infeasible even for large-scale fleets.

While real world testing provides the highest environmental and interaction fidelity, critical safety events are sparse in nominal traffic. Major naturalistic driving studies (NDS), such as the 100-Car Study [237], SeMiFOT [238], and the Ann

Arbor Safety Pilot [239], have successfully captured nominal driving behaviors and V2X interactions but confirmed that severe crash imminent scenarios remain rare outliers. Furthermore, in regions with strict regulatory constraints, such as China, road testing is often confined to closed proving grounds or limited open road pilots [240], further restricting the diversity of exposure. Consequently, real vehicle testing alone is insufficient for systematic safety verification and validation. In the V-model sense, controllable scenario execution in simulation, XiL, ViL, and proving-ground tests primarily serves verification, because it checks whether the ADS satisfies predefined safety requirements under concrete critical scenarios. These verification results can further contribute to validation when they are integrated with ODD coverage analysis, scenario representativeness assessment, and evidence for safety assurance [241].

3.2.3 Digital Twins as an Enabling Foundation. The virtual-real fusion architecture described above is methodologically grounded in digital twin (DT) technology. While DT originated as a core concept of Industry 4.0 [242] for manufacturing monitoring, its application in autonomous driving represents a paradigm shift from static simulation to dynamic, bi-directional cyber-physical synchronization [243, 244]. In the context of safety validation, a DT does not merely simulate the environment but maintains a real-time, high fidelity mapping between the physical test vehicle and its virtual counterpart, serving as the technical backbone for the augmented reality injection described in Fig. 8(b).

Current research has begun to explore DT-based validation, though often with limited scope compared to a full sensor fusion ViL. Methodologically, frameworks proposed in [245] and [246] outline the theoretical requirements for DT-based driving scenarios but lack extensive experimental validation. On the experimental side, existing implementations typically focus on specific subsystems rather than critical scenario replay. For instance, Solmaz et al. [247] validated trajectory planning algorithms using a DT hybrid setup in the EU INFRAMIX project, while Szalay et al. [248] demonstrated a V2X-based DT for sending virtual targets to an ego vehicle. However, these approaches often rely on object-level data injection (bypassing perception sensors) or focus heavily on communication latency rather than complex visual interaction [249, 250].

Compared with object-level ground truth injection, the proposed virtual-real fusion validation framework provides a more demanding validation pattern because virtual hazards are rendered into perception facing sensor streams rather than directly injected as object lists. This design closes the validation process at the sensor-level and enables the complete autonomy stack, from raw perception to control, to be evaluated under conditions closer to real road tests while preserving the safety and controllability of virtualization. The maturation of DT technology, especially in synchronization latency control and mixed reality rendering fidelity, is essential for virtual-real fusion ViL testing.

4 Challenges and Future Directions

4.1 Rare Critical Scenarios and Dataset Limitations

Critical, pre-crash, and near-miss scenarios are rare and follow a long-tail distribution, and they are insufficiently covered by existing datasets [235]. Many accident and traffic anticipation benchmarks focus on several frequent crash types, short temporal windows, and monocular RGB dashcam views. Thus, they provide limited coverage of adverse weather, complex road geometry, detailed vehicle dynamics, and cooperative V2X settings. Recent long-tail datasets and V2X or cooperative perception benchmarks [157, 251] have extended the coverage of rare events and interactions among multiple agents. However, most of them are designed for specific subtasks, e.g., perception, question answering, or motion prediction, rather than for closed-loop risk assessment across sensing, planning, and control.

A key direction is to build multimodal datasets with multiple interacting agents, focusing explicitly on rare near-miss and pre-crash events across highways, rural roads, and dense urban intersections. Such datasets could include

synchronized camera, LiDAR, and radar streams, vehicle dynamics signals, V2X messages, and BEV or HD maps. Scenario taxonomies, such as the thirty representative critical scenarios summarized in this survey, can provide a blueprint for balanced data collection and help expose systematic gaps in existing coverage. Online mining of fleet data based on surrogate safety indicators, including conflict measures and hard braking statistics, could be combined with accident boundary analysis and targeted logging around high-risk conditions. This would allow non-crash data, such as near misses and mild conflicts, to be captured together with actual crashes. In parallel, long-tail datasets for interactive situations and cooperative maneuvers, such as platooning, emergency corridors, and complex merging, could be curated and connected with generative scenario models. This can augment rare patterns in a controllable way, rather than relying on uniform random sampling.

4.2 Causal and Uncertainty-aware Risk Modeling

Many existing risk models are still mainly associative. They may conflate true causal accident factors with spurious background correlations, and they often lack well calibrated uncertainty estimates. Models that perform well on a given benchmark may become brittle under distribution shifts in behavior, weather, sensor quality, or adversarial perception perturbations [252, 253]. This limits their reliability as safety related triggers for AEB, trajectory replanning, and cooperative V2X maneuvers.

Future work could combine interaction graphs and WMs with explicit causal reasoning and uncertainty estimation. One promising direction is to formulate accident and near miss prediction as a structural causal model over agents, road layout, and environmental factors [254–256]. Counterfactual training based on diffusion perturbations, causal intervention samples, or synthetic counterfactual scenes can then be used to separate causal cues from spurious cues. In parallel, uncertainty estimation methods, including Bayesian deep networks, ensembles, and conformal prediction [257, 258], can produce calibrated risk scores and prediction sets. In this way, risk estimators can indicate not only the likelihood of a crash, but also the confidence of the prediction. These calibrated risk signals could be propagated into downstream planners and safety monitors through explicit risk budgets or probabilistic safety constraints. They could also be combined with causal attribution, such as identifying the agent or factor that contributes most to the risk, to support both online decision making and retrospective accident analysis.

4.3 From Accident Anticipation to Verifiable Closed-Loop Control

Many current methods are still evaluated mainly through accident anticipation in open-loop settings, where prediction accuracy has only a limited correlation with closed-loop driving safety and comfort. In contrast, end-to-end controllers are often optimized for route completion and comfort metrics [159, 259], but they usually lack explicit representations of accident risk or safety envelopes. As a result, dangerous failure modes may remain hidden until they appear as collisions in simulation or on the road. Recent long-tail closed-loop benchmarks partly reduce this gap by introducing human preference scores and multidimensional safety metrics. However, they still do not provide a general method for integrating accident predictors into controllers with verifiable safety properties.

Future research could integrate risk predictors directly into the control loop. Planners and controllers can use risk maps, time to accident distributions, and safety envelopes as constraints or cost terms in model predictive control, safe RL, or trajectory optimization. This would make risk an explicit input for online decision making, rather than an after the fact indicator. Accident anticipation modules can be trained jointly with closed-loop control policies, or through multitask learning, so that safety related outcomes such as collisions, near misses, and hard braking events can guide both perception and decision layers. Bridging methods, including 3D interaction aware traffic accident anticipation,

planning methods that combine WMs with VLMs [260, 261], future occupancy and semantic prediction, and risk guided policy optimization, could be further connected with formal safety analysis. Tools such as control barrier functions [262], Hamilton-Jacobi reachability [263], and conformal decision theory [264] can support the assessment of whether a closed-loop policy respects predefined safety envelopes under bounded uncertainty.

4.4 Scenario Testing, Simulation Fidelity, and Standardization

Current safety validation pipelines overemphasize road mileage accumulation and a small set of manually crafted scenarios. This provides limited statistical power for rare safety related events. Although scenario-based testing and XiL have been adopted in several industrial practices and standards, many simulators and generated scenarios remain limited in three aspects. First, they often lack behavioral diversity and interaction realism. Second, they simplify sensing degradation, fault mechanisms, and communication failures [265]. Third, they use fragmented scenario representations and interfaces. These gaps make it difficult to translate simulation results into defensible safety claims for real world deployment, or to compare algorithms under consistent and reproducible conditions.

A useful step is to align critical scenario WMs and generative models with scenario-based testing standards. Generative models and traffic simulators could be trained and calibrated to reproduce empirical failure envelopes observed in incidents and near misses, including human driver reactions, perception degradation, sensor faults, and communication losses, rather than only nominal traffic statistics. Scenario description languages and test catalogues defined in standards such as ISO 34502 and ISO 34504 [266] could be extended to describe connected vehicle and multiagent critical scenarios. They could also be linked to scenario databases that manage data provenance, parameter ranges, and coverage metrics. Simulation interfaces and cosimulation with V2X network simulators could be used to evaluate accident anticipation, safety, comfort, and communication robustness using a shared set of performance indicators.

4.5 From Scenario Replay to Scenario Evidence Engineering

Most existing validation approaches still treat scenarios mainly as replayable test cases, rather than as traceable evidence for safety assurance. They often evaluate ego vehicle safety and efficiency in isolation, with limited consideration of the surrounding traffic system and human stakeholders. Risk predictions are also difficult to audit, especially in cooperative situations such as yielding for emergency corridors, interacting with vulnerable road users, or resolving occlusions with other agents. As a result, system level safety metrics that account for the risks of all agents, fairness, traffic flow, and human interpretability are rarely optimized in an explicit and verifiable way.

Future research could move from scenario replay to scenario evidence engineering. First, critical scenario WMs and generative models could be calibrated against empirical failure envelopes observed in incidents and near misses, rather than matching only nominal traffic statistics. This requires modeling human reactions, negotiation patterns among multiple agents, perception degradation, sensor faults, and communication impairments. It also requires validating whether generated scenarios can reproduce the boundary conditions under which failures occur. Second, standardization could move beyond format compliance toward evidence traceability, so that every test scenario, parameter range, simulation setting, and outcome can be reproduced, audited, and referenced in safety claims. Third, simulation fidelity could be assessed as a core validation variable, rather than being treated as an implicit assumption. A unified closed-loop evaluation stack could combine vehicle dynamics, sensing and perception pipelines, and V2X network cosimulation to quantify safety, comfort, efficiency, fairness, and communication robustness under shared performance indicators.

5 Conclusion

This survey identifies five main findings on critical scenario handling for autonomous driving. First, a major safety bottleneck for ADS is not limited to nominal perception or trajectory tracking, but the handling of rare, interactive, and ODD-boundary scenarios where perception errors, prediction uncertainty, and control limits appear together.

Second, recent work is shifting from geometric safety indicators to risk grounding, uncertainty-aware control, and WM/VLM-enabled planning. However, many methods still stop at open-loop risk prediction and lack control-consumable interfaces, such as explicit constraints, calibrated uncertainty, or safety shields.

Third, current datasets and benchmarks are not yet fully aligned with closed-loop safety assurance. Accident datasets provide useful evidence for risk perception and semantic understanding, but most of them remain limited in 3D structure, multimodal sensing, and closed-loop executability. Recent benchmarks such as Bench2Drive, NAVSIM v2, and WOD-E2E show that many difficult failures arise in interaction-heavy behaviors, recovery from ego-state deviation, and long-tail human-preferred decisions rather than simple lane following.

Fourth, scenario generation has become more scalable through optimization, adversarial learning, diffusion models, and VLMs, but the generated scenarios often lack verifiable coverage, physical consistency, and traceability to safety claims. Thus, future critical scenario generation should not only create rare cases but also explain which safety requirement, ODD boundary, or failure mode each case is designed to test.

Finally, no single testing method is sufficient for ADS safety assurance. SiL and HiL mainly support requirement and implementation verification, while ViL, proving-ground tests, and real-vehicle experiments provide stronger validation evidence for integrated behavior.

A practical safety assurance process therefore requires a scenario-to-evidence chain that connects critical scenario definition, executable parameterization, closed-loop testing, and certification-oriented evidence. A central conclusion of this survey is that safety-critical scenario research should move beyond isolated risk detection and scenario generation toward traceable closed-loop safety assurance supported by verification and validation evidence.

References

[1] Zihan Fang, Zheng Lin, Senkang Hu, Hangcheng Cao, Yiqin Deng, Xianhao Chen, and Yuguang Fang. 2024. IC3M: In-Car Multimodal Multi-Object Monitoring for Abnormal Status of Both Driver and Passengers. *arXiv preprint arXiv:2410.02592* (2024).

[2] Zheng Lin, Lifeng Wang, Jie Ding, Bo Tan, and Shi Jin. 2022. Channel Power Gain Estimation for Terahertz Vehicle-to-Infrastructure Networks. *IEEE Commun. Lett.* 27, 1 (2022), 155–159.

[3] Haoxuan Yuan, Zhe Chen, Zheng Lin, Jinbo Peng, Yuhang Zhong, Xuanjie Hu, Songyan Xue, Wei Li, and Yue Gao. 2025. Constructing 4D Radio Map in LEO Satellite Networks with Limited Samples. *IEEE INFOCOM* (2025).

[4] Tianyang Duan, Zongyuan Zhang, Zheng Lin, Songxiao Guo, Xiuxian Guan, Guangyu Wu, Zihan Fang, Haotian Meng, Xia Du, Ji-Zhe Zhou, et al. 2025. LLM-Driven Stationarity-Aware Expert Demonstrations for Multi-Agent Reinforcement Learning in Mobile Systems. *arXiv preprint arXiv:2511.19368* (2025).

[5] Mingda Hu, Jingjing Zhang, Xiong Wang, Shengyun Liu, and Zheng Lin. 2024. Accelerating Federated Learning with Model Segmentation for Edge Networks. *IEEE Trans. Green Commun. Netw.* (2024).

[6] Jinbo Peng, Junwen Duan, Zheng Lin, Haoxuan Yuan, Yue Gao, and Zhe Chen. 2025. SigChord: Sniffing Wide Non-Sparse Multiband Signals for Terrestrial and Non-Terrestrial Wireless Networks. *arXiv preprint arXiv:2504.06587* (2025).

[7] Zihan Fang, Zheng Lin, Senkang Hu, Yihang Tao, Yiqin Deng, Xianhao Chen, and Yuguang Fang. 2025. Dynamic uncertainty-aware multimodal fusion for outdoor health monitoring. *arXiv preprint arXiv:2508.09085* (2025).

[8] Wei Wei, Zheng Lin, Xihui Liu, Hongyang Du, Dusit Niyato, and Xianhao Chen. 2025. Optimizing Split Federated Learning with Unstable Client Participation. *arXiv preprint arXiv:2509.17398* (2025).

[9] Zekai Sun, Xiuxian Guan, Zheng Lin, Yuhao Qing, Haoze Song, Zihan Fang, Zhe Chen, Fangming Liu, Heming Cui, Wei Ni, et al. 2025. Rrto: A high-performance transparent offloading system for model inference in mobile edge computing. *arXiv preprint arXiv:2507.21739* (2025).

[10] Yuxin Zhang, Zheng Lin, Zhe Chen, Zihan Fang, Wenjun Zhu, Xianhao Chen, Jin Zhao, and Yue Gao. 2024. Satfed: A resource-efficient leo satellite-assisted heterogeneous federated learning framework. *Engineering* (2024).

[11] Zheng Lin, Zhe Chen, Xianhao Chen, Wei Ni, and Yue Gao. 2026. HASFL: Heterogeneity-aware Split Federated Learning over Edge Computing Systems. *IEEE Trans. Mobile Comput.* (2026).

[12] Junfei Zhan, Haoxun Shen, Zheng Lin, and Tengjiao He. 2025. PRISM: Privacy-Aware Routing for Adaptive Cloud-Edge LLM Inference via Semantic Sketch Collaboration. *arXiv preprint arXiv:2511.22788* (2025).

[13] Xiuxian Guan, Zongyuan Zhang, Zheng Lin, Zekai Sun, Tianyang Duan, Zihan Fang, Rui Wang, Heming Cui, Wei Ni, Jun Luo, et al. 2026. FluxShard: Motion-Aware Feature Cache Reuse for Collaborative Video Analytics in Mobile Edge Computing. *arXiv preprint arXiv:2605.06027* (2026).

[14] Zheng Lin, Yuxin Zhang, Zhe Chen, Zihan Fang, Cong Wu, Xianhao Chen, Yue Gao, and Jun Luo. 2025. LEO-Split: A Semi-Supervised Split Learning Framework over LEO Satellite Networks. *IEEE Trans. Mobile Comput.* (2025).

[15] Zekai Sun, Xiuxian Guan, Zheng Lin, Zihan Fang, Xiangming Cai, Zhe Chen, Fangming Liu, Heming Cui, Jie Xiong, Wei Ni, et al. 2025. Intra-DP: A High Performance Collaborative Inference System for Mobile Edge Computing. *arXiv preprint arXiv:2507.05829* (2025).

[16] Weiwei Zhuang, Wangze Xie, Qi Zhang, Xia Du, Zihan Lin, Zheng Lin, Hanlin Cai, Jizhe Zhou, Zihan Fang, Chi-man Pun, et al. 2026. Physically-Induced Atmospheric Adversarial Perturbations: Enhancing Transferability and Robustness in Remote Sensing Image Classification. *arXiv preprint arXiv:2604.14643* (2026).

[17] Mingwei Hong, Zheng Lin, Zehang Lin, Lin Li, Miao Yang, Xia Du, Zihan Fang, Zhaolu Kang, Dianxin Luan, and Shunzhi Zhu. 2026. Conflict-Aware Client Selection for Multi-Server Federated Learning. *arXiv preprint arXiv:2602.02458* (2026).

[18] Zheng Lin, Lifeng Wang, Jie Ding, Yuedong Xu, and Bo Tan. 2022. Tracking and transmission design in terahertz V2I networks. *IEEE Trans. Wireless Commun.* 22, 6 (2022), 3586–3598.

[19] Haoxuan Yuan, Zhe Chen, Zheng Lin, Jinbo Peng, Zihan Fang, Yuhang Zhong, Zihang Song, and Yue Gao. 2025. SatSense: Multi-Satellite Collaborative Framework for Spectrum Sensing. *IEEE Trans. Cogn. Commun. Netw.* (2025).

[20] Jinbo Peng, Zhe Chen, Zheng Lin, Haoxuan Yuan, Zihan Fang, Lingzhong Bao, Zihang Song, Ying Li, Jing Ren, and Yue Gao. 2024. SUMS: Sniffing Unknown Multiband Signals under Low Sampling Rates. *IEEE Trans. Mobile Comput.* (2024).

[21] Haoxuan Yuan, Zhe Chen, Zheng Lin, Jinbo Peng, Zihan Fang, Yuhang Zhong, Zihang Song, Xiong Wang, and Yue Gao. 2023. Graph Learning for Multi-Satellite Based Spectrum Sensing. In *Proc. IEEE Int. Conf. Commun. Technol. (ICCT)*. 1112–1116.

[22] Zhiyuan Zhao, Zhe Chen, Zheng Lin, Wenjun Zhu, Kun Qiu, Chaoqun You, and Yue Gao. 2024. LEO Satellite Networks Assisted Geo-Distributed Data Processing. *IEEE Wireless Commun. Lett.* (2024).

[23] Gueltoum Bendiab, Amina Hameurlaine, Georgios Germanos, Nicholas Kolokotronis, and Stavros Shiaeles. 2023. Autonomous vehicles security: Challenges and solutions using blockchain and artificial intelligence. *IEEE Transactions on Intelligent Transportation Systems* 24, 4 (2023), 3614–3637.

[24] Zihan Fang, Zheng Lin, Zhe Chen, Xianhao Chen, Yue Gao, and Yuguang Fang. 2025. Automated Federated Pipeline for Parameter-Efficient Fine-Tuning of Large Language Models. *IEEE Trans. Mobile Comput.* (2025).

[25] Zheng Lin, Guanqiao Qu, Wei Wei, Xianhao Chen, and Kin K Leung. 2025. Adaptsfl: Adaptive Split Federated Learning in Resource-Constrained Edge Networks. *IEEE Trans. Netw.* (2025).

[26] Mai Le, Thien Huynh-The, Tan Do-Duy, Thai-Hoc Vu, Won-Joo Hwang, and Quoc-Viet Pham. 2025. Applications of Distributed Machine Learning for the Internet-of-Things: A Comprehensive Survey. *IEEE Communications Surveys & Tutorials* 27, 2 (2025), 1053–1100.

[27] Zheng Lin, Yuxin Zhang, Zhe Chen, Zihan Fang, Xianhao Chen, Praneeth Vepakomma, Wei Ni, Jun Luo, and Yue Gao. 2025. HSplitLoRA: A Heterogeneous Split Parameter-Efficient Fine-Tuning Framework for Large Language Models. *arXiv preprint arXiv:2505.02795* (2025).

[28] Ons Aouedi, Thai-Hoc Vu, Alessio Sacco, Dinh C Nguyen, Kandaraj Piamrat, Guido Marchetto, and Quoc-Viet Pham. 2024. A survey on intelligent Internet of Things: Applications, security, privacy, and future directions. *IEEE communications surveys & tutorials* 27, 2 (2024), 1238–1292.

[29] Tianyue Zheng, Ang Li, Zhe Chen, Hongbo Wang, and Jun Luo. 2023. Autofed: Heterogeneity-aware federated multimodal learning for robust autonomous driving. In *Proceedings of the 29th annual international conference on mobile computing and networking*. 1–15.

[30] Jiayi Yi, Woojoo Kim, Dengbo He, and Chunxi Huang. 2026. Drivers' mental models of advanced driver assistance systems: A systematic review of conceptualization, associated factors, and intervention strategies. *Transportation Research Part F: Traffic Psychology and Behaviour* 118 (2026), 103529.

[31] Chen Sang, Sihan Gao, Xingwang Zhang, Haixin Zhang, Zhekang Dong, Yi Chen, and Junfan Wang. 2026. A Lightweight Dynamic Gesture Recognition Network for Advanced Driver Assistance Systems. *IEEE Internet of Things Journal* 13, 1 (2026), 787–800.

[32] Ziyu Song and Haitao Ding. 2023. Modeling car-following behavior in heterogeneous traffic mixing human-driven, automated and connected vehicles: considering multitype vehicle interactions. *Nonlinear dynamics* 111, 12 (2023), 11115–11134.

[33] Ruifo Zhang, Zhengyu Tan, Zemin Lin, Ruiying Zhang, and Chenhui Liu. 2025. Exploring the trust and behavior of experienced advanced driver assistance system drivers: An on-road study. *Accident Analysis & Prevention* 217 (2025), 108071.

[34] Guoqiang Li, Xudong Zhang, Hongliang Guo, Basilio Lenzo, and Ningyuan Guo. 2023. Real-time optimal trajectory planning for autonomous driving with collision avoidance using convex optimization. *Automotive Innovation* 6, 3 (2023), 481–491.

[35] Ziyu Song, Haitao Ding, Leila Jamel, Jing Yang, Muhammad Attique Khan, Juan M. Gorriz, Jamel Baili, and Lip Yee Por. 2025. Smart-City Spatiotemporal Data-Driven Trajectory Prediction for Autonomous Vehicles via Attention Mechanisms and Self-Supervised Learning. *IEEE Transactions on Consumer Electronics* 71, 4 (2025), 11834–11845.

[36] Yanwen Yang, Natnael M Negash, and James Yang. 2025. Recent advances in interactive driving of autonomous vehicles: Comprehensive review of approaches. *Automotive Innovation* 8, 2 (2025), 304–334.

[37] Gabriel Nativel-Fontaine, Véronique Lespinet-Najib, Robin Cazes, Camille Dupetit, Colin De Gasquet, Mathieu Chevrie, François Aïoun, and Luciano Ojeda. 2023. Exploration of the acceptability of different behaviors of an autonomous vehicle in so-called conflict situations. *Accident Analysis & Prevention* 186 (2023), 107041.

[38] Makoto Chikaraishi, Diana Khan, Banri Yasuda, and Akimasa Fujiwara. 2020. Risk perception and social acceptability of autonomous vehicles: A case study in Hiroshima, Japan. *Transport Policy* 98 (2020), 105–115.

[39] Wei Ji, Quan Yuan, Gang Cheng, Shengnan Yu, Min Wang, Zefang Shen, and Tiantong Yang. 2023. Traffic accidents of autonomous vehicles based on knowledge mapping: A review. *Journal of traffic and transportation engineering (English edition)* 10, 6 (2023), 1061–1073.

[40] Matthew Schwall, Tom Daniel, Trent Victor, Francesca Favaro, and Henning Hohnhold. 2020. Waymo public road safety performance data. *arXiv preprint arXiv:2011.00038* (2020).

[41] Zhige Chen, Zhigang Zhang, Qizheng Su, Kai Yang, Yandong Wu, Lei He, and Xiaolin Tang. 2025. Object detection for autonomous vehicles under adverse weather conditions. *Expert Systems with Applications* (2025), 128994.

[42] Jonas Mirlach, Lei Wan, Andreas Wiedholz, Hannan Ejaz Keen, and Andreas Eich. 2025. R-livit: A lidar-visual-thermal dataset enabling vulnerable road user focused roadside perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 28375–28384.

[43] Yaoye Zhu, Zhe Wang, and Yan Wang. 2025. MamV2XCalib: V2X-based Target-less Infrastructure Camera Calibration with State Space Model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 26696–26705.

[44] Luke Chen, Junyao Wang, Trier Mortlock, Pramod Khargonekar, and Mohammad Abdullah Al Faruque. 2025. Hyperdimensional uncertainty quantification for multimodal uncertainty fusion in autonomous vehicles perception. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 22306–22316.

[45] Xuesong Wang, Dingming Qin, Salvatore Cafiso, Kyle Kangzhi Liang, and Xiaolei Zhu. 2021. Operational design domain of autonomous vehicles at skewed intersection. *Accident Analysis & Prevention* 159 (2021), 106241.

[46] Taiwan News. 2020. Video shows Tesla on autopilot slam into truck on Taiwan highway. https://www.taiwannews.com.tw/news/3943199.

[47] David Shepardson. 2021. Police say Autopilot not believed in use in Detroit Tesla crash. https://jp.reuters.com/article/business/technology/police-say-autopilot-not-believed-in-use-in-detroit-tesla-crash-idUSKBN2B9070/.

[48] Reuters. 2021. U.S. safety agency opens probe of Tesla fatal crash in California. https://www.reuters.com/business/autos-transportation/us-safety-agency-opens-probe-tesla-fatal-crash-california-2021-05-12/.

[49] Reuters. 2024. Driverless Waymo car hits cyclist in San Francisco, causes minor scratches. https://www.reuters.com/world/us/driverless-waymo-car-hits-cyclist-san-francisco-causes-minor-scratches-2024-02-07/.

[50] Reuters. 2025. Xiaomi's EV in fatal accident, company says cooperating with police. https://jp.reuters.com/business/autos/JU6ZQKUDAVPNLOWLG25PFIN3NE-2025-04-01/.

[51] National Highway Traffic Safety Administration. 2025. *Part 573 Safety Recall Report 25E-029*. Technical Report. U.S. Department of Transportation. https://static.nhtsa.gov/odi/rcl/2025/RCLRPT-25E029-4731.PDF

[52] The Straits Times. 2025. Baidu robotaxi with passenger falls into construction pit in China, raising safety concerns. https://www.straitstimes.com/asia/east-asia/baidu-robotaxi-falls-into-construction-pit-in-china-raising-safety-concerns.

[53] Stefan Riedmaier, Thomas Ponn, Dieter Ludwig, Bernhard Schick, and Frank Diermeyer. 2020. Survey on scenario-based safety assessment of automated vehicles. *IEEE access* 8 (2020), 87456–87477.

[54] Xinhai Zhang, Jianbo Tao, Kaige Tan, Martin Törngren, José Manuel Gaspar Sánchez, Muhammad Rusyadi Ramli, Xin Tao, Magnus Gyllenhammar, Franz Wotawa, Naveen Mohan, et al. 2022. Finding critical scenarios for automated driving systems: A systematic mapping study. *IEEE Transactions on Software Engineering* 49, 3 (2022), 991–1026.

[55] Wenhao Ding, Chejian Xu, Mansur Arief, Haohong Lin, Bo Li, and Ding Zhao. 2023. A survey on safety-critical driving scenario generation—a methodological perspective. *IEEE Transactions on Intelligent Transportation Systems* 24, 7 (2023), 6971–6988.

[56] Jon Perez-Cerrolaza, Jaume Abella, Markus Borg, Carlo Donzella, Jesús Cerquides, Francisco J Cazorla, Cristofer Englund, Markus Tauber, George Nikolakopoulos, and Jose Luis Flores. 2024. Artificial intelligence for safety-critical systems in industrial and transportation domains: A survey. *Comput. Surveys* 56, 7 (2024), 1–40.

[57] Jingyuan Zhao, Yuyan Wu, Rui Deng, Susu Xu, Jinpeng Gao, and Andrew Burke. 2025. A survey of autonomous driving from a deep learning perspective. *Comput. Surveys* 57, 10 (2025), 1–60.

[58] Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. 2025. Understanding world or predicting future? a comprehensive survey of world models. *Comput. Surveys* 58, 3 (2025), 1–38.

[59] Saeid Nahavandi, Roohallah Alizadehsani, Darius Nahavandi, Shady Mohamed, Navid Mohajer, Mohammad Rokonuzzaman, and Ibrahim Hossain. 2025. A comprehensive review on autonomous navigation. *Comput. Surveys* 57, 9 (2025), 1–67.

[60] Wei Zhou, Li Yang, Lei Zhao, Runyu Zhang, Yifan Cui, Hongpu Huang, Kun Qie, and Chen Wang. 2025. Vision technologies with applications in traffic surveillance systems: A holistic survey. *Comput. Surveys* 58, 3 (2025), 1–47.

[61] Yuan Gao, Mattia Piccinini, Yuchen Zhang, Dingrui Wang, Korbinian Moller, Roberto Brusnicki, Baha Zarrouki, Alessio Gambi, Jan Frederik Totz, Kai Storms, et al. 2026. Foundation models in autonomous driving: A survey on scenario generation and scenario analysis. *IEEE Open Journal of Intelligent Transportation Systems* (2026).

[62] 2022. *ISO 21448:2022 — Road vehicles — Safety of the intended functionality.* International Standard ISO 21448:2022. International Organization for Standardization, Geneva, Switzerland. First edition.

[63] Erwin de Geldera, Jan-Pieter Paardekoopera, Arash Khabbaz Saberia, Hala Elrofaia, et al. 2020. Ontology for scenarios for the assessment of automated vehicles. *arXiv preprint arXiv:2001.11507* (2020).

[64] Eleonora Andreotti, Pinar Boyraz, and Selpi Selpi. 2020. Mathematical definitions of scene and scenario for analysis of automated driving systems in mixed-traffic simulations. *IEEE Transactions on Intelligent Vehicles* 6, 2 (2020), 366–375.

[65] 2022. *ISO 34502:2022 — Road vehicles — Test scenarios for automated driving systems — Scenario-based safety evaluation framework*. International Standard. International Organization for Standardization, Geneva, Switzerland. ISO 34502:2022(E).

[66] Sven Hallerbach, Yiqun Xia, Ulrich Eberle, and Frank Koester. 2018. Simulation-based identification of critical scenarios for cooperative and automated vehicles. *SAE International Journal of Connected and Automated Vehicles* 1, 2018-01-1066 (2018), 93–106.

[67] Qunying Song, Avner Bensoussan, and Mohammad Reza Mousavi. 2025. Synthetic versus real: an analysis of critical scenarios for autonomous vehicle testing. *Automated Software Engineering* 32, 2 (2025), 37.

[68] Degan Zhang, Wenjing Wang, Jie Zhang, Ting Zhang, Xuejie Ren, and Lu Chen. 2025. New Method of Vehicular Network Content Distribution Based on Edge Caching and Catch Fish Optimization Strategy. *IEEE Transactions on Reliability* 74, 4 (2025), 5190–5204.

[69] Shiyi Liang, Xinyuan Chang, Changjie Wu, Huiyuan Yan, Yifan Bai, Xinran Liu, Hang Zhang, Yujian Yuan, Shuang Zeng, Mu Xu, et al. 2026. Persistent autoregressive mapping with traffic rules for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 6862–6870.

[70] Henda Sfaxi, Imene Lahyani, Sami Yangui, and Mouna Torjmen. 2024. Latency-aware and proactive service placement for edge computing. *IEEE Transactions on Network and Service Management* 21, 4 (2024), 4243–4254.

[71] Sushank Chaudhary, Abhishek Sharma, Sunita Khichar, Yahui Meng, and Jyoteesh Malhotra. 2024. Enhancing autonomous vehicle navigation using SVM-based multi-target detection with photonic radar in complex traffic scenarios. *Scientific Reports* 14, 1 (2024), 17339.

[72] Jianfeng Wu, Xingyu Xing, Lu Xiong, and Junyi Chen. 2024. Accelerated testing and evaluation of autonomous vehicles based on dual surrogates. *Automotive Innovation* 7, 3 (2024), 390–402.

[73] Linda Pipkorn, Joshua Domeyer, Bruce Mehler, Bryan Reimer, and Pnina Gershon. 2025. Decoding the silent dialogue: Unveiling driver-pedestrian communication dynamics with a hidden Markov model. *Transportation Research Part F: Traffic Psychology and Behaviour* 109 (2025), 965–976.

[74] Shuning Tang, Yajie Zou, Shubo Wu, Yuanchang Xie, and Yunlong Zhang. 2025. Comparing Car-Following Behavior Patterns of Human-Driven Vehicles and Autonomous Vehicles in a Mixed Traffic Environment. *IEEE Transactions on Intelligent Transportation Systems* 26, 5 (2025), 6814–6830.

[75] Shengpeng Zhang and Taeoh Tak. 2024. Risk analysis of autonomous vehicle test scenarios using a novel analytic hierarchy process method. *IET Intelligent Transport Systems* 18, 5 (2024), 794–807.

[76] Kevin Tom Kurian, Nishant Rajesh, Erjen Lefeber, Jeroen Ploeg, Nathan van de Wouw, Igo Besselink, and Mohsen Alirezaei. 2025. How unsafe was the scenario? A criticality measure for scenario-based testing of automated vehicles. In *Proceedings of the 2025 IEEE Intelligent Vehicles Symposium (IV)*. 1905–1912.

[77] Yixuan Li, Xuesong Wang, Tianyi Wang, Lishengsa Yue, and Qian Liu. 2026. Characteristics analysis of autonomous vehicle pre-crash scenarios. *Accident Analysis & Prevention* 224 (2026), 108285.

[78] Yongqiang Lu, Hongjie Ma, Edward Smart, and Hui Yu. 2025. Enhancing Autonomous Driving Decision: A Hybrid Deep Reinforcement Learning-Kinematic-Based Autopilot Framework for Complex Motorway Scenes. *IEEE Transactions on Intelligent Transportation Systems* 26, 3 (2025), 3198–3209.

[79] Piotr Wzorek, Krzysztof Blachut, Kamil Jeziorek, and Tomasz Kryjak. 2025. Live Demonstration: Real-Time Event-Data Processing with Graph Convolutional Neural Networks and SoC FPGA. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 5103–5104.

[80] Enfu Huang, Zhanshan Zhao, Jiao Yin, Jinli Cao, and Hua Wang. 2026. Transformer-enhanced adaptive graph convolutional network for traffic flow prediction. *ACM Transactions on Intelligent Systems and Technology* 17, 3 (2026), 1–24.

[81] Qichao Liu, Heye Huang, Shiyue Zhao, Lei Shi, Soyoung Ahn, and Xiaopeng Li. 2026. RiskNet: interaction-aware risk forecasting for autonomous driving in long-tail scenarios. *Transportation Research Part E: Logistics and Transportation Review* 205 (2026), 104478.

[82] Kai Yang, Shen Li, Ming Wang, and Xiaolin Tang. 2025. Interactive decision-making integrating graph neural networks and model predictive control for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems* 26, 5 (2025), 6991–7005.

[83] Hung Duy Nguyen, Duc Thinh Le, Tung Lam Nguyen, and Minh Nhat Vu. 2025. Robust Model Predictive Control-Based Recurrent Neural Networks for Autonomous Vehicles in Avoidance Collisions. *IEEE Access* 13 (2025), 106115–106128.

[84] Hao Chen, Chongfeng Wei, Yinhua Liu, Chuan Hu, and Xi Zhang. 2024. Vulnerable traffic participant trajectory prediction based on gate recurrent unit-attention and ameliorative social force model. *IEEE Transactions on Transportation Electrification* 10, 4 (2024), 9396–9405.

[85] Ping Lu, Hui Xu, and Bo Hu. 2025. A Transformer Optimized Planner for Autonomous Vehicle On-Ramping Merging Task. *IEEE Transactions on Industrial Electronics* 72, 12 (2025), 14437–14447.

[86] Pincan Zhao, Changle Li, Xinrui Zhang, F Richard Yu, and Yuchuan Fu. 2025. Intelligent cooperative sensing for connected and autonomous vehicles: An improved decision transformer approach. *IEEE Internet of Things Journal* 12, 11 (2025), 15424–15437.

[87] Jianwu Fang, Jiahuan Qiao, Jianru Xue, and Zhengguo Li. 2023. Vision-based traffic accident detection and anticipation: A survey. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 4 (2023), 1983–1999.

[88] Haicheng Liao, Bin Rao, Haoyu Sun, Chengyue Wang, Qing Chang, Shengbo Eben Li, Chengzhong Xu, and Zhenning Li. 2025. Chain-of-Thought Guided Multimodal Large Language Models for Scene-Aware Accident Anticipation in Autonomous Driving. *IEEE Transactions on Intelligent Transportation Systems* 26, 11 (2025), 19371–19380.

[89] Haicheng Liao, Yongkang Li, Zhenning Li, Zilin Bian, Jaeyoung Lee, Zhiyong Cui, Guohui Zhang, and Chengzhong Xu. 2024. Real-time accident anticipation for autonomous driving through monocular depth-enhanced 3D modeling. *Accident Analysis & Prevention* 207 (2024), 107760.

[90] Jianwu Fang, Lei-Lei Li, Zhedong Zheng, Hongkai Yu, Jianru Xue, Zhengguo Li, and Tat-Seng Chua. 2025. EQ-TAA: Equivariant Traffic Accident Anticipation via Diffusion-Based Accident Video Synthesis. *arXiv preprint arXiv:2506.10002* (2025).

[91] Haicheng Liao, Yongkang Li, Chengyue Wang, Yanchen Guan, Kahou Tam, Chunlin Tian, Li Li, Chengzhong Xu, and Zhenning Li. 2024. When, where, and what? A benchmark for accident anticipation and localization with large language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 8–17.

[92] Jingyu Li, Bozhou Zhang, Xin Jin, Jiankang Deng, Xiatian Zhu, and Li Zhang. 2025. ImagiDrive: A Unified Imagination-and-Planning Framework for Autonomous Driving. *arXiv preprint arXiv:2508.11428* (2025).

[93] Zihao Sheng, Zilin Huang, Yen-Jung Chen, Yansong Qu, Yuhao Luo, Yue Leng, and Sikai Chen. 2025. SafePLUG: Empowering Multimodal LLMs with Pixel-Level Insight and Temporal Grounding for Traffic Accident Understanding. *arXiv preprint arXiv:2508.06763* (2025).

[94] Xingcheng Liu, Bin Rao, Yanchen Guan, Chengyue Wang, Haicheng Liao, Jiaxun Zhang, Chengyu Lin, Meixin Zhu, and Zhenning Li. 2025. Predict and Resist: Long-Term Accident Anticipation under Sensor Noise. *arXiv preprint arXiv:2511.08640* (2025).

[95] Yiran Zhang, Zhongxu Hu, Haohan Yang, Shanhe Lou, and Chen Lv. 2025. A Planner-Agnostic Monitor for Behaviour Feasibility of Autonomous Vehicles Using a Bayesian Discriminator. *IEEE Transactions on Intelligent Transportation Systems* 26, 10 (2025), 17096–17109.

[96] H. Fu, D. Zhang, Z. Zhao, J. Cui, D. Liang, C. Zhang, D. Zhang, H. Xie, B. Wang, and X. Bai. 2025. Orion: A Holistic End-to-End Autonomous Driving Framework by Vision-Language Instructed Action Generation. *arXiv preprint arXiv:2503.19755* (2025).

[97] Y. Li, K. Xiong, X. Guo, F. Li, S. Yan, G. Xu, L. Zhou, L. Chen, H. Sun, B. Wang, et al. 2025. RecogDrive: A Reinforced Cognitive Framework for End-to-End Autonomous Driving. *arXiv preprint arXiv:2506.08052* (2025).

[98] X. Tian, J. Gu, B. Li, Y. Liu, Y. Wang, Z. Zhao, K. Zhan, P. Jia, X. Lang, and H. Zhao. 2025. DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models. In *Conference on Robot Learning*. PMLR, 4698–4726.

[99] B. Jiang, S. Chen, B. Liao, X. Zhang, W. Yin, Q. Zhang, C. Huang, W. Liu, and X. Wang. 2024. Senna: Bridging Large Vision-Language Models and End-to-End Autonomous Driving. *arXiv preprint arXiv:2410.22313* (2024).

[100] Wenhui Huang, Songyan Zhang, Qihang Huang, Zhidong Wang, Zhiqi Mao, Collister Chua, Zhan Chen, Long Chen, and Chen Lv. 2026. AutoMoT: A Unified Vision-Language-Action Model with Asynchronous Mixture-of-Transformers for End-to-End Autonomous Driving. *arXiv preprint arXiv:2603.14851* (2026).

[101] S. Zhang, W. Huang, Z. Chen, C. J. Collister, Q. Huang, and C. Lv. 2025. OpenREAD: Reinforced Open-Ended Reasoning for End-to-End Autonomous Driving with LLM-as-Critic. *arXiv preprint arXiv:2512.01830* (2025).

[102] S. Zhang, W. Huang, Z. Gao, H. Chen, and C. Lv. 2024. WiseAD: Knowledge Augmented End-to-End Autonomous Driving with Vision-Language Model. *arXiv preprint arXiv:2412.09951* (2024).

[103] Z. Zhou, T. Cai, S. Z. Zhao, Y. Zhang, Z. Huang, B. Zhou, and J. Ma. 2025. AutoVLA: A Vision-Language-Action Model for End-to-End Autonomous Driving with Adaptive Reasoning and Reinforcement Fine-Tuning. *arXiv preprint arXiv:2506.13757* (2025).

[104] K. Renz, L. Chen, E. Arani, and O. Sinavski. 2025. SimLingo: Vision-Only Closed-Loop Autonomous Driving with Language-Action Alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 11993–12003.

[105] Y. Wang, W. Luo, J. Bai, Y. Cao, T. Che, K. Chen, Y. Chen, J. Diamond, Y. Ding, W. Ding, et al. 2025. Alpamayo-R1: Bridging Reasoning and Action Prediction for Generalizable Autonomous Driving in the Long Tail. *arXiv preprint arXiv:2511.00088* (2025).

[106] X. Zhou, X. Han, F. Yang, Y. Ma, V. Tresp, and A. Knoll. 2025. OpenDriveVLA: Towards End-to-End Autonomous Driving with Large Vision Language Action Model. *arXiv preprint arXiv:2503.23463* (2025).

[107] Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, and Xing Wei. 2025. FutureSightDrive: Thinking Visually with Spatio-Temporal CoT for Autonomous Driving. *arXiv preprint arXiv:2505.17685* (2025).

[108] Zhida Zhao, Talas Fu, Yifan Wang, Lijun Wang, and Huchuan Lu. 2025. From Forecasting to Planning: Policy World Model for Collaborative State-Action Prediction. *arXiv preprint arXiv:2510.19654* (2025).

[109] Yingyan Li, Shuyao Shang, Weisong Liu, Bing Zhan, Haochen Wang, Yuqi Wang, Yuntao Chen, Xiaoman Wang, Yasong An, Chufeng Tang, et al. 2025. DriveVLA-W0: World Models Amplify Data Scaling Law in Autonomous Driving. *arXiv preprint arXiv:2510.12796* (2025).

[110] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv preprint arXiv:2406.09418* (2024).

[111] Weiming Hu, Xuejuan Xiao, Dan Xie, and Tieniu Tan. 2003. Traffic accident prediction using vehicle tracking and trajectory analysis. In *Proceedings of the 2003 IEEE International Conference on Intelligent Transportation Systems*, Vol. 1. 220–225.

[112] Zhenyu Shan, Qianqian Zhu, and Danna Zhao. 2017. Vehicle collision risk estimation based on RGB-D camera for urban road. *Multimedia systems* 23, 1 (2017), 119–127.

[113] Fu-Hsiang Chan, Yu-Ting Chen, Yu Xiang, and Min Sun. 2016. Anticipating accidents in dashcam videos. In *Asian conference on computer vision*. 136–153.

[114] Wentao Bao, Qi Yu, and Yu Kong. 2020. Uncertainty-based traffic accident anticipation with spatio-temporal relational learning. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2682–2690.

[115] Hoon Kim, Kangwook Lee, Gyeongjo Hwang, and Changho Suh. 2021. Predicting vehicle collisions using data collected from video games. *Machine Vision and Applications* 32, 4 (2021), 93.

[116] Wentao Bao, Qi Yu, and Yu Kong. 2021. Drive: Deep reinforced accident anticipation with visual explanation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7619–7628.

[117] Muhammad Monjurul Karim, Yu Li, and Ruwen Qin. 2022. Toward explainable artificial intelligence for early anticipation of traffic accidents. *Transportation research record* 2676, 6 (2022), 743–755.

[118] Deesha Chavan, Dev Saad, and Debarati B Chakraborty. 2021. COLLIDE-PRED: prediction of on-road collision from surveillance videos. *arXiv preprint arXiv:2101.08463* (2021).

[119] Arnav Vaibhav Malawade, Shih-Yuan Yu, Brandon Hsu, Deepan Muthirayan, Pramod P Khargonekar, and Mohammad Abdullah Al Faruque. 2022. Spatiotemporal scene-graph embedding for autonomous vehicle collision prediction. *IEEE Internet of Things Journal* 9, 12 (2022), 9379–9388.

[120] Muhammad Monjurul Karim, Yu Li, Ruwen Qin, and Zhaozheng Yin. 2022. A dynamic spatial-temporal attention network for early anticipation of traffic accidents. *IEEE Transactions on Intelligent Transportation Systems* 23, 7 (2022), 9590–9600.

[121] Muhammad Monjurul Karim, Zhaozheng Yin, and Ruwen Qin. 2023. An attention-guided multistream feature fusion network for early localization of risky traffic agents in driving videos. *IEEE Transactions on Intelligent Vehicles* 9, 1 (2023), 1792–1803.

[122] Tianhang Wang, Kai Chen, Guang Chen, Bin Li, Zhijun Li, Zhengfa Liu, and Changjun Jiang. 2023. GSC: A graph and spatio-temporal continuity based framework for accident anticipation. *IEEE Transactions on Intelligent Vehicles* 9, 1 (2023), 2249–2261.

[123] Jianwu Fang, Lei-Lei Li, Kuan Yang, Zhedong Zheng, Jianru Xue, and Tat-Seng Chua. 2022. Cognitive accident prediction in driving scenes: A multimodality benchmark. *arXiv preprint arXiv:2212.09381* (2022).

[124] Yuntao Chen, Yuqi Wang, and Zhaoxiang Zhang. 2025. Drivinggpt: Unifying driving world modeling and planning with multi-modal autoregressive transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 26890–26900.

[125] Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiao-Xiao Long, et al. 2025. Epona: Autoregressive diffusion world model for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 27220–27230.

[126] Xingtai Gui, Meijie Zhang, Tianyi Yan, Wencheng Han, Jiahao Gong, Feiyang Tan, Cheng-zhong Xu, and Jianbing Shen. 2026. Bridging scene generation and planning: Driving with world model via unifying vision and motion representation. *arXiv preprint arXiv:2603.14948* (2026).

[127] Zhexiao Xiong, Xin Ye, Burhan Yaman, Sheng Cheng, Yiren Lu, Jingru Luo, Nathan Jacobs, and Liu Ren. 2026. UniDrive-WM: Unified Understanding, Planning and Generation World Model For Autonomous Driving. *arXiv preprint arXiv:2601.04453* (2026).

[128] Shuyao Shang, Bing Zhan, Yunfei Yan, Yuqi Wang, Yingyan Li, Yasong An, Xiaoman Wang, Jierui Liu, Lu Hou, Lue Fan, et al. 2026. DynVLA: Learning World Dynamics for Action Reasoning in Autonomous Driving. *arXiv preprint arXiv:2603.11041* (2026).

[129] Hui Lv, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. 2021. Localizing anomalies from weakly-labeled videos. *IEEE transactions on image processing* 30 (2021), 4505–4515.

[130] Yu Yao, Mingze Xu, Yuchen Wang, David J Crandall, and Ella M Atkins. 2019. Unsupervised traffic accident detection in first-person videos. In *2019 IEEE/RSJ International conference on intelligent robots and systems (IROS)*. IEEE, 273–280.

[131] Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Yuchen Wang, Ella Atkins, and David J Crandall. 2022. DoTA: Unsupervised detection of traffic anomaly in driving videos. *IEEE transactions on pattern analysis and machine intelligence* 45, 1 (2022), 444–459.

[132] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, and Hongkai Yu. 2021. DADA: Driver attention prediction in driving accident scenarios. *IEEE transactions on intelligent transportation systems* 23, 6 (2021), 4959–4971.

[133] Jiaming Zhang, Kailun Yang, and Rainer Stiefelhagen. 2021. Exploring event-driven dynamic context for accident scene segmentation. *IEEE Transactions on Intelligent Transportation Systems* 23, 3 (2021), 2606–2622.

[134] Hoon Kim, Kangwook Lee, Gyeongjo Hwang, and Changho Suh. 2019. Crash to not crash: Learn to identify dangerous vehicles using a simulator. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 978–985.

[135] Tianqi Wang, Sukmin Kim, Ji Wenxuan, Enze Xie, Chongjian Ge, Junsong Chen, Zhenguo Li, and Ping Luo. 2024. Deepaccident: A motion and accident prediction benchmark for v2x autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 5599–5606.

[136] Li Xu, He Huang, and Jun Liu. 2021. SUTD-TrafficQA: A question answering benchmark and an efficient network for video reasoning over traffic events. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9878–9888.

[137] Jianwu Fang, Lei-lei Li, Junfei Zhou, Junbin Xiao, Hongkai Yu, Chen Lv, Jianru Xue, and Tat-Seng Chua. 2024. Abductive ego-view accident video understanding for safe driving perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22030–22040.

[138] Yixuan Zhou, Long Bai, Sijia Cai, Bing Deng, Xing Xu, and Heng Tao Shen. 2025. Tau-106k: A new dataset for comprehensive understanding of traffic accident. In *The Thirteenth International Conference on Learning Representations*.

[139] Chirag Parikh, Deepti Rawat, Tathagata Ghosh, Ravi Kiran Sarvadevabhatla, et al. 2025. RoadSocial: A Diverse VideoQA Dataset and Benchmark for Road Event Understanding from Social Video Narratives. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 19002–19011.

[140] Zhenghao Xing, Hao Chen, Binzhu Xie, Jiaqi Xu, Ziyu Guo, Xuemiao Xu, Jianye Hao, Chi-Wing Fu, Xiaowei Hu, and Pheng-Ann Heng. 2025. EchoTraffic: Enhancing traffic anomaly understanding with audio-visual insights. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 19098–19108.

[141] Yongjun Zhang, Pengcheng Shi, and Jiayuan Li. 2024. Lidar-based place recognition for autonomous driving: A survey. *Comput. Surveys* 57, 4 (2024), 1–36.

[142] Ankit Parag Shah, Jean-Bapstite Lamare, Tuan Nguyen-Anh, and Alexander Hauptmann. 2018. CADP: A novel dataset for CCTV traffic camera based accident analysis. In *Proceedings of the 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 1–9.

[143] Hirokatsu Kataoka, Teppei Suzuki, Shoko Oikawa, Yasuhiro Matsui, and Yutaka Satoh. 2018. Drive video analysis for the detection of traffic near-miss incidents. In *2018 IEEE International Conference on robotics and automation (ICRA)*. 3421–3428.

[144] Tomoyuki Suzuki, Hirokatsu Kataoka, Yoshimitsu Aoki, and Yutaka Satoh. 2018. Anticipating traffic accidents with adaptive loss and large-scale incident db. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3521–3529.

[145] Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Basura Fernando, Lars Petersson, and Lars Andersson. 2018. VIENA: A driving anticipation dataset. In *Asian Conference on Computer Vision*. 449–466.

[146] Rixing Zhu, Jianwu Fang, Hongke Xu, and Jianru Xue. 2019. Progressive temporal-spatial-semantic analysis of driving anomaly detection and recounting. *Sensors* 19, 23 (2019), 5098.

[147] Sanjay Haresh, Sateesh Kumar, M Zeeshan Zia, and Quoc-Huy Tran. 2020. Towards anomaly detection in dashcam videos. In *Proceedings of the 2020 IEEE Intelligent Vehicles Symposium (IV)*. 1407–1414.

[148] Trung-Nghia Le, Shintaro Ono, Akihiro Sugimoto, and Hiroshi Kawasaki. 2020. Attention R-CNN for accident detection. In *Proceedings of the 2020 IEEE intelligent vehicles symposium (IV)*. 313–320.

[149] Tackgeun You and Bohyung Han. 2020. Traffic accident benchmark for causality recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 540–556.

[150] Thakare Kamalakar Vijay, Debi Prosad Dogra, Heeseung Choi, Gipyo Nam, and Ig-Jae Kim. 2022. Detection of road accidents using synthetically generated multi-perspective accident videos. *IEEE Transactions on Intelligent Transportation Systems* 24, 2 (2022), 1926–1935.

[151] Chunsheng Liu, Zijian Li, Faliang Chang, Shuang Li, and Jincan Xie. 2021. Temporal shift and spatial attention-based two-stream network for traffic risk assessment. *IEEE Transactions on Intelligent Transportation Systems* 23, 8 (2021), 12518–12530.

[152] Yajun Xu, Huan Hu, Chuwen Huang, Yibing Nan, Yuyao Liu, Kai Wang, Zhaoxiang Liu, and Shiguo Lian. 2025. TAD: A Large-Scale Benchmark for Traffic Accidents Detection From Video Surveillance. *IEEE Access* 13 (2025), 2018–2033.

[153] Haohan Luo and Feng Wang. 2023. A simulation-based framework for urban traffic accident detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5.

[154] Daniel Moura, Shizhan Zhu, and Orly Zvitia. 2025. Nexar Dashcam Collision Prediction Dataset and Challenge. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2583–2591.

[155] Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. 2024. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)* 37 (2024), 819–844.

[156] Wei Cao, Marcel Hallgarten, Tianyu Li, Daniel Dauner, Xunjiang Gu, Caojun Wang, Yakov Miron, Marco Aiello, Hongyang Li, Igor Gilitschenski, et al. 2025. Pseudo-simulation for autonomous driving. *arXiv preprint arXiv:2506.04218* (2025).

[157] Runsheng Xu, Hubert Lin, Wonseok Jeon, Hao Feng, Yuliang Zou, Liting Sun, John Gorman, Kate Tolstaya, Sarah Tang, Brandyn White, et al. 2025. WOD-E2E: Waymo Open Dataset for End-to-End Driving in Challenging Long-tail Scenarios. *arXiv preprint arXiv:2510.26125* (2025).

[158] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An open urban driving simulator. In *Proceedings of the Conference on Robot Learning*. 1–16.

[159] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. 2021. nuPlan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810* (2021).

[160] Penghao Wu, Xiaosong Jia, Li Chen, Junchi Yan, Hongyang Li, and Yu Qiao. 2022. Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. *Advances in Neural Information Processing Systems* 35 (2022), 6119–6132.

[161] Xiaosong Jia, Penghao Wu, Li Chen, Jiangwei Xie, Conghui He, Junchi Yan, and Hongyang Li. 2023. Think twice before driving: Towards scalable decoders for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21983–21994.

[162] Xiaosong Jia, Yulu Gao, Li Chen, Junchi Yan, Patrick Langechuan Liu, and Hongyang Li. 2023. Driveadapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7953–7963.

[163] Jiang-Tian Zhai, Ze Feng, Jihao Du, Yongqiang Mao, Jiang-Jiang Liu, Zichang Tan, Yifu Zhang, Xiaoqing Ye, and Jingdong Wang. 2023. Rethinking the Open-Loop Evaluation of End-to-End Autonomous Driving in nuScenes. *arXiv preprint arXiv:2305.10430* (2023).

[164] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. 2023. Planning-oriented Autonomous Driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 17853–17862.

[165] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. 2023. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8340–8350.

[166] Xiaosong Jia, Junqi You, Zhiyuan Zhang, and Junchi Yan. 2025. DriveTransformer: Unified Transformer for Scalable End-to-End Autonomous Driving. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[167] Zhenxin Li, Shihao Wang, Shiyi Lan, Zhiding Yu, Zuxuan Wu, and Jose M Alvarez. 2025. Hydra-next: Robust closed-loop driving with open-loop training. *arXiv preprint arXiv:2503.12030* (2025).

[168] Julian Zimmerlin, Jens Beißwenger, Bernhard Jaeger, Andreas Geiger, and Kashyap Chitta. 2024. Hidden biases of end-to-end driving datasets. *arXiv preprint arXiv:2412.09602* (2024).

[169] Haochen Liu, Tianyu Li, Haohan Yang, Li Chen, Caojun Wang, Ke Guo, Haochen Tian, Hongchen Li, Hongyang Li, and Chen Lv. 2025. Reinforced Refinement with Self-Aware Expansion for End-to-End Autonomous Driving. *arXiv preprint arXiv:2506.09800* (2025).

[170] Yingqi Tang, Zhuoran Xu, Zhaotie Meng, and Erkang Cheng. 2025. Hip-ad: Hierarchical and multi-granularity planning with deformable attention for autonomous driving in a single decoder. *arXiv preprint arXiv:2503.08612* (2025).

[171] Liuhan Yin, Runkun Ju, Guodong Guo, and Erkang Cheng. 2025. DiffRefiner: Coarse to Fine Trajectory Planning via Diffusion Refinement with Semantic Interaction for End to End Autonomous Driving. *arXiv preprint arXiv:2511.17150* (2025).

[172] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. 2021. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7077–7087.

[173] Zhenxin Li, Kailin Li, Shihao Wang, Shiyi Lan, Zhiding Yu, Yishen Ji, Zhiqi Li, Ziyue Zhu, Jan Kautz, Zuxuan Wu, et al. 2024. Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation. *arXiv preprint arXiv:2406.06978* (2024).

[174] Wenhao Yao, Zhenxin Li, Shiyi Lan, Zi Wang, Xinglong Sun, Jose M Alvarez, and Zuxuan Wu. 2025. DriveSuprim: Towards Precise Trajectory Selection for End-to-End Planning. *arXiv preprint arXiv:2506.06659* (2025).

[175] Renju Feng, Ning Xi, Duanfeng Chu, Rukang Wang, Zejian Deng, Anzheng Wang, Liping Lu, Jinxiang Wang, and Yanjun Huang. 2025. Artemis: Autoregressive end-to-end trajectory planning with mixture of experts for autonomous driving. *arXiv preprint arXiv:2504.19580* (2025).

[176] Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, et al. 2025. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 12037–12047.

[177] Simon Ulbrich, Till Menzel, Andreas Reschka, Fabian Schuldt, and Markus Maurer. 2015. Defining and substantiating the terms scene, situation, and scenario for automated driving. In *Proceedings of the 2015 IEEE 18th international conference on intelligent transportation systems*. 982–988.

[178] Till Menzel, Gerrit Bagschik, and Markus Maurer. 2018. Scenarios for Development, Test and Validation of Automated Vehicles. In *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 1821–1827.

[179] Gerrit Bagschik, Till Menzel, and Markus Maurer. 2018. Ontology based scene creation for the development of automated vehicles. In *Proceedings of the 2018 IEEE intelligent vehicles symposium (IV)*. 1813–1820.

[180] Julian Bock, R Krajewski, L Eckstein, J Klimke, J Sauerbier, and A Zlocki. 2018. Data basis for scenario-based validation of HAD on highways. In *27th Aachen colloquium automobile and engine technology*. 8–10.

[181] Hermann Winner, Karsten Lemmer, Thomas Form, and Jens Mazzega. 2018. PEGASUS—First steps for the safe introduction of automated driving. In *Road Vehicle Automation 5*. 185–195.

[182] Elias Rocklage, Heiko Kraft, Abdullah Karatas, and Jörg Seewig. 2017. Automated scenario generation for regression testing of autonomous vehicles. In *Proceedings of the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. 476–483.

[183] Jianli Duan, Feng Gao, and Yingdong He. 2020. Test scenario generation and optimization technology for intelligent driving systems. *IEEE Intelligent Transportation Systems Magazine* 14, 1 (2020), 115–127.

[184] Florian Klück, Yihao Li, Mihai Nica, Jianbo Tao, and Franz Wotawa. 2018. Using ontologies for test suites generation for automated and autonomous driving functions. In *Proceedings of the 2018 IEEE International symposium on software reliability engineering workshops (ISSREW)*. 118–123.

[185] Cumhur Erkan Tuncali, Theodore P Pavlic, and Georgios Fainekos. 2016. Utilizing S-TaLiRo as an automatic test generation framework for autonomous vehicles. In *Proceedings of the 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*. 1470–1475.

[186] Halil Beglerovic, Michael Stolz, and Martin Horn. 2017. Testing of autonomous vehicles using surrogate models and stochastic optimization. In *Proceedings of the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. 1–6.

[187] Peter Du and Katherine Driggs-Campbell. 2019. Finding diverse failure scenarios in autonomous systems using adaptive stress testing. *SAE International Journal of Connected and Automated Vehicles* 2, 12-02-04-0018 (2019), 241–251.

[188] Moritz Klischat, Edmond Irani Liu, Fabian Holtke, and Matthias Althoff. 2020. Scenario factory: Creating safety-critical traffic scenarios for automated vehicles. In *Proceedings of the 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. 1–7.

[189] Hala Elrofai, Daniël Worm, and Olaf Op den Camp. 2016. Scenario identification for validation of automated driving functions. In *Advanced Microsystems for Automotive Applications 2016: Smart Systems for the Automobile of the Future*. 153–163.

[190] Robert Krajewski, Tobias Moers, Dominik Nerger, and Lutz Eckstein. 2018. Data-driven maneuver modeling using generative adversarial networks and variational autoencoders for safety validation of highly automated vehicles. In *Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2383–2390.

[191] Zhiyuan Wei, Helai Huang, Guoqing Zhang, Rui Zhou, Xiaolong Luo, Shiqi Li, and Hanchu Zhou. 2025. Interactive Critical Scenario Generation for Autonomous Vehicles Testing Based on In-Depth Crash Data Using Reinforcement Learning. *IEEE Transactions on Intelligent Vehicles* 10, 3 (2025), 1471–1482.

[192] Mingxing Peng, Kehua Chen, Xusen Guo, Qiming Zhang, Hui Zhong, Meixin Zhu, and Hai Yang. 2025. Diffusion Models for Intelligent Transportation Systems: A Survey. *IEEE Transactions on Intelligent Transportation Systems* 26, 12 (2025), 21526–21543.

[193] Haohong Lin, Xin Huang, Tung Phan, David Hayden, Huan Zhang, Ding Zhao, Siddhartha Srinivasa, Eric Wolff, and Hongge Chen. 2025. Causal composition diffusion model for closed-loop traffic generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 27542–27552.

[194] Jiacheng Chen, Ziyu Jiang, Mingfu Liang, Bingbing Zhuang, Jong-Chyi Su, Sparsh Garg, Ying Wu, and Manmohan Chandraker. 2025. AutoScape: Geometry-Consistent Long-Horizon Scene Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 25700–25711.

[195] Jongsuk Kim, Jaeyoung Lee, Gyojin Han, Dong-Jae Lee, Minki Jeong, and Junmo Kim. 2025. SynAD: Enhancing Real-World End-to-End Autonomous Driving Models through Synthetic Data Integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 25197–25206.

[196] Rui Zhou, Zhiyuan Wei, Jiang Bian, Qianyuan Yu, Shan Tian, Helai Huang, Gui Gui, and Lu Xing. 2025. MARL-Based High-Risk Multivehicle Scenario Generation for Autonomous Vehicle Safety Testing. *IEEE Internet of Things Journal* 12, 22 (2025), 46928–46940.

[197] Lulu Jia, Dezhen Yang, Yi Ren, Cheng Qian, Qiang Feng, Bo Sun, and Zili Wang. 2024. A dynamic test scenario generation method for autonomous vehicles based on conditional generative adversarial imitation learning. *Accident Analysis & Prevention* 194 (2024), 107279.

[198] Haohong Lin, Xin Huang, Tung Phan, David Hayden, Huan Zhang, Ding Zhao, Siddhartha Srinivasa, Eric Wolff, and Hongge Chen. 2025. Causal composition diffusion model for closed-loop traffic generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 27542–27552.

[199] Zihao Li, Xinyuan Cao, Xiangbo Gao, Kexin Tian, Keshu Wu, Mohammad Anis, Hao Zhang, Keke Long, Jiwan Jiang, Xiaopeng Li, et al. 2025. Simulating the Unseen: Crash Prediction Must Learn from What Did Not Happen. *arXiv preprint arXiv:2505.21743* (2025).

[200] Yuewen Mei, Tong Nie, Jian Sun, and Ye Tian. 2025. Llm-attacker: Enhancing closed-loop adversarial scenario generation for autonomous driving with large language models. *arXiv preprint arXiv:2501.15850* (2025).

[201] Mingxing Peng, Yuting Xie, Xusen Guo, Ruoyu Yao, Hai Yang, and Jun Ma. 2025. Ld-scene: Llm-guided diffusion for controllable generation of adversarial safety-critical driving scenarios. *arXiv preprint arXiv:2505.11247* (2025).

[202] Yuewen Mei, Tong Nie, Jian Sun, and Ye Tian. 2025. Seeking to collide: Online safety-critical scenario generation for autonomous driving with retrieval augmented large language models. *arXiv preprint arXiv:2505.00972* (2025).

[203] Tong Nie, Jian Sun, and Wei Ma. 2025. Exploring the roles of large language models in reshaping transportation systems: A survey, framework, and roadmap. *Artificial Intelligence for Transportation* 1 (2025), 100003.

[204] Jiangfan Liu, Yongkang Guo, Fangzhi Zhong, Tianyuan Zhang, Zonglei Jing, Siyuan Liang, Jiakai Wang, Mingchuan Zhang, Aishan Liu, and Xianglong Liu. 2025. Adversarial Generation and Collaborative Evolution of Safety-Critical Scenarios for Autonomous Vehicles. *arXiv preprint arXiv:2508.14527* (2025).

[205] Zihao Sheng, Zilin Huang, Yansong Qu, Yue Leng, Sruthi Bhavanam, and Sikai Chen. 2025. CurricuVLM: Towards Safe Autonomous Driving via Personalized Safety-Critical Curriculum Learning with Vision-Language Models. *arXiv preprint arXiv:2502.15119* (2025).

[206] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2446–2454.

[207] Quanyi Li, Zhenghao Peng, Lan Feng, Qihang Zhang, Zhenghai Xue, and Bolei Zhou. 2022. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence* 45, 3 (2022), 3461–3475.

[208] Qiujing Lu, Meng Ma, Ximiao Dai, Xuanhan Wang, and Shuo Feng. 2024. Realistic corner case generation for autonomous vehicles with multimodal large language model. *arXiv preprint arXiv:2412.00243* (2024).

[209] Daniel Krajzewicz, Jakob Erdmann, Michael Behrisch, Laura Bieker, et al. 2012. Recent development and applications of SUMO-Simulation of Urban MObility. *International journal on advances in systems and measurements* 5, 3&4 (2012), 128–138.

[210] Erfan Aasi, Phat Nguyen, Shiva Sreeram, Guy Rosman, Sertac Karaman, and Daniela Rus. 2025. Generating out-of-distribution scenarios using language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*. 10616–10623.

[211] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11621–11631.

[212] Shenyu Zhang, Jiaguo Tian, Zhengbang Zhu, Shan Huang, Jucheng Yang, and Weinan Zhang. 2025. Drivegen: Towards infinite diverse traffic scenarios with large models. *arXiv preprint arXiv:2503.05808* (2025).

[213] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. 2023. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493* (2023).

[214] Kotagiri Ramamohanarao, Hairuo Xie, Lars Kulik, Shanika Karunasekera, Egemen Tanin, Rui Zhang, and Eman Bin Khunayn. 2016. Smarts: Scalable microscopic adaptive road traffic simulator. *ACM Transactions on Intelligent Systems and Technology* 8, 2 (2016), 1–22.

[215] Qiujing Lu, Xuanhan Wang, Yiwei Jiang, Guangming Zhao, Mingyue Ma, and Shuo Feng. 2025. OmniTester: Multimodal Large Language Model Driven Scenario Testing for Autonomous Vehicles. *Automotive Innovation* 8, 4 (2025), 838–852.

[216] Chejian Xu, Aleksandr Petiushko, Ding Zhao, and Bo Li. 2025. Diffscene: Diffusion-based safety-critical scenario generation for autonomous vehicles. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 39. 8797–8805.

[217] Jinxiong Lu, Shoaib Azam, Gokhan Alcan, and Ville Kyrki. 2024. Data-Driven Diffusion Models for Enhancing Safety in Autonomous Vehicle Traffic Simulations. *arXiv preprint arXiv:2410.04809* (2024).

[218] Yuting Xie, Xianda Guo, Cong Wang, Kunhua Liu, and Long Chen. 2024. AdvDiffuser: Generating Adversarial Safety-Critical Driving Scenarios via Guided Diffusion. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 9982–9988.

[219] Wei-Jer Chang, Francesco Pittaluga, Masayoshi Tomizuka, Wei Zhan, and Manmohan Chandraker. 2024. SAFE-SIM: Safety-Critical Closed-Loop Traffic Simulation with Diffusion-Controllable Adversaries. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 242–258.

[220] Zhiyu Huang, Zixu Zhang, Ameya Vaidya, Yuxiao Chen, Chen Lv, and Jaime Fernández Fisac. 2024. Versatile scene-consistent traffic scenario generation as optimization with diffusion. *arXiv preprint arXiv:2404.02524* 3 (2024).

[221] Ziyuan Zhong, Davis Rempe, Yuxiao Chen, Boris Ivanovic, Yulong Cao, Danfei Xu, Marco Pavone, and Baishakhi Ray. 2023. Language-Guided Traffic Simulation via Scene-Level Diffusion. In *Proceedings of the Conference on Robot Learning (CORL)*, Vol. 229. 144–177.

[222] Zipeng Guo, Yuchen Zhou, and Chao Gou. 2024. DrivingGen: Efficient Safety-Critical Driving Video Generation with Latent Diffusion Models. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*. 1–6.

[223] Cheng Li, Keyuan Zhou, Tong Liu, Yu Wang, Mingqiao Zhuang, Huan-ang Gao, Bu Jin, and Hao Zhao. 2025. AVD2: Accident Video Diffusion for Accident Video Description. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*. 13289–13296.

[224] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. 2013. Vision meets robotics: The kitti dataset. *The international journal of robotics research* 32, 11 (2013), 1231–1237.

[225] Bozhou Zhang, Jingyu Li, Nan Song, and Li Zhang. 2026. Perception in plan: Coupled perception and planning for end-to-end autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 12376–12384.

[226] Santiago Matalonga, Domenico Amalfitano, Martin Solari, Jean Carlo Rossa Hauck, and Guilherme Horta Travassos. 2025. Testing Context-Aware Software Systems From the Voices of the Automotive Industry. *IEEE Transactions on Industrial Informatics* 21, 5 (2025), 3705–3716.

[227] Yi Huang, Zhan Qu, Lihui Jiang, Bingbing Liu, and Hongbo Zhang. 2026. Prioritizing perception-guided self-supervision: A new paradigm for causal modeling in end-to-end autonomous driving. *Advances in Neural Information Processing Systems* 38 (2026), 49956–49982.

[228] Haochen Liu, Tianyu Li, Haohan Yang, Li Chen, Caojun Wang, Ke Guo, Haochen Tian, Hongchen Li, Hongyang Li, and Chen Lv. 2026. Reinforced Refinement With Self-Aware Expansion for End-to-End Autonomous Driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 48, 5 (2026), 5774–5792.

[229] Yuan Wang. 2020. *Research on lateral obstacle avoidance control strategy and hardware-in-the-loop of intelligent vehicle.* Master's Thesis.

[230] Xiang-mo Zhao, Jing-jun Cheng, Zhi-gang Xu, W Wang, RM Wang, GuanQun Wang, Y Zhu, GP Wang, Y Zhou, and NF Chen. 2019. An indoor rapid-testing platform for autonomous vehicle based on vehicle-in-the-loop simulation. *China Journal of Highway and Transport* 32, 6 (2019), 124–136.

[231] Herbert Schuette and Peter Waeltermann. 2005. Hardware-in-the-loop testing of vehicle dynamics controllers–a technical survey. *SAE transactions* (2005), 593–609.

[232] Zhaodong Zhou, Christopher Rother, and Jun Chen. 2023. Event-Triggered Model Predictive Control for Autonomous Vehicle Path Tracking: Validation Using CARLA Simulator. *IEEE Transactions on Intelligent Vehicles* 8, 6 (2023), 3547–3555.

[233] Bing Zhu, Peixing Zhang, Jian Zhao, and Weiwen Deng. 2022. Hazardous Scenario Enhanced Generation for Automated Vehicle Testing Based on Optimization Searching Method. *IEEE Transactions on Intelligent Transportation Systems* 23, 7 (2022), 7321–7331.

[234] Bing Zhu, Yuhang Sun, Jian Zhao, Sumin Zhang, Peixing Zhang, and Dongjian Song. 2022. Millimeter-Wave Radar in-the-Loop Testing for Intelligent Vehicles. *IEEE Transactions on Intelligent Transportation Systems* 23, 8 (2022), 11126–11136.

[235] Henry X Liu and Shuo Feng. 2024. Curse of rarity for autonomous vehicles. *nature communications* 15, 1 (2024), 4808.

[236] H. Winner, W. Wachenfeld, and P. Junietz. 2018. Validation and introduction of autonomous driving. In *Automotive Systems Engineering II*. Springer, Cham, 177–196.

[237] Gregory M Fitch and Jonathan M Hankey. 2012. Investigating improper lane changes: driver performance contributing to lane change near-crashes. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 56. 2231–2235.

[238] Francois Dion, Ralph Robinson, et al. 2009. *Sweden/Michigan naturalistic field operational test-phase 1: benefits of origin and destination information in IntelliDrive data sets.* Technical Report. University of Michigan. Transportation Research Institute.

[239] Ding Zhao, Huei Peng, Henry Lam, Shan Bao, Kazutoshi Nobukawa, David J LeBlanc, and Christopher S Pan. 2015. Accelerated evaluation of automated vehicles in lane change scenarios. In *Dynamic Systems and Control Conference*, Vol. 57243. American Society of Mechanical Engineers, V001T17A002.

[240] Guangming Xiong, Peiyun Zhou, Shengyan Zhou, Xijun Zhao, Haojie Zhang, Jianwei Gong, and Huiyan Chen. 2010. Autonomous driving of intelligent vehicle BIT in 2009 future challenge of China. In *Proceedings of the 2010 IEEE Intelligent Vehicles Symposium*. 1049–1053.

[241] Motoyuki Akamatsu, Paul Green, and Klaus Bengler. 2013. Automotive technology and human factors research: Past, present, and future. *International journal of vehicular technology* 2013, 1 (2013), 526180.

[242] Mohd Javaid, Abid Haleem, and Rajiv Suman. 2023. Digital twin applications toward industry 4.0: A review. *Cognitive Robotics* 3 (2023), 71–92.

[243] Alexander Barbie, Niklas Pech, Wilhelm Hasselbring, Sascha Flögel, Frank Wenzhöfer, Michael Walter, Elena Shchekinova, Marc Busse, Matthias Türk, Michael Hofbauer, et al. 2021. Developing an underwater network of ocean observation systems with digital twin prototypes—a field report from the baltic sea. *IEEE Internet Computing* 26, 3 (2021), 33–42.

[244] Samir Khan, Michael Farnsworth, Richard McWilliam, and John Erkoyuncu. 2020. On the requirements of digital twin-driven autonomous maintenance. *Annual Reviews in Control* 50 (2020), 13–28.

[245] Chaohui Wu, Zhenzheng Liu, Ke Shi, et al. 2021. Research on the application of digital twin construction and virtual reality fusion in driving scene. *Journal of System Simulation* 33, 02 (2021), 295–305.

[246] Jingyu Liu, Hui Ma, Xuewen Zhang, et al. 2021. Research progress of virtual simulation test technology for autonomous driving. *China Science and Technology Paper* 16, 06 (2021), 571–584.

[247] Selim Solmaz, Martin Rudigier, and Marlies Mischinger. 2020. A vehicle-in-the-loop methodology for evaluating automated driving functions in virtual traffic. In *Proceedings of the 2020 IEEE Intelligent Vehicles Symposium (IV)*. 1465–1471.

[248] Zsolt Szalay, Dániel Ficzere, Viktor Tihanyi, Ferenc Magyar, Gábor Soós, and Pál Varga. 2020. 5G-enabled autonomous driving demonstration with a V2X scenario-in-the-loop approach. *Sensors* 20, 24 (2020), 7344.

[249] Hongyang Sun, Qinglin Yang, Jiawei Wang, Zhen Xu, Chen Liu, Yida Wang, Kun Zhan, Hujun Bao, Xiaowei Zhou, and Sida Peng. 2025. Hierarchy UGP: Hierarchy Unified Gaussian Primitive for Large-Scale Dynamic Scene Reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 26252–26262.

[250] Muhammad Shahbaz. 2025. *From Failure to Fidelity: Enabling Scalable Sim2Real LiDAR Perception Through Realistic Digital Twins.* Ph. D. Dissertation. University of Central Florida.

[251] Walter Zimmer, Gerhard Arya Wardana, Suren Sritharan, Xingcheng Zhou, Rui Song, and Alois C Knoll. 2024. Tumtraf v2x cooperative perception dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 22668–22677.

[252] Tianyue Zheng, Jingzhi Hu, Rui Tan, Yinqian Zhang, Ying He, and Jun Luo. 2024. {π-Jack}:{Physical-World} Adversarial Attack on Monocular Depth Estimation with Perspective Hijacking. In *33rd USENIX Security Symposium (USENIX Security 24)*. 7321–7338.

[253] Kangqiao Zhao, Shuo Huai, Xurui Song, and Jun Luo. 2026. Cheating Stereo Matching in Full-scale: Physical Adversarial Attack against Binocular Depth Estimation in Autonomous Driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 13190–13198.

[254] Luca Giamattei, Antonio Guerriero, Roberto Pietrantuono, and Stefano Russo. 2024. Causality-driven testing of autonomous driving systems. *ACM Transactions on Software Engineering and Methodology* 33, 3 (2024), 1–35.

[255] Huijia Sun, Christopher M Poskitt, Yang Sun, Jun Sun, and Yuqi Chen. 2024. ACAV: a framework for automatic causality analysis in autonomous vehicle accident recordings. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*. 1–13.

[256] Lei-lei Li, Jianwu Fang, Junbin Xiao, Shanmin Pang, Hongkai Yu, Chen Lv, Jianru Xue, and Tat-Seng Chua. 2025. Causal-Entity Reflected Egocentric Traffic Accident Video Synthesis. *arXiv preprint arXiv:2506.23263* (2025).

[257] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*. 1050–1059.

[258] Huiqun Huang, Sihong He, and Fei Miao. 2025. CUQDS: Conformal Uncertainty Quantification Under Distribution Shift for Trajectory Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 17422–17430.

[259] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. 2024. End-to-End Autonomous Driving: Challenges and Frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 12 (2024), 10164–10183.

[260] Julong Wei, Shanshuai Yuan, Pengfei Li, Qingda Hu, Zhongxue Gan, and Wenchao Ding. 2024. Occllama: An occupancy-language-action generative world model for autonomous driving. *arXiv preprint arXiv:2409.03272* (2024).

[261] Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. 2024. Occworld: Learning a 3d occupancy world model for autonomous driving. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 55–72.

[262] Jiankun Sun, Jun Yang, and Zhigang Zeng. 2024. Safety-critical control with control barrier function based on disturbance observer. *IEEE Trans. Automat. Control* 69, 7 (2024), 4750–4756.

[263] Milan Ganai, Sicun Gao, and Sylvia L Herbert. 2024. Hamilton-jacobi reachability in reinforcement learning: A survey. *IEEE Open Journal of Control Systems* 3 (2024), 310–324.

[264] Jordan Lekeufack, Anastasios N Angelopoulos, Andrea Bajcsy, Michael I Jordan, and Jitendra Malik. 2024. Conformal decision theory: Safe autonomous decisions from imperfect predictions. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. 11668–11675.

[265] Shuo Huai, Zhixin Xie, and Jun Luo. 2026. R2DShield: Robust Object Detection in Real-Time via Bayesian Input Shielding. In *Proceedings of the 2026 ACM/IEEE International Conference on Embedded Artificial Intelligence and Sensing Systems*. 1114–1128.

[266] 2024. *ISO 34504:2024 — Road vehicles — Test scenarios for automated driving systems — Scenario categorization.* International Standard. International Organization for Standardization, Geneva, Switzerland. ISO 34504:2024(E).