AI-07-08
7th session of the IWG on AI
March 12, 2026
Provisional agenda item 4

Submitted by the expert from China

Henry Liu[1], Zhong Cao[1], Xintao Yan[1], Shuo Feng[1], and Qiujing Lu[1]

[1]Affiliation not available

May 30, 2025

Autonomous Vehicles: A Critical Review (2004–2024) and a Vision for the Future

Henry X. Liu, *Senior Member, IEEE*, Zhong Cao, *Member, IEEE*, Xintao Yan *Member, IEEE*, Shuo Feng, *Member, IEEE*, and Qiujing Lu

***Abstract*—Autonomous vehicles (AVs) are among the most transformative technologies of the 21st century, reshaping our vision of transportation and mobility. Since the debut of the first prototypes in 2004, rapid technological breakthroughs have significantly advanced AV capabilities, culminating in real-world deployments such as robotaxi services. However, progress has recently slowed, and the path toward achieving safe, scalable, high-level autonomy remains uncertain. Meanwhile, a new wave of AI innovation, particularly in generative AI, is beginning to reshape the AV landscape. At this pivotal moment, critical reflection is necessary. This article highlights key achievements over the past two decades, focusing on major trends propelled by advances in AI. It traces the evolution from rule-based to data-driven approaches, reflecting a shift in design philosophy from "Autonomous Driving by Design" to "Autonomous Driving through Discriminative Learning." To frame this discussion, the article identifies safety and scalability as the two fundamental driving forces behind AV development, examining how successive paradigm shifts have sought to address them and envisioning future systems grounded in "Autonomous Driving through Generative Learning." This work presents a system-level perspective on key milestones, uncovers the forces driving progress and ongoing challenges, and offers visions and insights into future research directions.**

***Index Terms*—Autonomous Driving, Safety, Scalability, Deep Learning, Generative AI.**

I. Introduction

AUTONOMOUS Vehicles (AVs) represent one of the most transformative technologies of the 21st century, with the potential to revolutionize transportation systems, improve road safety, and reshape urban mobility. Since the inception of AV technology in the early 2000s with the DARPA Grand Challenge, significant progress has been made over the past two decades, moving from theoretical concepts to limited commercial deployment. However, despite these advancements, the path toward achieving safe and scalable high-level autonomy remains elusive. As the field enters a critical juncture, it demands reflection, reassessment, and renewed innovation to address persistent challenges and realize its full potential.

This research was partially funded by the U.S. National Science Foundation through the Mcity 2.0 Project (CMMI #2223517). *(Henry X. Liu, Zhong Cao, Xintao Yan, and Shuo Feng contributed equally to this work.) (Corresponding author: Henry X. Liu.)*

Henry X. Liu is with the Department of Civil and Environmental Engineering and the University of Michigan Transportation Research Institute, University of Michigan, Ann Arbor, MI 48109 USA, (e-mail: henryliu@umich.edu).

Zhong Cao and Xintao Yan are with the Department of Civil and Environmental Engineering, University of Michigan, Ann Arbor, MI 48109 USA, (e-mail: zhcao@umich.edu, xintaoy@umich.edu).

Shuo Feng and Qiujing Lu are with the Department of Automation, Tsinghua University, Beijing 100084, China, (e-mail: fshuo@tsinghua.edu.cn, qiujinglu@mail.tsinghua.edu.cn).

A. Understanding the Current Status of Autonomous Vehicles

After two decades of development, the architecture of modern AV systems has converged around three core modules: perception, planning, and control. This modular paradigm has guided both academic research and industry applications, with deep learning playing a central role. Recently, end-to-end (E2E) approaches—especially those leveraging large-scale neural networks trained on human demonstrations or labeled datasets—have emerged as promising alternatives, offering the potential for higher performance but at the cost of increased opacity.

Driven by advancements in artificial intelligence, increased availability of high-quality datasets, and greater computational power, AV systems have achieved significant performance milestones. These developments have facilitated the emergence of commercial AV services across various levels of automation. Notable examples include Waymo's fully driverless Level 4 ride-hailing services in cities like San Francisco, Los Angeles, and Phoenix, which now deliver over 150,000 paid trips and 1 million autonomous miles weekly [1]. Baidu has also rolled out robotaxi services across multiple cities in China, delivering nearly 900,000 rides in Q2 2024 [2], while Tesla's Autopilot and Full Self-Driving (FSD) offer widespread Level 2 capabilities.

These developments demonstrate the feasibility of small-scale deployments for AV systems. However, they come with significant constraints. Current deployments often rely on extensive high-definition mapping, manual rule-setting tailored to specific regions, and exhaustive testing to address unique local conditions. This highly localized and customized approach, with tailored operational design domains (ODDs), makes scalability a daunting challenge. Deployment typically begins in densely populated cities with predictable returns on investment, and service areas are often limited to simpler ODDs, such as restricted times of day and favorable traffic conditions. Moreover, safety remains a critical concern. A high-profile incident involving Cruise in 2023—where a pedestrian was dragged by a driverless vehicle—led to a six-month suspension of operations and the company's eventual shutdown by General Motors in 2024 [3]. While AV companies frequently claim safety performance exceeding that of human drivers[1,2], such claims are typically self-reported, with limited transparency or third-party validation. This contributes to ongoing public

[1]https://www.tesla.com/VehicleSafetyReport/

[2]https://waymo.com/blog/2023/12/waymo-significantly-outperforms-comparable-human-benchmarks-over-7-million/

skepticism, as reflected in a 2025 AAA survey showing that only 13% of U.S. drivers are comfortable riding in AVs [4].

B. Motivation and Uniqueness of This Survey

Despite two decades of rapid development, progress toward high-level autonomy has begun to plateau. Incremental refinements alone have proven insufficient to overcome the fundamental challenges of safety and scalability. For instance, it took Waymo nearly a decade to launch its first fully driverless service in Phoenix in 2019, and expansion to just two additional cities followed over the next five years. At the same time, a new wave of innovation—driven by large language models (LLMs) and foundation models—is beginning to influence the AV research landscape. Yet, their role in advancing the long-standing goals of autonomous driving remains uncertain and largely underexplored.

In this context, there is a growing need for a comprehensive and critical assessment of where the field stands and where it is headed. This survey aims to fulfill that need. Our goal is twofold: to provide a retrospective analysis of key developments in AV technology and to offer a forward-looking framework that can guide future innovation. The survey is designed to be accessible to newcomers seeking an entry point into the AV field, while also offering experienced researchers a structured synthesis of emerging trends and open challenges.

This survey distinguishes itself from existing literature in several key ways:

- **In-Depth Analysis of Critical Challenges**: While existing surveys often catalog technical advances, we center our discussion around two persistent bottlenecks—safety and scalability—and explore how these challenges have shaped and constrained the evolution of AV systems.
- **Systematic Overview of Technological Development**: Rather than focusing on isolated components or specific methodologies, our survey provides a holistic view of the technological evolution of AV systems. We categorize the existing developmental journey into AV 1.0 (Autonomous Driving by Design) and AV 2.0 (Autonomous Driving through Discriminative Learning) and propose the future AV 3.0 (Autonomous Driving through Generative Learning). This structured framework allows us to trace how development strategies have shifted over time—from rule-based engineering to data-driven learning, and now toward generative models.
- **Vision for the Future**: Our survey not only consolidates existing knowledge but also provides a forward-looking perspective. We articulate our vision for overcoming the core challenges faced by AV systems and summarize initial exploration towards AV 3.0. Our goal is to spark new ideas and catalyze research that addresses the fundamental roadblocks to safe and scalable autonomy.

C. Structure of This Survey

The remainder of this paper is organized as follows: In Section II, we analyze the two fundamental driving forces of AV development—safety and scalability—and review how each phase of AV evolution has attempted to address them. Sections III and IV present comprehensive reviews of the AV 1.0 and AV 2.0 paradigms, including key milestones, representative methods, and remaining limitations. In Section V, we offer our perspective on the transition toward AV 3.0, summarizing early explorations and discussing future research directions aimed at enabling high-level autonomy. Finally, Section VI concludes the survey and reflects on the implications of the technological trajectory we have outlined.

II. Driving Forces for AV Development: Safety and Scalability

A. Safety and Scalability Requirements for AV

Safety and scalability have been the two core driving forces behind the development of AVs over the past two decades. The foundational motivation for AV technology arises from a persistent and well-documented issue: human error accounts for approximately 94% of traffic crashes, as reported by the National Motor Vehicle Crash Causation Survey [5]. The central vision for AVs is to create intelligent machine drivers—free from distraction, fatigue, or impaired judgment—that can significantly reduce crash rates and improve overall road safety. Consequently, from early experimental prototypes to today's commercial pilot programs, the pursuit of safe and scalable AV systems has consistently guided both research agendas and industry strategies.

The safety objective for AVs is twofold: (1) statistically outperform human drivers in terms of crash rates and fatality rates, delivering a measurable safety improvement over the existing transportation system; (2) guarantee the avoidance of commonplace driving errors, specifically those that a non-impaired, attentive, highly skilled human driver would rarely make. By achieving these two goals, AVs can fulfill their original promise of eliminating the 94% of crashes attributed to human error and offer superior, reliable road safety at a societal scale.

In parallel, scalability remains a critical requirement for realizing the full potential of AV technology. The goal is to develop AV systems capable of operating across broad operational design domains (ODDs) with minimal adaptation or operational overhead, while consistently upholding high performance of safety. Central to this capability is robust model generalization—the ability of learned models to perform reliably in diverse, previously unseen environments without retraining and exhaustive location-specific tuning, enabling large-scale, cost-effective deployment. Meeting this challenge is essential for bringing the vision of Level 5 autonomy closer to reality.

B. Curse of Dimensionality and Curse of Rarity: Root Causes of the Dilemma

Two fundamental challenges underlie the difficulties in achieving AV safety and scalability: the Curse of Dimensionality (CoD) and the Curse of Rarity (CoR) [6]. Together, these factors give rise to the persistent dilemma of safety versus scalability. The real-world driving environment is inherently high-dimensional, encompassing complex road geometries,

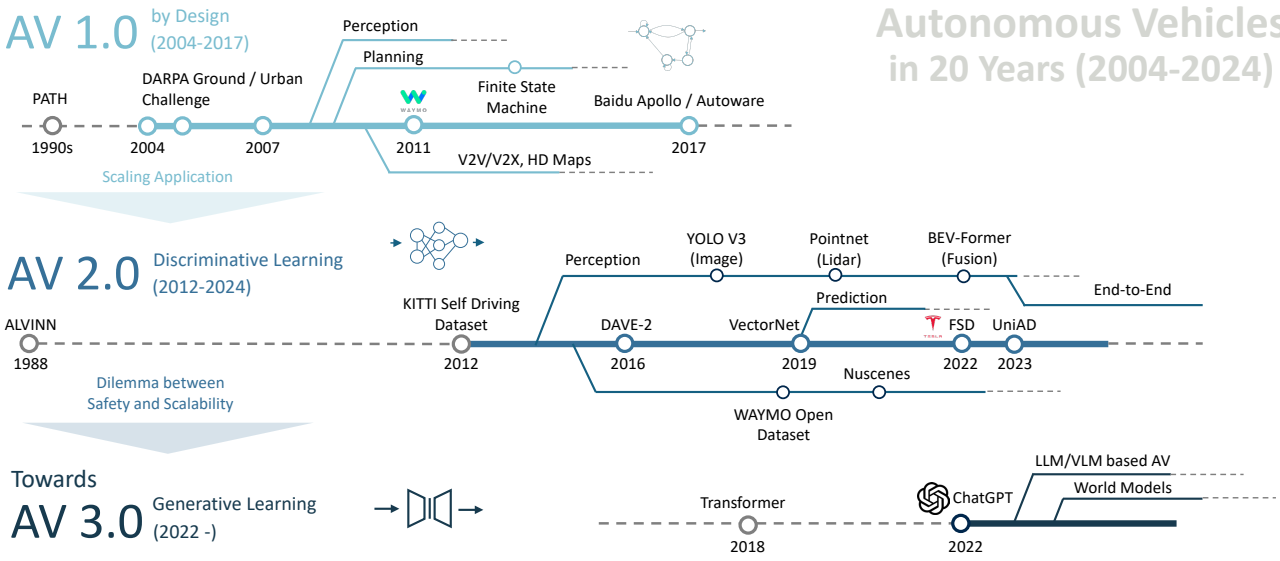


Fig. 1. Autonomous Vehicles in 20 Years (2004-2024)

heterogeneous traffic infrastructures, diverse traffic rules, dynamic road users, variable weather conditions, and constantly changing temporal contexts. As a result, the operational state space that an AV must handle becomes exceedingly large, leading to the CoD problem, where the amount of data needed to adequately represent and generalize across all possible situations increases exponentially with the number of relevant variables.

On top of this, the Curse of Rarity (CoR) poses a critical challenge. As emphasized in [6], autonomous vehicle (AV) training typically requires exhaustive data collection and road testing to ensure safety. However, the AV's performance cannot be guaranteed in scenarios that have not been explicitly encountered during training. Given the continuous and highly diverse nature of real-world driving, AVs are vulnerable not only to rare, safety-critical events but also to failures in seemingly routine situations—those that are well-represented in the training data but not captured in precisely the same form. A recent survey of self-driving accidents [7] revealed that 96.45% of failures occurred under normal conditions, such as daylight and dry roads, and 93.2% involved no dangerous behaviors. This paradox highlights the limitations of current AV generalization capabilities, significantly inflating the cost of training and testing, and contributing to persistent public skepticism toward autonomous driving technologies.

The compounding effects of CoD and CoR create a fundamental dilemma in autonomous driving. To improve safety, developers often resort to constraining the ODD, running more physical testing, and adding cost in the form of more sensors, HD maps, compute power, teleoperators, and other hazard countermeasures. While beneficial to safety, these measures undermine AV scalability to high-volume implementations in widely variable usage environments. Over the past two decades, addressing this tension between safety and scalability has been one of the most significant research and engineering challenges in the field, driving a series of paradigm shifts in AV system design, as discussed next.

C. AV Development Stages

To address the intertwined challenges of safety and scalability, the AV research community has undergone several major methodological paradigm shifts over the past two decades. Although there is no universally accepted framework for demarcating the stages of AV development—and the boundaries between them are inherently fluid—we adopt a categorization informed by foundational paradigms in the broader field of artificial intelligence. Based on this perspective, we delineate the evolution of AV systems into three distinct stages, as shown in Fig. 1.

1) AV 1.0: Autonomous Driving by Design: In this early phase, the primary objective was to demonstrate AV functionality by building complete, rule-based AV pipelines. These systems were mostly manually engineered with explicit, human-crafted logic for perception, decision-making, and control. While these early designs enabled proof-of-concept demonstrations and the execution of simple driving tasks, they lacked adaptability and robustness in complex, dynamic, and unpredictable real-world environments. As a result, AV 1.0 systems suffered from limited safety performance and scalability, constrained by the inherent limitations of rule-based systems and the CoD problem.

2) AV 2.0: Autonomous Driving through Discriminative Learning: This stage marked a paradigm shift toward data-driven, learning-based approaches, particularly leveraging deep learning within a discriminative framework. Core AV components such as perception, prediction, and planning increasingly rely on learning mappings from observed inputs to labeled outputs — such as inferring object detections from sensor data or generating control actions from traffic scenes.

Deep neural networks offered far superior representational capacity than rule-based systems, significantly alleviating the CoD issue by enabling AV systems to handle high-dimensional, complex driving environments. However, discriminative learning methods focus on modeling conditional

distributions, expressed as $p(y|x)$, where the output y is conditioned on the given input x. This approach inherently assumes that (1) the available training dataset x sufficiently covers the full space of possible driving situations, and (2) the learned function $p(y|x)$ can generalize to the complete joint space of (x, y) based on this partial dataset. In practice, extensive experience during the AV 2.0 phase revealed that these assumptions break down in autonomous driving. Real-world driving occurs in an open, dynamic, and highly unpredictable environment, where collected (x, y) pairs only capture a narrow subset of possible situations. Models trained on such discrete, limited datasets struggle to generalize to unseen or rare situations, which are often the most safety-critical. This limitation raises significant safety concerns and leaves AV 2.0 systems vulnerable to out-of-distribution events — the long-tail cases at the heart of the CoR problem. Although AV 2.0 systems delivered notable improvements in scalability and performance over rule-based predecessors, they continued to face fundamental challenges: poor generalization in long-tail scenarios, a lack of true scene understanding and reasoning capabilities, and limited interpretability. These shortcomings ultimately constrained their safety and scalability in real-world driving environments.

3) Towards AV 3.0: Autonomous Driving through Generative Learning: The field is now entering a third stage—AV 3.0—driven by the growing recognition that the limitations of AV 2.0 stem fundamentally from the discriminative learning paradigm that defines it. A notable shift in AV 3.0 lies in its emphasis not only on aligning generated trajectories with human demonstrations, but also on capturing the internal structure that emerges during the trajectory generation process. This perspective follows the generative learning paradigm, which aims to model the joint distribution between inputs and outputs, i.e., $p(x, y)$, rather than merely learning a conditional mapping. These internal relationships may reflect physical constraints, contextual consistencies, or spatial regularities—providing the foundation for scene understanding and reasoning about how environment dynamics influence driving actions. In AV 3.0, the agent is no longer trained on isolated instances; instead, it aims to generalize across interrelated scenarios through structured and regularized modeling, enabling it to learn from, rather than merely mimic, observed data.

Notably, not all generative learning architectures—such as Transformers or Diffusion models—are inherently suited for autonomous driving tasks. Although these models are capable of learning internal relationships, the representations they capture may not align with the specific relational structures required for driving. The compatibility between model architecture and task structure plays a crucial role in determining model effectiveness. For instance, Transformers perform well in natural language processing due to their ability to model long-range dependencies via attention mechanisms, while earlier generative models such as GANs struggled in this regard. Likewise, Diffusion models have shown superior performance in image generation tasks compared to Transformers, owing to their better adaptation to global and local image features. These examples highlight that architectural suitability is highly task-dependent. Therefore, the central challenge in realizing AV 3.0 lies in designing a learnable generative architecture that can acquire the foundational internal relationships essential to driving—relationships that enable inherently scalable and safe autonomous behavior.

Recent research related to the AV 3.0 paradigm, particularly those efforts involving vision-language models (VLMs) and world models, attempt to leverage their pre-trained internal relationships across driving scenarios. However, we observe that relying on existing pre-trained architectures remains insufficient to meet the scalability and safety requirements of autonomous driving. Further research on how to design structured and regularized models—more explicitly aligned with the unique characteristics of driving tasks—may serve as a critical foundation for the future development of autonomous vehicle systems.

III. AV 1.0: Autonomous Driving by Design

A. Overview

Autonomous driving technology was initially developed to improve traffic safety. During the 1950s-1990s, vehicles had become essential for transportation but were also a leading cause of traffic accidents, driving the development of vehicle safety technologies. One of these research ideas aimed to assist human driving and mitigate human errors, called driver assistance systems (DAS) [8]. It typically obtained the vehicle and surrounding environment status and then intervened to help control the vehicle when necessary, enabling limited autonomous operation. This also inspired researchers' exploration of the vehicles with more intelligence.

Early research on intelligent vehicles encompassed a range of objectives, from developing comprehensive intelligent systems integrating multiple technologies to focusing on specific functionalities like lane keeping and obstacle detection, laying the groundwork for modern autonomous driving technologies. A landmark initiative was the launch of the Partners for Advanced Transit and Highways (PATH) program in 1986, led by the University of California, Berkeley, which aimed to improve traffic flow and safety through advanced vehicle technologies [9]. PATH was a pioneer in developing early automated and connected vehicle systems, including traffic detection technologies and wireless communication infrastructure. In parallel, Advanced Driver Assistance Systems (ADAS) were actively developed to support intelligent driving functions such as Advanced Emergency Braking (AEB), cruise control, adaptive cruise control, and lane-keeping assistance [10]. Other research efforts focused on enabling precise control of vehicles along predefined trajectories, leading to significant progress in vehicle dynamic control technologies [11]. A notable milestone from this era was the development of ALVINN (Autonomous Land Vehicle In a Neural Network), which used a neural network trained on camera images to steer a vehicle [12]. Often cited as a foundational example of data-driven autonomous driving, ALVINN marked an early integration of machine learning into vehicle control. During this formative period, research objectives, system architectures, and methodological approaches varied widely across projects. A unified technological framework for autonomous

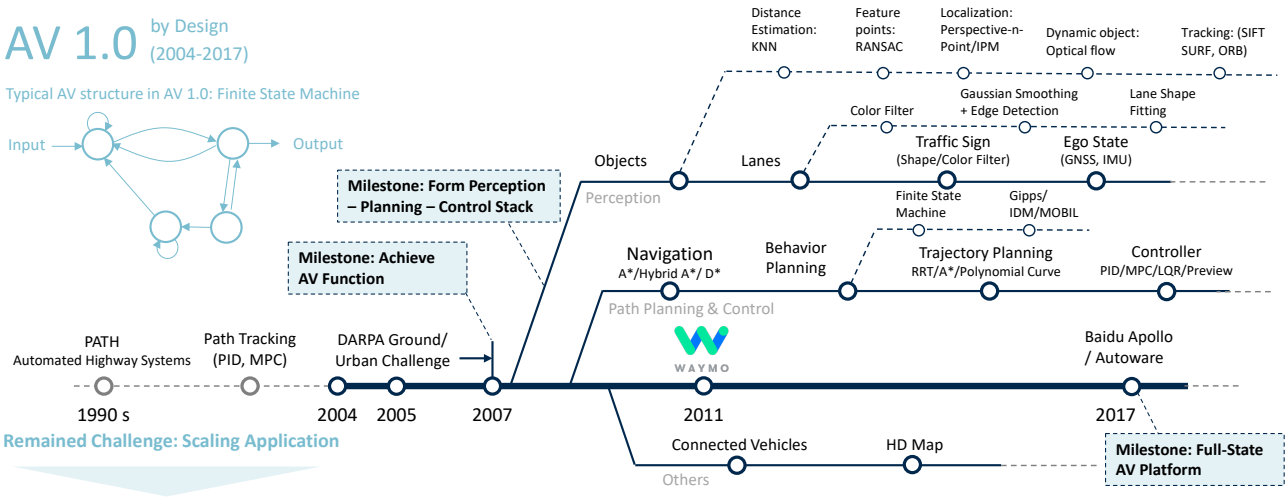


Fig. 2. AV 1.0: Autonomous Driving by Design

driving had yet to emerge, and consensus within the research community regarding the path forward remained elusive.

A turning point occurred with the 2007 DARPA Urban Challenge [13], which required participating teams to design autonomous vehicles capable of completing a fully self-driving journey in an urban environment. During the challenge, vehicles needed to perceive their surroundings, avoid dynamic obstacles, plan the path, control the vehicle, and arrive at the destination. This challenge attracted participants from Stanford [14], MIT [15], and CMU [16] et al. After the challenge, all teams published technical papers detailing their approaches to achieving autonomous driving. Some teams subsequently began commercial development of autonomous driving technologies, e.g., Waymo (Google). This marked the emergence of a consensus on the architecture of the first-generation autonomous driving systems, i.e., AV 1.0, and initiated focused research on various system modules shown in Fig. 2. The following sections introduced the AV 1.0 and related subsequent studies.

B. Architecture

The widely accepted structure of autonomous driving systems primarily includes perception, planning, control, and others. This system structure was also adopted by most DARPA Urban Challenge teams, shown in Fig. 3. The rationale behind this segmentation lies in the distinct objectives and functions of each module, as well as the fact that the technologies required to achieve these objectives were often from closely related fields. This section follows this structure to review the related research works.

C. Perception

The goal of the perception module is to obtain the driving conditions, including static and dynamic road objects, road lanes, traffic signs, and the ego vehicle's kinematic states. These methods are closely related to the sensors equipped on autonomous vehicles. In the DARPA Urban Challenge,

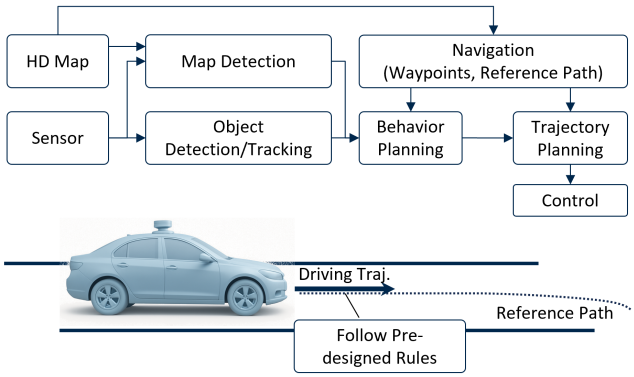


Fig. 3. A typical structure in AV 1.0: The agent relies on a predefined set of rules to generate driving trajectories.

commonly used environmental perception sensors included LiDAR, cameras, radar, GPS, and IMU. This paper does not provide an in-depth discussion of sensors; readers may refer to [17] for related work. In subsequent sections, we will focus on perception methods that leveraging these sensors.

1) Objects: The initial concept of object detection focused on identifying the drivable area, which simply required detecting areas without elevated objects on the ground. LiDAR proved to be an ideal sensor for this task, as it emitted dense point clouds around the vehicle and measured the distance of each point from the vehicle. By applying methods such as K-Nearest Neighbors (KNN) [18] clustering, objects elevated above the ground could be identified. This approach did not require object recognition or classification, yet it provided reliable collision avoidance for vehicles. As a result, LiDAR was widely adopted in early autonomous vehicles.

Images can also be used to identify environmental obstacles. The primary approach involves extracting feature points from the image and estimating their relative position to the vehicle using the physical characteristics of the camera, thus identifying objects elevated above the ground. A classic method for

detecting image feature points is RANSAC (Random Sample Consensus) [19].

The approach for evaluating relative positions depends on the number of cameras installed on the vehicle. Stereo cameras use the relative distance between two cameras with PnP (Perspective-n-Point) [20], while monocular cameras rely on the camera's height above the ground and its intrinsic parameters using IPM (Inverse Perspective Mapping) [21]. The measurement accuracy of stereo cameras generally depends on the precision of feature point matching between the two images, whereas monocular cameras are more reliant on calibration accuracy.

For dynamic objects, their velocity was crucial for autonomous driving, requiring the perception module to perform sequential measurements of these objects. Two common methods were widely employed for this purpose: 1) Optical Flow Method [22]: This technique analyzed the differences between two consecutive frames to generate an optical flow, which was used to estimate the motion of dynamic obstacles; 2) Feature Point Matching Method: Algorithms such as Scale-Invariant Feature Transform (SIFT), Speeded Up Robust Feature (SURF), and Orientated FAST and Robust BRIEF (ORB) [23] matched feature points between consecutive frames, and the displacement of these points was used to estimate their velocity and, consequently, the motion of the object. Among these, the feature point matching method generally offers higher accuracy and is therefore more widely adopted in practice.

There were also some feature-based vehicle detection methods, e.g., detecting the vehicle plates by their shape and color, and then detecting the car [24]. However, the localization of vehicles was significantly below acceptable levels, not to mention the challenges in recognizing pedestrians and uncommon objects. Thus, measuring the drivable area was the main target at that time.

2) Lanes: Lane detection usually relies on images, with feature-based methods falling into three main categories: color-based, edge-based, and shape-based approaches. Color-based methods identify lane markings by setting color thresholds to extract yellow and white pixels from the road surface. Edge-based approaches first apply a blurring process to the image and then extract edges from the blurred color gradients to identify lane lines. Shape-based methods assume that the lane lines take the form of straight or continuous curved lines. Based on this assumption, they fit key feature points on the road surface to derive the lane boundaries.

The details of these methods can be found in [25], which also includes the pre-processing of the image. All three methods rely on some assumptions about the attributes of lane lines. When these assumptions are not met, the accuracy of detection often suffers. Consequently, these methods struggled to achieve high precision. During this period, enhanced digital maps and high-precision maps (HD maps) [26] collected in advance were commonly used to obtain lane line information, adopted by most DARPA Urban Challenge teams.

3) Traffic Signs: Similarly, traffic sign detection also relies on images and sign-specific features, including shape, color, symbols, and text. Shapes, symbols, and text are typically identified using template-matching methods [27], while color features are detected through color filtering techniques. Since the types of traffic signs are usually fixed, these methods generally achieve good detection performance.

However, the challenge lies in understanding the applicable range of detected traffic signs. For instance, detecting a speed limit sign that does not apply to the current lane may cause the vehicle to unnecessarily slow down, potentially leading to safety risks. As a result, enhanced digital maps and HD maps have still been widely used for traffic sign detection to ensure reliability.

4) Ego states: Ego vehicle's state usually includes its position, orientation, velocity, and acceleration, which are measured using onboard sensors. Position estimation is usually achieved through satellite-based navigation systems (GNSS) [28], with higher-precision differential real-time kinematic positioning RTK-GPS [29] often employed, achieving accuracy within 10 cm. The vehicle's heading is calculated using two GPS devices placed at the front and rear of the vehicle. Velocity and acceleration can be derived from positional differences or measured using wheel speed sensors and inertial measurement units (IMU) [30].

Ego vehicle's state information is critical not only for vehicle planning and control but also for querying map data based on the vehicle's position. Therefore, the accuracy of vehicle state estimation is essential for the effectiveness of autonomous driving systems.

D. Path Planning and Control

After obtaining the driving environment state, the autonomous vehicle should plan a reasonable route and control itself to reach the destination. This process typically involves four components: navigation, behavior planning, trajectory planning, and control. Navigation is responsible for mapping an origin-to-destination path within the road network, while high-level behavior planning generates abstract commands such as yielding, passing, or lane changing. Trajectory planning, on the other hand, is focused on generating detailed vehicle trajectories, while motion control translates these trajectories into actionable throttle, brake, and steering commands. These components generally operate sequentially but often overlap in functionality. For example, some systems integrate behavior planning with trajectory planning, where trajectory planning inherently includes decision-making. Similarly, some systems consider trajectory generation during navigation planning. This paper provides an overview of the objectives and common approaches for these four modules. However, it should be noted that not all systems necessarily include all four components.

1) Navigation: Navigation methods had been explored before autonomous driving. The goal is to determine the optimal route from a starting point to a destination, with graph searching methods, e.g., A* and D*. Related work can be found in [31]. Traditional navigation systems typically output a sequence of road segments and intersections. However, early autonomous driving algorithms required denser route information. To address this, navigation systems for autonomous

driving usually adopt waypoint-based or reference trajectory methods, where waypoints or trajectories connect the start and end points. During driving, autonomous vehicles can follow these waypoints or trajectories and avoid objects on the road. This largely simplifies the autonomous driving problem. The DARPA Urban Challenge also utilized the waypoint method for releasing driving tasks.

However, ideal autonomous vehicles should plan the path according to the real-time environment instead of always following a pre-designed trajectory. Thus, as autonomous driving algorithms improved, this method was gradually phased out. For related research, please refer to [32].

2) Behavior Planning: Driving behavior is critical as it reflects the intelligence of autonomous vehicles. The development of behavior planning methods has progressed through three key stages: finite state machines, behavior modeling, and general behavior planning.

In the DARPA Urban Challenge, most teams adopted the finite state machine (FSM) approach for behavior decision-making [14]. The FSM method categorized all environmental states and assigned specific behaviors to each category to achieve decision-making. The advantage of this approach was its ability to associate vehicle behavior with environmental states while ensuring that different behaviors did not interfere with each other. This simplified the decision-making process for autonomous driving while enabling more complex vehicle behaviors.

This method was later extended into scenario-based autonomous driving [33], where vehicle behaviors were further refined by constructing extensive scenario libraries. This approach significantly influenced the development of autonomous driving systems for a long time thereafter. However, as the scenario libraries expanded, researchers began to encounter challenges such as overly complex scenario descriptions, unclear scenario boundaries, and difficulties in defining certain scenarios.

Therefore, many studies introduced driver behavior models to integrate multiple scenarios, for example, the Integrated Driving Model (IDM) [34] and Minimizing Overall Braking Induced by Lane Changes (MOBIL) [35] models, derived from microscopic traffic simulation models. These models define vehicles' longitudinal and lateral behaviors, enabling tasks like car-following and lane-changing, thereby merging numerous multi-lane scenarios into a single scenario.

Simultaneously, behavior algorithms tailored to specific scenarios were proposed, addressing tasks such as merging or exiting highways, intersections, and yielding to pedestrians. However, due to the complexity of scenario segmentation, no widely accepted models for specific scenarios have been proposed, limiting further development of this approach.

To overcome the limitations of scenario-specific behavior models, general behavior planning methods have been developed to enable broader adaptability across diverse driving contexts. Among these, sampling-based approaches have gained prominence. These methods frame behavior generation as an optimization problem: finding a feasible and cost-effective trajectory within the vehicle's reachable state space.

A representative class is structured sampling-based methods, such as lattice-based planning [36], [37]. These approaches discretize the continuous state space into a lattice of candidate states and use precomputed motion primitives to ensure dynamically feasible paths. Cost functions - which account for safety, comfort, traffic rules, and task-specific goals - guide the selection of the optimal trajectory from the set of candidates.

3) Trajectory Planning: The goal of trajectory planning is to generate a path or trajectory from the current or a specific starting point to a target point. Here, a path usually means a sequence of positions, while a trajectory also contains the speed at these positions. This section will generally use the term trajectory for short. A trajectory should avoid collisions with surrounding objects and remain smooth to be tracked by the controller. However, the definition and scale of trajectory planning vary significantly across different systems.

In early systems, trajectories were often static or updated only when necessary, resulting in large-scale trajectory plans that sometimes included navigation functionality. As behavior decision-making modules became more sophisticated, trajectories needed to adapt dynamically to behavioral changes. This increased the frequency of trajectory generation and reduced the planning scale. Consequently, algorithm complexity decreased while real-time performance improved.

The emergence of general behavior models further advanced and unified trajectory planning algorithms, often integrating trajectory planning and decision-making into a single module. This section introduces some representative algorithms of these three phases.

In the DARPA Urban Challenge, most teams used the classical A* algorithm [14] for trajectory planning, while MIT's team adopted the Rapidly-Exploring Random Tree (RRT) [15] method. Both algorithms computed a collision-free trajectory from the starting point to the target point on a planar surface. The A* algorithm iteratively searched for the optimal path from the start to the goal by generating a trajectory through pre-established nodes in the space, evaluating the cost to reach each node. However, if the nodes were too dense, the method became inefficient, whereas overly sparse nodes resulted in poor-quality trajectories. In contrast, the RRT algorithm sampled points randomly in the space and refined the trajectory iteratively, balancing efficiency and performance. This made RRT the state-of-the-art method during that phase.

Nonetheless, both methods had limitations, particularly their inability to account for dynamic environmental information. Additionally, RRT was more challenging to integrate with behavior decision-making modules. As a result, RRT did not become a mainstream algorithm for autonomous driving systems.

With better behavior planning modules, trajectory planning modules mostly aim to execute specific behaviors. During this time, algorithms such as A* and Hybrid A* remained commonly used. Meanwhile, methods based on predefined curves were introduced to accelerate trajectory generation while ensuring smoothness. These curve types include polynomial curves, B-spline interpolations, and Bézier curves [38].

The computation of these curves is typically based on constraints from the start point, end point, and intermediate

control points, allowing for the rapid generation of trajectories. The advantage of curve-based methods lies in their high computational efficiency and guaranteed smoothness. However, their limitation is the inability to represent complex trajectory shapes. Consequently, in more complicated scenarios, such as parking lots, algorithms like A* still work well.

Trajectory planning algorithms later evolved and converged into the Lattice algorithm [39], which is based on spatial sampling and uses quintic polynomials for smoothing. This method offers significant flexibility for various behaviors while maintaining high efficiency, making it widely adopted in self-driving systems.

4) Control: A control module is designed to enable the vehicle to track the planned trajectory accurately. PID controllers [40] and optimal control techniques (like MPC) [41]–[43] were widely adopted. These methods ensured stable and predictable control performance, particularly in urban driving contexts. This paper does not delve into these methods in detail; readers may refer to [44] for related research.

It should be noted that during the development of self-driving cars, some vehicle control algorithms incorporated environmental information to add safeguards or correct unsmooth trajectories. These methods emerged due to the limited capabilities of decision-making and trajectory planning modules, necessitating assistance from the control module. However, as self-driving capabilities improved, the control module returned to focusing solely on the execution of its control objectives.

E. Others

In addition to the essential modules required for self-driving functions, certain supportive modules have also been developed independently. Here, we focus on introducing two key modules: connected vehicles and enhanced digital maps.

1) Connected Vehicles: Connected vehicle technology was initially developed to collect dynamic traffic information, with the main technical challenge being the establishment of a stable and reliable communication network [45].

For self-driving cars, the most critical role of connected vehicles lies in enhancing perception capabilities in situations of insufficient sensor performance. Using communication methods, the connected vehicle network can provide accurate and reliable information about surrounding vehicles and pedestrians. In addition, connected vehicle technology offers potential solutions to challenges such as autonomous vehicle platooning [46], cooperative perception [47], and cooperative decision-making [48].

2) HD maps: Accurate lane and traffic sign detection remains challenging for onboard sensors, making High-definition (HD) maps [26], [49] an important tool in the AV by providing highly detailed, pre-constructed representations of the road environment. These maps are used to augment real-time perception systems and support localization, navigation, and decision-making [50], [51], particularly in complex or unfamiliar environments. HD maps typically include precise lane-level information, road geometry, traffic signs, and other static elements such as intersections and crosswalks.

However, the creation and continuous updating of HD maps require substantial computational resources and data [52], presenting ongoing challenges for the field. More detailed information on HD mapping can be found in [53]. To address these, simultaneous localization and mapping (SLAM) has become a crucial technology, enabling real-time map creation and localization in previously unexplored areas. SLAM facilitates the real-time update of maps, helping to tackle dynamic changes such as road construction or environmental alterations, which are critical for maintaining map accuracy and consistency. For more detailed discussions on SLAM, see [54].

3) Open-source systems: A key milestone of the AV 1.0 stage was the release of open-source AV platforms, which enabled rapid research and development. Notable systems such as Autoware [55] and Baidu Apollo [56] provided modular, full-stack pipelines covering core AV functions, including perception, planning, and control. These platforms allowed researchers and developers to experiment with, customize, and extend AV functionalities, accelerating innovation and fostering collaboration across academia and industry.

F. Discussions

In the AV1.0 stage, autonomous driving technology achieved the following objectives: 1) The hardware of autonomous vehicles could support self-driving capabilities. 2) The community formed a well-adopted autonomous driving system consisting of perception, planning, and control modules. 3) Autonomous driving vehicles demonstrated self-driving ability under some given conditions.

However, the remaining issues to be addressed during this stage were: 1) The accuracy and reliability of environmental perception remained insufficient, compromising driving safety. 2) There were still concerns regarding the trustworthiness of large-scale deployment of autonomous driving technology.

IV. AV 2.0: Autonomous Driving through Discriminative Learning

A. Overview

The transition from AV 1.0 (Autonomous Driving by Design) to AV 2.0 (Autonomous Driving through Discriminative Learning) marks a pivotal evolution in AV development. Unlike the hand-crafted rules-based methods that dominated the AV 1.0 era, AV 2.0 embraces data-driven approaches, leveraging large-scale real-world datasets, including human-labeled annotations and driving demonstrations, to train both modular and end-to-end (E2E) systems. This shift automated feature extraction and function design, enabling more scalable and adaptable solutions to meet the complexity of real-world driving environments. Methods from this era predominantly followed a discriminative learning paradigm, mapping inputs directly to outputs based on labeled data pairs rather than depending on explicitly encoded rules or model-based designs.

The motivation for this transition stemmed from the inherent limitations of AV 1.0. While AV 1.0 established a foundational architecture—comprising perception, planning, and control—its model-based methods struggled to cope with the

CoD of real-world driving. The diverse behaviors of road users and the dynamic nature of the environment posed significant challenges for heuristic-driven approaches, limiting their scalability and intelligence. As a result, AV 1.0 systems largely remained confined to structured environments or conceptual demonstrations.

Two major factors catalyzed the emergence of AV 2.0. First, the rise of deep learning (DL) revolutionized approaches to high-dimensional tasks. Landmark achievements such as AlexNet [57] demonstrated the potential of DL for computer vision, while AlphaGo [58] highlighted the promise of reinforcement learning (RL)[3] for decision-making in complex domains. Second, the availability of large-scale, high-quality datasets—such as KITTI [59], [60] and the Waymo Open Dataset [61], [62]—enabled data-driven methods to advance rapidly.

Together, these technological and data-driven breakthroughs laid the foundation for transformative growth in AV 2.0, leading to two major milestones: the development of AI-enhanced modular AV systems and the emergence of E2E AV pipelines.

B. Architecture

The AV 2.0 architecture initially extended the modular structure established in AV 1.0, retaining the core components of perception, planning, and control. However, each module evolved substantially, with data-driven learning replacing much of the handcrafted design. The architectural strategy followed a divide-and-conquer approach: each component was developed and optimized independently using large-scale data, and later integrated into a complete system. A typical structure of AV 2.0 is shown in Fig. 5.

As research progressed, E2E pipelines emerged, encompassing either fully differentiable modular systems or complete black-box models trained in a holistic manner. The E2E paradigm offered the potential to minimize compounded errors and preserve valuable information across components by optimizing the full AV stack jointly. Alongside these architectural innovations, this period also saw the rapid development of a broad ecosystem of supporting tools, including large-scale datasets, standardized benchmarks, and high-fidelity simulators, all of which accelerated AV development and deployment.

Broadly, AV 2.0 can be characterized by the following key advancements:

- **Perception**: The perception module shifted entirely to learning-based approaches. Breakthroughs in deep learning for computer vision drove advancements in object detection, semantic segmentation, and sensor fusion, utilizing data from cameras, LiDARs, and radars.
- **Planning**: While model-based methods remained valuable for their robustness and interpretability, learning-based approaches, particularly imitation learning and reinforcement learning, gained traction for planning tasks, offering greater adaptability in complex environments.
- **End-to-End (E2E) Pipeline**: While modular pipelines offer advantages in terms of interpretability and ease of development and debugging, E2E architectures offer the potential for higher upper-bound performance by jointly optimizing across the full AV stack.
- **Testing and Evaluation**: As AV technology advanced, safety testing and evaluation became critical for deployment readiness. AV 2.0 introduced functional safety checks and scenario-based evaluations to assess the performance and reliability of AV hardware and software components.
- **Supporting Tools**: Publicly available sensing and trajectory datasets were instrumental in benchmarking and advancing perception and planning modules. The emergence of high-fidelity simulators has provided virtual environments for training and evaluation of AV systems.

We will follow this structure to review key research efforts during the AV 2.0 era.

C. Perception

1) Early Milestones: Image-Based Perception with CNNs: The wave of deep learning-based perception systems began in 2012 with the introduction of AlexNet [57], which leveraged Convolutional Neural Networks (CNNs) for image classification. This breakthrough spurred numerous influential works that extended CNNs to object detection, semantic segmentation, and tracking tasks. Notable milestones include OverFeat [63], the R-CNN series [64]–[67], YOLO series [68]–[70], SPP [71], FCN [72], SSD [73], CenterNet [74], and Deep SORT [75]. These methods were often built on foundational architectures such as AlexNet [57], VGG-16 [76], Inception [77], and ResNet [78]. Image-based perception systems significantly improved the ability of AVs to sense and interpret their surroundings using camera inputs, achieving unprecedented scalability and accuracy compared to traditional handcrafted methods. For a comprehensive overview of image-based perception methods, refer to [79], [80].

2) Moving to 3D: LiDAR-Based Perception: LiDAR is critical for AV perception, providing accurate depth measurements and reliable performance in diverse lighting and weather conditions, complementing camera-based systems. While cameras capture rich semantic details, they struggle in low-light or adverse environments, whereas LiDAR excels in geometric precision and robustness. This synergy enables multimodal approaches that combine the strengths of both sensors. LiDAR ushered in the era of 3D perception, where methods such as PointNet [81], PointNet++ [82], VoxelNet [83], PIXOR [84], and PointPillars [85] leveraged the 3D point clouds generated by LiDAR. These approaches adapted 2D CNN concepts to 3D CNN architectures to extract meaningful features from spatial data. The complementary nature of 2D images (rich in semantics) and 3D point clouds (precise in geometry and distance) inspired multimodal perception methods. Groundbreaking works like MV3D [86], Frustum PointNet [87], ContFuse [88], and AVOD [89] combined camera and LiDAR data to achieve more holistic perception capabilities. For a detailed exploration of LiDAR-based and multimodal perception, see [90], [91].

[3]Conceptually, reinforcement learning is not part of the discriminative learning paradigm. Reviewed here due to its brief exploration during AV 2.0, despite a lack of widespread adoption.

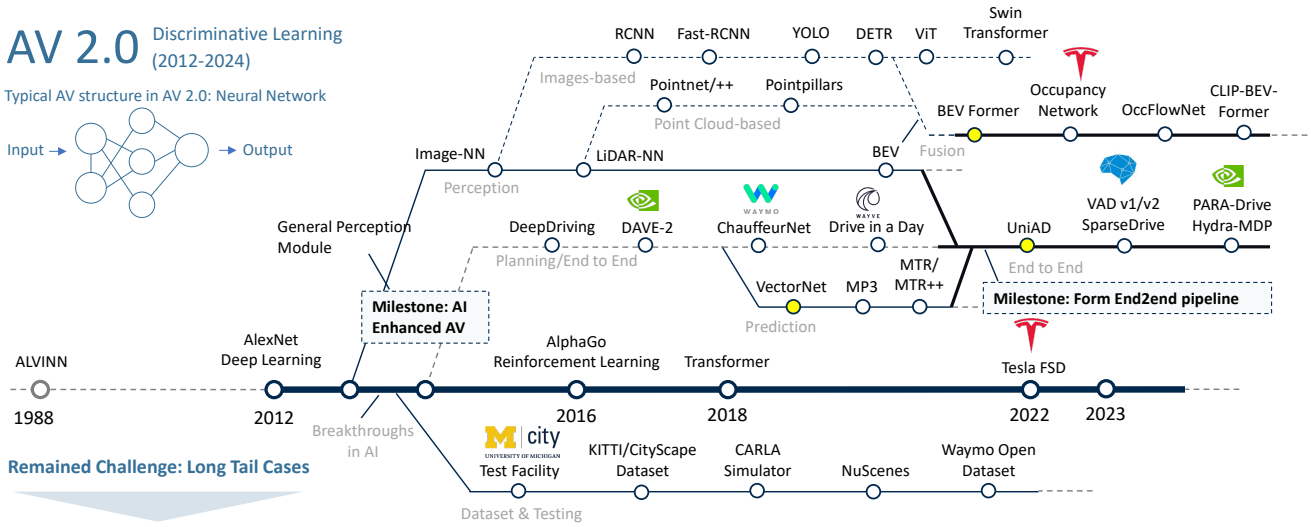


Fig. 4. AV 2.0: Autonomous Driving by Discriminative Learning

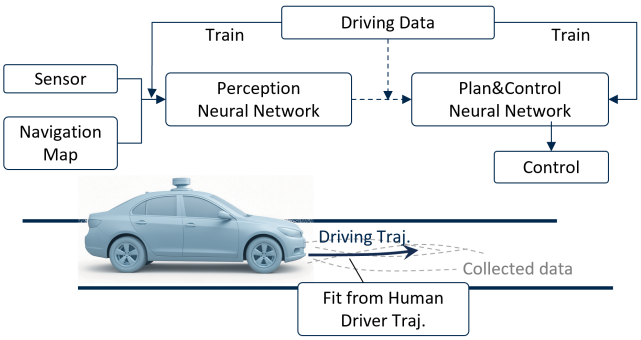


Fig. 5. A typical structure in AV 2.0: The agent is supervised using human driving data, learning to generate trajectories that closely match those demonstrated by human drivers. Perception and planning modules may be jointly learned within a unified network, commonly referred to as an end-to-end driving framework.

3) From CNN to Transformer: The Evolution of Perception Backbones: In earlier methods, CNNs dominated perception tasks, serving as the backbone for data processing and feature extraction. However, the introduction of the Transformer architecture [92] in 2017 revolutionized the field, initially transforming natural language processing and later gaining traction in computer vision. The Transformer's attention mechanism allowed for better handling of long-range dependencies and context, which CNNs struggled with due to their local receptive fields. Early applications in perception, such as DETR [93], Deformable DETR [94], and Conditional DETR [95], adopted the Transformer's encoder-decoder structure, delivering simplified architectures with fewer handcrafted components like anchor generation and non-maximum suppression (NMS) in previous method [96]. Furthermore, Vision Transformer (ViT) [97] marked a shift to pure Transformer architectures, applying Transformer principles to sequences of image patches. Over time, variants like Swin Transformer [98], Twins [99], and TNT [100] emerged, cementing the Transformer as a superior backbone for perception tasks. Representative works such as SegFormer [101] further demonstrated its capability to outperform traditional CNN-based methods in detection, segmentation, and tracking. For a detailed review of Transformers in vision tasks, refer to [96], [102].

4) Emergence of Bird's-Eye View Representations: Bird's-Eye View (BEV) representations emerged during the AV 2.0 stage as a powerful alternative to traditional front-view perception, offering benefits such as seamless sensor fusion, reduced occlusion and scale issues, and natural alignment with planning and control. Early works like PIXOR [84] and HDNET [103] laid the foundation, but BEV gained widespread traction following Tesla's adoption of BEV-based vector space construction [104]. Transformer-based models such as BEVFormer [105] and BEVFusion [106] further pushed the frontier by leveraging multi-view camera and LiDAR data.

As BEV became the mainstream perception approach, development split into two major directions: 2D-to-3D lifting and direct 3D projection. The former estimates depth and unprojects features into 3D space, as in BEVFormerV2 [107], introducing a two-stage BEV detection framework with perspective view supervision. The latter projects predefined 3D coordinate volumes onto 2D views and aggregates image features, exemplified by SparseBEV [108], which introduces a scale-adaptive attention mechanism and a spatio-temporal sampling strategy, significantly enhancing the detector's adaptability in both BEV and image space.

Recent advances also incorporate language as a complementary modality to enhance semantic understanding. BEVTSR [109] enables scene retrieval via cross-modal BEV-language embeddings, while CLIP-BEVFormer [110] applies contrastive learning to integrate language supervision into BEV features with minimal ground truth. For a complete review of BEV perception, see [111].

5) From 2D to 3D Holistic Representations: Occupancy-based perception, rooted in classic Occupancy Grid Mapping (OGM) [112], has evolved into 3D occupancy prediction—a richer scene representation that classifies voxels as free, occupied, or unobserved, often with semantic labels. Unlike BEV

or OGM, modern 3D occupancy methods remove dependence on range sensors and static scene assumptions, offering more complete spatial understanding.

Monocular and vision-centric approaches have recently advanced, including MonoScene [113], Tesla's occupancy network [114], and TPVFormer [115], achieving LiDAR-level performance using only cameras. Models like SurroundOcc [116], OccFormer [117], and FB-OCC [118] enhance multi-view fusion and spatial reasoning. Cam4DOcc [119] further extends occupancy into the spatiotemporal domain for dynamic scene prediction.

Due to the high cost of dense voxel annotations, methods such as Occ3D [120], OpenOccupancy [121], and SelfOcc [122] explore dense label generation and self-supervised learning. NeRF-based techniques like OccNerf [123], RenderOcc [124], and OccFlowNet [125] push the frontier by reconstructing 3D volumes from 2D or multi-view images with minimal supervision. For a more comprehensive overview of occupancy prediction, refer to [126].

D. Planning

The AV 2.0 stage was marked by active exploration in planning systems, with numerous approaches proposed but no definitive consensus established. The overall pipeline remained consistent with the AV 1.0 structure, comprising navigation, behavior planning, trajectory planning, and control. Among these, navigation and control saw minimal change, while significant advancements occurred in behavior and trajectory planning. In fact, the boundary between behavior and trajectory planning became increasingly blurred, and in many AV 2.0 systems, the two are treated jointly. Accordingly, we do not distinguish them explicitly in this discussion.

We categorize AV 2.0 planning methods into three major groups: *Imitation Learning*, *Reinforcement Learning*, and *Others*. The first two represent dominant learning-based paradigms: learning from expert demonstrations or through trial-and-error interaction with the environment. The *Others* category includes formal methods and hybrid approaches that integrate multiple paradigms or adopt alternative formulations.

1) Imitation Learning: Imitation Learning (IL) gained significant traction during the AV 2.0 stage, allowing AVs to learn from human driving demonstrations. Unlike model-based methods in AV 1.0, IL-based approaches train neural networks directly on real-world driving data, bypassing the need for handcrafted rules.

Modularized planning pipelines typically involve two key modules: trajectory prediction and planning. In trajectory prediction, IL-based methods learn to forecast the behavior of surrounding agents—such as vehicles, pedestrians, and cyclists—to enable safe and effective planning. By leveraging large-scale driving datasets, these methods use supervised learning to map observed states to future trajectories. Notable advancements included Social LSTM [127], MultiPath [128], Multipath++ [129], VectorNet [130], TNT [131], DenseTNT [132], MotionLM [133], MTR [134], MTR++ [135], and QCNet [136]. VectorNet [130], in particular, introduced a vectorized representation of road networks, which was later widely used by other methods, improving model efficiency and scalability.

Once trajectory prediction has been completed, the planning module leverages the predicted agent behaviors to plan the ego vehicle's trajectory by mimicking human driving behaviors. During training, the system incorporates additional loss terms related to collision avoidance, off-road deviations, and map adherence, in addition to the standard IL imitation loss, to optimize the future trajectory. However, the distinction between trajectory prediction and planning is not always clear-cut. Many approaches, such as ChauffeurNet [137], LookOut [138], PiP [139], MP3 [140], and DIPP [141], jointly predict both the behavior of surrounding agents and the ego vehicle's future trajectory. While IL has enabled significant advances, it is still hampered by challenges such as compounding errors [142], which can lead to the degradation of performance in AV sequential decision-making processes.

2) Reinforcement Learning: The rise of reinforcement learning (RL) during the AV 2.0 stage brought a promising new direction for the AV planning. Unlike IL, which relies on expert demonstrations, RL empowers agents to autonomously explore environments and learn through trial and error, enabling AVs to potentially achieve superhuman performance in complex and dynamic driving tasks. Inspired by RL's success in Atari games [143], AlphaGo [58], and AlphaGo Zero [144], extensive research [145]–[152] has explored RL applications to AV planning using value-based, policy-based, and actor-critic methods in both simulated and real-world environments. Notable studies, such as [146], developed RL frameworks for controlling vehicles in simulators, while [147] applied Deep Deterministic Policy Gradient (DDPG) to control a full-sized robotic vehicle in real-world environments.

While RL has demonstrated promising results, its real-world deployment faces challenges such as data inefficiency, the sim-to-real gap, and the lack of theoretical safety guarantees. To address these issues, recent developments in safe RL [153] and confidence-aware RL [154] have focused on ensuring continuous performance improvement and safety in dynamic environments. A comprehensive survey on RL applications in AV planning can be found in [155].

3) Others: Formal methods have also been explored as a means to develop safe and theoretically sound planning algorithms. These methods focus on ensuring safety by providing guarantees about AV behavior. Reachability-based methods [156] calculate the reachable sets of surrounding agents, allowing formal verification of the AV's planned trajectories. However, these methods tend to be conservative, often over-approximating reachable sets, and they struggle to scale when the number of surrounding agents increases. Similarly, safety-envelope methods, such as Responsibility-Sensitive Safety (RSS) [157], rely on assumptions about the behavior of surrounding agents and construct safety layers to ensure safe operation of the AV. While these methods provide formal safety guarantees, their reliance on predefined rules and assumptions can limit their applicability in dynamic, real-world environments.

Hybrid methods have also garnered attention, combining the strengths of different approaches to address the limi-

tations of individual methods. For instance, methods that integrate model-based planners with RL [158] combine the interpretability and robustness of model-based methods with the flexibility of RL, enabling more adaptable planning in complex environments. Hybrid approaches that combine IL with RL [159], [160] exploit IL for pretraining and RL for fine-tuning, achieving better generalization in dynamic situations. Furthermore, Generative Adversarial Imitation Learning (GAIL), which combines IL with Generative Adversarial Network (GAN), has been applied to model human driving behaviors [161]–[164]. These hybrid strategies seek to balance interpretability, robustness, and performance, allowing for more effective and scalable solutions in real-world autonomous driving tasks.

To further enhance the robustness of AV decision-making, efforts are increasingly directed toward improving system transparency and interpretability, making it easier to debug and optimize the system. One approach is attention visualization [165], where the system highlights the most relevant elements to its decision-making process, offering a visual insight into its reasoning. Another strategy involves designing interpretable tasks [166] that decode latent feature representations into meaningful insights like reconstructing input characteristics. Furthermore, some models generate explanations in natural language [167], [168] for their predictions, making the underlying processes more accessible to human understanding.

E. E2E Pipeline

End-to-End (E2E) pipelines fundamentally differ from traditional modular architectures, which design perception, planning, and control components independently. The core idea of E2E systems is to jointly optimize the entire autonomous driving stack, either through a fully differentiable modular system or as a complete black-box model, holistically trained using data-driven approaches such as IL. This unified training framework enables integrated task optimization and allows for seamless information flow across components.

Early E2E approaches demonstrated the feasibility of directly mapping raw sensor inputs to control commands using deep neural networks. Notably, Nvidia's DAVE-2 [169] and other contemporaneous systems [147], [170], [171] employed CNNs to predict steering commands from front-facing camera images, achieving short-range autonomous driving in real-world settings. These pioneering efforts showcased the intuitive appeal of E2E learning by removing the need for hand-engineered intermediate representations and separate modules.

As the field evolved, researchers began integrating modular principles into E2E systems while preserving end-to-end trainability. For instance, UniAD [172] introduced a differentiable modular design where perception, prediction, and planning components were jointly optimized using auxiliary IL losses. This approach maintained the interpretability and structure of modular pipelines while benefiting from unified learning and optimization.

Building on this foundation, recent systems further enhanced performance and generalization through architectural innovations. VAD [173] introduced a fully vectorized scene representation, avoiding rasterized inputs and post-processing, thus improving both interpretability and accuracy. OccNet [174] employed a multi-view, vision-centric architecture leveraging occupancy maps and object detection outputs to improve planning accuracy. VADv2 [175] modeled the planning policy as an environment-conditioned, nonstationary stochastic process, using a probabilistic field to map the action space to probability distributions and significantly improving closed-loop performance. SparseDrive [176] addressed the inefficiencies of dense BEV representations by introducing a sparse perception module and a parallel planner with hierarchical decision layers, achieving notable gains in safety and computational efficiency.

Despite these advancements, challenges remain. Most current E2E systems lack safety guarantees and struggle to generalize reliably in complex driving scenarios. In particular, probabilistic planning under uncertainty with safety constraints is still an open problem in end-to-end autonomous driving.

F. Testing and Validation

Testing and validation are critical phases in the development lifecycle of AV, essential for AV developers, customers, and regulators. In AV 2.0, testing and validation are focused on two main aspects: functional safety evaluation and scenario-based testing. The primary goal was to identify and address potential failures before deployment, refining system reliability and safety. The complexity of AV systems significantly amplifies the challenge of validation. Modern AVs can incorporate up to 100 million lines of code—far surpassing the 14 million lines in a Boeing 787 aircraft [177]. Validating the safety of such intricate systems requires advanced methodologies and robust testing frameworks [178], [179].

1) Functional Safety Evaluation: Traditional automotive testing programs, such as the New Car Assessment Program (NCAP) [180], initially focused on vehicle crashworthiness and have recently expanded to include Advanced Driver Assistance Systems (ADAS) features like Automatic Emergency Braking (AEB). For higher-level AVs, the emphasis shifted toward functional safety, guided by standards such as ISO 26262 [181] for Functional Safety and ISO 21448 [182] for Safety of the Intended Functionality (SOTIF). The objective is to ensure the absence of unreasonable risks that arise from hardware or software malfunctions or insufficiencies, for example, a malfunctioning sensor (FuSa), or a miss detection of objects (SOTIF). Functional safety evaluation adopts a vehicle-centric view, examining the reliability, robustness, and sufficiency of AV hardware and software components. Additionally, UL 4600 [183] offers structured guidelines for building comprehensive safety cases tailored to AV systems.

2) Scenario-based Testing: To assess system-level safety, scenario-based testing emerged as a major focus in AV validation. This approach involves executing a wide range of testing scenarios, from simple to complex, often within controlled environments, to evaluate AV safety performance. Recent standards like ISO 34502 [184] provide high-level frameworks to formalize scenario-based evaluation. Generating effective test scenarios is a key challenge. Early research employed combinatorial techniques to manage the vast scenario space

[185], while later works proposed adaptive sampling [186] and worst-case scenario generation [187] to probe AV performance limits. To accelerate testing, advanced scenario-generation techniques have been developed to efficiently identify critical scenarios that stress AV systems [33], [188], [189]. The scenario-based testing can be conducted in either simulations or closed testing facilities like Mcity [190].

However, scenario-based testing has limitations. Most scenarios involve only a few dynamic agents and short time horizons, which may not capture the complexity of real-world driving. Moreover, successful performance in predefined scenarios does not guarantee the absence of failures in untested conditions, as comprehensive coverage remains difficult to achieve. Another complementary line of research involves formal verification, which seeks to provide mathematically rigorous safety guarantees. While formal methods offer strong theoretical assurances, they often face scalability challenges in with complex systems and are constrained by assumptions about the correctness of sensor and perception subsystems [191].

G. Supporting tools

The transition to AV 2.0, characterized by data-driven methodologies, was facilitated by the development of various supporting tools. These tools enabled the shift from rule-based systems to data-centric approaches by addressing critical needs in data accessibility, benchmarking, and simulation. Below, we discuss the major supporting tools and their contributions to the AV ecosystem.

1) Datasets: A key driving factor in the AV 2.0 stage was the emergence of large-scale, high-quality, real-world datasets from diverse sources. These datasets provided critical human-labeled data and demonstrations for training data-driven models. The first wave of sensing datasets, such as KITTI [59], [60], paved the way for tasks like detection, tracking, and segmentation. Subsequent datasets, including Cityscapes [192], BDD100K [193], and ApolloScape [194], expanded the scope of data available for AV perception.

Major AV companies also contributed significantly by releasing datasets like Argoverse [195], [196], nuScenes [197] and nuPlan [198], Waymo Open Dataset [199], [200], and Lyft L5 Dataset [201]. These datasets, captured by fleets of vehicles equipped with advanced sensors, incorporated not only raw sensing data but also high-definition maps and the trajectories of surrounding traffic participants, enabling advancements in AV development.

Infrastructure-based datasets, collected using mounted cameras or drones, provided a complementary perspective to vehicle-based datasets. Examples include NGSIM [202], INTERACTION [203], highD [204], rounD [205], inD [206], MSight [207], CitySim [208], I-24 Motion [209], and pNEUMA [210]. These datasets offered richer contextual information about the AV surrounding environment, capturing interactions among multiple road users from a broader vantage point. However, infrastructure-based datasets were typically limited in size, often spanning only a few hours. In contrast, vehicle-based datasets released by companies were significantly larger, often comprising hundreds or even thousands of hours of data. For a more detailed discussion and comprehensive overview of these datasets, see surveys such as [211], [212].

2) Benchmarks: Benchmarks and challenges were established to promote fair comparisons and consistent improvements in AV systems. Notable initiatives included the KITTI Vision Benchmark Suite [59] for perception tasks and the Waymo Open Data Challenges [213], [214], which span perception, planning, and simulation domains. The nuPlan Planning Challenge [215] introduced benchmarks for both open-loop and closed-loop planning evaluation. Other notable benchmarks and challenges, such as the CARLA Challenge [216], Autonomous Grand Challenge [217], and Mcity AV Challenge [218], fostered innovation by providing transparent leaderboards and encouraging the development of robust AV solutions. These events not only highlighted the technological progress in AV systems but also helped to align industry and academic efforts towards common goals.

3) Simulation: Simulation platforms played a critical role in complementing real-world datasets by providing synthetic data and virtual environments for AV training and testing. High-fidelity simulators, such as CARLA [219], AirSim [220], and LGSVL [221], offered full-stack simulation systems, including sensor simulation, traffic interaction modeling, and vehicle dynamics. Among these, CARLA stood out as a leading open-source platform, extensively used for developing perception and decision-making modules, as well as training E2E learning-based driving systems, due to its comprehensive functionality and rich customization capabilities. Other simulators, such as TORCS [222], SUMO [223], HighwayEnv [224], Gazebo [225], CarSim [226], and MetaDrive [227], specialized in specific functionalities, offering focused solutions for particular aspects of AV development. Some of these simulators can be co-simulated with comprehensive platforms like CARLA, enabling more versatile and detailed simulation setups. The effectiveness of simulators depended heavily on their fidelity, with significant research devoted to improving sensor [228]–[230] and traffic simulation [231]–[233] fidelity.

H. Discussions

AV 2.0 represents over a decade of rapid progress, fueled by advancements in deep learning and the availability of large-scale real-world datasets. This era fundamentally transformed perception, shifting from handcrafted feature extraction to neural network-based solutions. The widespread adoption of BEV representations and Transformer-based architectures in recent years reflects a growing convergence toward standardized design patterns across the field. However, decision-making remained fragmented, with no clear consensus between model-based and learning-based approaches. Meanwhile, control modules largely continued to rely on traditional model-based techniques. The emergence of E2E pipelines during this stage further exemplified the momentum toward holistic learning systems, with unified architectures gaining attention for their conceptual elegance and promising performance. At the same time, commercial products matured from

experimental prototypes to real-world applications. Notable examples include Tesla's Full Self-Driving (FSD), a Level 2 system, and Waymo, which operates at a higher autonomy level and has launched Robotaxi services in several cities in the United States. Similarly, Baidu's Apollo Go expanded Robotaxi services across multiple cities in China, signaling broader deployment efforts.

Despite the significant progress of AV 2.0, current systems still struggle to handle long-tail events, particularly those that fall outside the training distribution. These limitations have highlighted the need for for AV 3.0: a new paradigm that leverages generative AI to enable safe autonomy at scale.

V. Towards AV 3.0: Autonomous Driving through Generative Learning

A. Technical challenges in AV 2.0

Under the data-driven paradigm of AV 2.0, autonomous vehicles have made remarkable progress. However, we find some inherent limitations of AV 2.0 as follows:

a) Limited training data v.s. Infinite driving scenarios: AV 2.0 relies on collecting data to cover as many driving scenarios as possible. However, real-world driving scenarios are actually infinite due to the diverse and stochastic nature of driving tasks. This intrinsic complexity imposes fundamental limitations on the generalization capabilities of data-driven autonomous driving systems.

b) Statistical model training v.s. Rare scenario safety: AV 2.0 models are typically biased toward high-frequency scenarios during training, since these contribute more significantly to the optimization of the training objective. However, rare scenarios are often the primary contributors to safety-critical failures in autonomous driving. Relying solely on naturally collected driving data cannot overcome the "curse of rarity" problem.

c) Discrete data v.s. Continuous space: Even in common scenarios with dense training data, AV 2.0 systems may still fail, as the training data consist of independent, discrete samples that cannot fully cover the continuous space of real-world driving scenarios. The performance of AV systems cannot be reliably guaranteed when real-world scenarios deviate even slightly from the training data.

We believe that the root cause of these limitations lies in the discriminative learning paradigm that defines AV 2.0. In detail, most AV 2.0 systems are trained to predict actions that closely match the ground truth, effectively learning to estimate and reconstruct the conditional distribution of outputs given inputs, i.e., $p(y|x)$. This formulation, which conditions on x, treats the training data as independent and discrete samples, without capturing any inherent relationships between them. Thus, such an approach ensures performance only in trained cases. Thus, AV 2.0 models must rely on the collection of exhaustive datasets and large-scale road testing to compensate for the lack of generalization and to ensure safety.

B. Vision and Paradigm for AV 3.0

Fundamentally addressing these limitations requires a paradigm shift in autonomous driving. We thus introduce AV 3.0, a new generation of autonomous driving systems grounded in the generative learning paradigm.

The most significant shift in AV 3.0's objective is that it no longer trains on isolated cases, but rather learns the inherent relationships across driving scenarios, enabling scalable and trustworthy autonomous driving. It adopts the generative learning paradigm, which directly models the joint distribution $p(x, y)$, thereby capturing both the mapping from inputs to outputs and the underlying dependencies between them. These inherent relationships may capture physical characteristics, stable contextual dependencies, or spatial structures. In other words, AV 3.0 should be capable of understanding these relationships and reasoning about how they influence the final driving actions.

The shift to the AV 3.0 paradigm offers significant benefits. Unlike traditional approaches that rely heavily on extensive data collection, AV 3.0 leverages learned internal relationships to enable extrapolation from limited demonstration examples, allowing the system to scale across a wide range of similar scenarios. Moreover, these structured internal relationships support continual learning by preserving previously acquired driving knowledge. As a result, the system can integrate new experiences and continuously adapt to novel scenarios without degrading prior performance. Crucially, AV 3.0 endows the agent with a key capability—predicting and navigating previously unseen scenarios while consistently maintaining reliable collision avoidance, especially under high-risk and time-critical conditions. These features position AV 3.0 as a promising paradigm for enabling inherently scalable and safe autonomous driving.

Recent studies have started exploring AV 3.0 paradigm, representing initial progress toward generative and relationship-aware autonomy. Fig. 6 summarizes the recent works, and Fig. 7 shows a typical structure in AV 3.0. The following section will review these studies.

C. Enhancing AV system with LLMs/VLMs/MLLMs

LLMs, VLMs, and broader multimodal large language models (MLLMs) have demonstrated the ability to capture stable inherent relationships within natural language and across modalities—particularly between language and vision—often referred to as commonsense knowledge. Hereafter, we use MLLMs as a unified term for LLMs, VLMs, and multimodal language models, unless stated otherwise. Leveraging these stable relationships in the context of autonomous driving has the potential to significantly enhance AV performance in unseen scenarios, aligning closely with the vision of AV 3.0.

Firstly, these pre-trained models can be utilized to describe diverse driving conditions [234], [235], interpret traffic rules [236], and represent ongoing contextualized events [237]. They serve as intermediaries to translate high-level semantic information into actionable commands that autonomous driving systems can interpret, thereby enabling AVs to handle scenarios that were not explicitly considered during system design.

These models and their associated modeling techniques can also be integrated into perception [238], motion prediction [133], scenario formulation [239], behavior planning

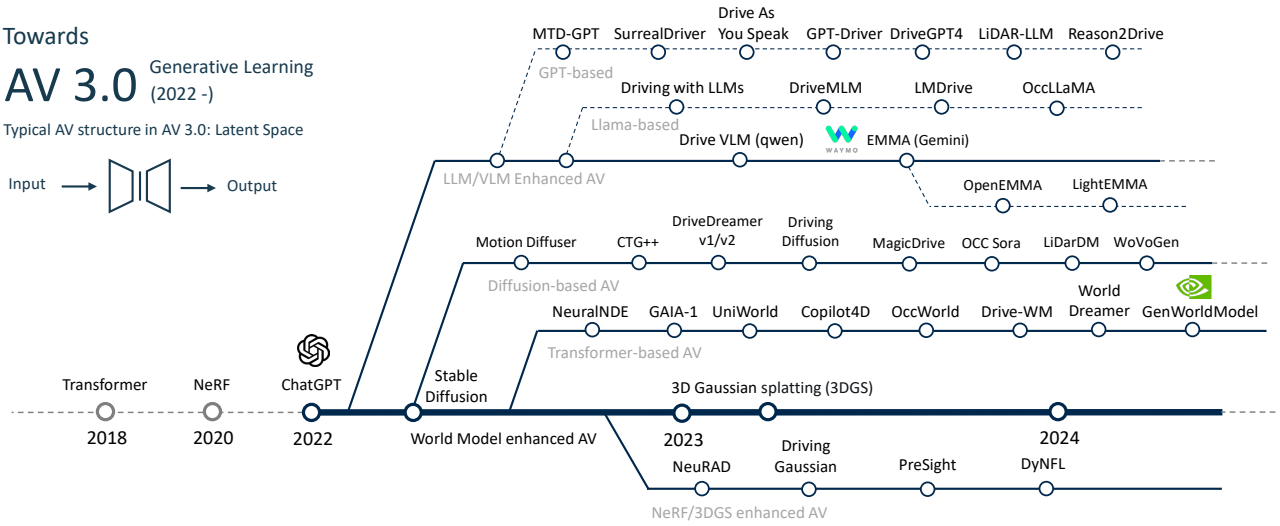


Fig. 6. Towards AV 3.0: Autonomous Driving by Generative Learning

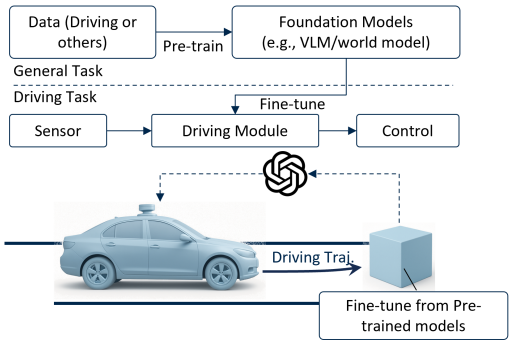


Fig. 7. A typical structure in AV 3.0: The agent is equipped with a pre-trained foundation model (e.g., a vision-language model), which can be fine-tuned or adapted for specific autonomous driving tasks.

modules [240], and driver interaction [241] within autonomous driving systems. In most of these approaches, the module-specific inputs are first converted into language-like representations, allowing LLMs to perform reasoning or generation in the linguistic space. The outputs are then transformed back into the original modality or control format, enabling compatible integration with existing AV system components. Some works further distill the outputs of LLMs to improve the performance of downstream modules and overall driving capabilities [242].

Finally, these models can also be utilized during the training process of autonomous driving systems—for example, to support critical data selection [243] or to assist with object annotation [244]. However, these approaches remain heavily constrained by the limitations of the pre-trained models themselves. In particular, the hallucination issue inherent to large models raises significant safety concerns. In our own study [245], we evaluated the driving-related capabilities of 12 different VLMs and found that their current performance still falls short of practical deployment requirements.

D. AV Agents Built upon MLLMs

Instead of simply incorporating MLLMs into existing AV architectures, some approaches aim to retrain or adapt MLLMs as the foundation for building autonomous driving agents. A key distinction from the previous section is that the MLLM is now assigned the role of making final control decisions, moving beyond its earlier use as an auxiliary module for perception or semantic reasoning.

For instance, DriveGPT4 [246], DriveMLM [247], and LanguageMPC [248] enable the MLLMs to analyze scenes and predict planning actions from a predefined action set. MTD-GPT [249] formulates the decision-making process as a sequence modeling task, while [250], [251] leverage driving data along with advanced MLLMs for scene understanding and evaluation via question-answering (QA) task. GPT-Driver [252] takes a different approach by framing motion planning as a language modeling problem, offering a token-based interpretation of the planning process from the perspective of GPT-style models.

Driving with LLMs [253], proposed by Wayve, introduces an architecture that embeds vectorized driving inputs into an LLM, combined with a two-stage pretraining and fine-tuning strategy to enhance driving-specific reasoning and control. Furthermore, models such as DriveVLM [254], which fine-tunes QWen-VL [255], and EMMA [256], which fine-tunes Gemini [257], along with the follow-up works OpenEMMA [258] and LightEMMA [245], demonstrate the feasibility of adapting general-purpose MLLMs for driving tasks. DriveMLM [247] and LMDrive [259] also fine-tune LLaMA [260] as a backbone, incorporating an additional vision encoder to improve driving performance.

Additionally, some works optimized the MLLM-based AV agents for higher computational efficiency, e.g., DriveVLM [254] and Senna [240]. Some novel datasets and benchmarks are also proposed to further enhance the reasoning capacity of VLMs in the driving domain [261], [262]. LANGPROP [263] applies LLMs to optimize code within driving applications

to improve the computational efficiency and performance of autonomous driving algorithms.

However, it should be acknowledged that while these works explore various ways to apply MLLMs to autonomous driving tasks, they fall short of fundamentally addressing the core challenges of autonomous driving. Moreover, there is still a lack of consensus on the precise role that MLLMs should play within the autonomous driving framework.

E. World Model-based AV

Another line of exploration involves training models to capture the a compact, structured representation of environment dynamics that supports planning and control in AV systems, referred to as world models [264]. These world models are often combined with reinforcement learning [265] or model predictive control (MPC) [266] to enhance scalability and generalization in unfamiliar scenarios.

World models can be constructed through various modalities, including video prediction [267]–[269], point cloud prediction [270], occupancy prediction [271], [272], and abstract state prediction [273]. OccLLaMA [274] extends traditional occupancy prediction by incorporating 4D scene understanding, temporal reasoning, and motion planning, trained on large-scale real-world datasets. By leveraging the generative capacity of world models, the generalization capability of AV systems in handling previously unseen scenarios has been significantly enhanced. For example, DriveDreamer2 [275] integrates LLMs with diffusion models to generate controllable driving scene videos from natural language inputs—including rare events such as sudden overtaking—and achieves over a 4% improvement in perception performance. In addition, several studies have explored integrating world models with RL [276], MPC [277], and continual learning [154] to enhance an AV agent's ability to handle corner cases. These approaches leverage imagination-based rollouts from world models to improve sample efficiency, long-horizon reasoning, and adaptability in complex or rare driving scenarios.

Some works have explored the design of generative architectures for building world models. Most of these approaches are based on diffusion models [278]–[284] or Transformer-based models [285]–[287]. These works improve the accuracy and realism of world models from different perspectives, such as spatial-temporal consistency, physical plausibility, or controllable generation.

In contrast to generic world models, a subset of research targets autonomous driving–specific world models, which primarily focus on capturing and predicting the behavior of surrounding agents. Notable examples include NeuralNDE [164], MTR/++ [134], [135], SMART [288], MotionDiffuser [289] and CTG++ [290]. These works focus directly on modeling the behavior of agents that have the greatest impact on autonomous driving, while intentionally ignoring irrelevant scene details such as roadside trees or lighting variations. As a result, they offer a more direct and efficient world model to improve autonomous driving performance.

Another subset of research focuses on constructing highly accurate 3D representations of the environment, aiming to capture as much spatial detail as possible in order to reconstruct every element in the scene. A representative work is Neural Radiance Fields (NeRF) [291] and its derivatives [292]–[296], which aim to reconstruct photorealistic 3D environments. The other is based on 3D Gaussian Splatting [297], with follow-up work [298]–[300] improving real-time rendering and geometry fidelity for dynamic scenes. Yet, both NeRF and 3DGS approaches tend to offer limited diversity and require considerable effort to produce multi-camera videos that align with real-world lighting, weather conditions, and other factors. Generalizing to entirely novel views remains challenging.

Despite recent progress, the application of world models to autonomous driving still faces a significant gap: many existing efforts remain closely tied to traditional perception and motion prediction paradigms. Such approaches often carry over the design mindset of modular pipelines or adopt end-to-end formulations that integrate perception, prediction, and control into a single, monolithic structure. However, there has been limited investigation into two critical challenges following the concept of world models: (1) how to transform multi-modal sensory inputs into a unified and compatible latent representation, and (2) how such latent representations can be effectively leveraged to guide decision-making and action generation. As previously emphasized, world models are not simply predictive tools—they are integral components of broader decision-making frameworks. From this perspective, the development of world models for autonomous driving remains in its infancy. A world model that can reliably handle the complexity, uncertainty, and real-time demands of autonomous driving has not yet been realized.

F. Testing and evaluation with AV 3.0

The advent of AV 3.0 marks a pivotal shift in AV development, aiming to achieve safe autonomy at scale. This progression necessitates a transformation in testing and evaluation methodologies. During the AV 2.0 era, safety assessments were predominantly vehicle-centric, focusing on functional safety and scenario-based testing to verify individual vehicle responses under controlled conditions. While effective for isolated cases, these methods are insufficient for evaluating the complex interactions and systemic behaviors critical for large-scale deployment. AV 3.0 introduces a paradigm shift towards evaluating *behavioral safety* through *environment-based testing*. This approach emphasizes assessing how autonomous vehicles interact within dynamic and unpredictable real-world environments, accounting for the behaviors of other road users and varying traffic conditions. By focusing on the broader context, environment-based testing aims to ensure that AVs can safely and reliably operate across diverse scenarios, thereby supporting scalable and trustworthy autonomous driving.

1) Behavioral Safety Evaluation: Behavioral safety evaluation focuses on assessing the AV's capability to make appropriate decisions when the system is functionally sufficient and error-free. During the evaluation, the AV will be treated as a black-box system, and the evaluation examines both the AV's own safety performance and its influence on the overall safety impact of the traffic environment. For instance, if another vehicle suddenly cuts into the AV's lane, behavioral

safety evaluation assesses whether the AV reacts appropriately (e.g., braking or lane changing) without triggering secondary conflicts, such as rear-end collisions. Behavioral safety also inherently encompasses evaluation for proactive risk mitigation, such as avoiding prolonged driving in blind spots to reduce cut-in risks. The critical role of behavioral safety evaluation is enabling real-world AV readiness at scale, which is also emphasized in the CertiCAV Assurance Paper by Connected Places Catapult of the United Kingdom [301].

2) Environment-based Testing: Behavioral safety must be assessed in dynamic, closed-loop environments with continuous agent interactions over long time horizons. This *environment-based testing* enables observation of emergent behaviors and the collection of statistical safety measurements (e.g., crash rates). While real-world testing is valuable, it is costly and time-consuming. In contrast, simulation offers scalability, safety, repeatability, and cost-efficiency. Recent advances in generative simulation such as TeraSim [164], [302], GAIA-1 and 2 [286], [303], SMART [288], CAT-K [304], and OMNIVERSE [305] provide promising tools for constructing rich virtual environments.

However, rare safety-critical events make mileage-based brute-force testing impractical, even in simulation—it may take billions of miles to observe sufficient events [306]. To address this, recent work has explored accelerated testing environments [307], [308], which significantly improve efficiency while preserving evaluation unbiasedness. Nevertheless, simulation fidelity, especially in sensor modeling and agent behavior, remains a critical challenge for bridging the sim-to-real gap and ensuring trustworthy results. Additionally, as AV systems grow more complex, maintaining the coverage and adaptability of accelerated testing environments is an ongoing research frontier.

A major advantage of environment-based testing is its ability to uncover unknown unsafe events. When AVs are evaluated in spatiotemporally continuous traffic environments, they are exposed to a virtually infinite range of situations, making it possible to uncover unknown system deficiencies that are difficult to detect through predefined, fragmented, and short-horizon scenario-based tests. Importantly, behavioral safety evaluation is not limited to single AV testing but extends to evaluations involving multiple AVs operating at scale. It captures a diverse range of interactions between AVs and background vehicles (BVs), including AV-AV, AV-BV, and BV-BV dynamics. This comprehensive approach enables a systemic assessment of the safety impacts AVs may have across the broader transportation ecosystem.

G. Future Direction

Despite the promising potential of these new technologies, a clear technical roadmap for fundamentally addressing the safety and scalability of autonomous vehicles has yet to be established. In the following, we outline several ideas and proposals under the new generative learning paradigm.

a) Autonomous Driving Foundation Models built upon LLM/VLM: LLMs and VLMs have demonstrated great success as generative models in natural language and vision-language processing, and their underlying principles align well with the objectives of AV 3.0. However, autonomous driving goes beyond a language task; it requires not only contextual reasoning, but also spatial reasoning, dynamics awareness, and more. Therefore, a generative foundation model for autonomous vehicles should not only leverage LLMs or VLMs, but also introduce a novel modality specifically designed for driving tasks. Research into what structures are most suitable for modeling driving behaviors, and how to effectively train a unified Vision-Language-Driving foundation model, remains an important direction for future exploration.

b) Physical Encoded World Models: World models naturally fit the AV 3.0 paradigm if they are able to capture stable relationships across driving scenarios. However, existing world models remain insufficient for safety-critical applications. To meaningfully enhance the safety and scalability of AV systems, world models must explicitly encode physical constraints, dynamics, and environmental interactions. Therefore, investigating how to incorporate physical laws into world models, or how to construct high-fidelity world models that accurately reflect real-world complexities, represents a critical direction for realizing the vision of AV 3.0.

c) Driving-Task-Oriented Novel Generative Model: LLMs and world models were not originally developed for driving tasks. As a result, they naturally require further adaptation, and their final performance in autonomous driving remains uncertain. However, the driving task itself possesses many unique characteristics—for example, the need for spatial-behavior relationship understanding, and the inherent influence of traffic rules and signals on behavior. These relationships are difficult to capture effectively using general-purpose learning architectures. Therefore, a promising direction in AV 3.0 is to develop new generative model structures specifically aligned with the nature of driving tasks, enabling more targeted and efficient training for autonomous vehicles.

VI. Conclusion

In this paper, we have presented a comprehensive overview of the technological evolution of AV systems over the past two decades. To systematically understand the paradigm shifts, we first analyzed the two fundamental driving forces behind AV development: safety and scalability. The ultimate goal for AVs is to build intelligent systems capable of safely operating across diverse ODDs with minimal adaptation effort, achieving superior safety performance compared to human drivers and improving overall transportation system safety. However, AV development has proven far more challenging than initially anticipated, facing two key compounding difficulties: the CoD, arising from the inherent complexity of real-world environments, and the CoR, stemming from the long-tail distribution of rare but critical events. Trade-offs between safety and scalability have driven continuous innovation and methodological evolution throughout the past two decades.

To address these challenges, the AV research community has undergone major methodological paradigm shifts. We categorized these into three primary phases: AV 1.0 (Autonomous Driving by Design), AV 2.0 (Autonomous Driving

through Discriminative Learning), and AV 3.0 (Autonomous Driving through Generative Learning). AV 1.0 marked the early stage, where rule-based and heuristic-driven methods were employed for initial trials. AV 2.0 emerged with the rise of deep learning, switching the development philosophy to data-driven approaches. This transition enabled the development of modular AV systems trained on vast real-world datasets and sparked interest in end-to-end architectures aimed at building more intelligent and scalable systems. Throughout these phases, we have highlighted key milestones, offering a clear understanding of past advancements and providing a valuable reference for readers.

Today, we are at a critical inflection point, transitioning toward the next phase of AV development. Looking ahead, we believe that the next major shift will be the move from discriminative imitation to generative learning, heralding the era of AV 3.0. For AVs to reach their full potential, they must develop true driving intelligence—encompassing understanding, reasoning, and decision-making capabilities—rather than relying solely on imitation and memorization from training data. The challenge is that, regardless of how large the training dataset is, it remains finite, discrete, and subject to long-tail distribution, while real-world driving incorporates an infinite number of potential situations. This renders purely imitative approaches inherently insufficient. To address this, a paradigm shift is needed—one that moves from discriminative modeling to generative modeling, enabling AVs to learn the inherent relationships between scene context, agent behaviors, and driving outcomes. Early explorations into generative models, such as LLMs/VLMs and world models, mark the beginning of this transformation. We envision that the development of autonomous driving foundation models, physical encoded world models, driving-task-oriented generative models, and behavioral safety assessment frameworks will play pivotal roles in this transformative shift.

In conclusion, we hope this paper provides a comprehensive roadmap for understanding the evolution of AV technologies, identifies the key challenges, and outlines promising future directions. By reflecting on the past and critically examining ongoing challenges, we aim to inspire future research that will overcome current limitations and pave the way for fully autonomous, safe, and scalable AV systems.

ACKNOWLEDGMENT

This research was partially funded by the U.S. National Science Foundation through the Mcity 2.0 Project (CMMI #2223517). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the official policy or position of the U.S. government.

The authors would like to thank Dr. Boqi Li, Dr. Jiawei Wang, and Greg Stevens from the University of Michigan, Ann Arbor, for their valuable comments and suggestions, which contributed to improving the quality of this paper.

REFERENCES

[1] Waymo, 2024. [Online]. Available: https://x.com/Waymo/status/1851365483972538407

[2] Baidu, 2024. [Online]. Available: https://ir.baidu.com/news-releases/news-release-details/baidu-announces-second-quarter-2024-results

[3] T. Thadani, "Cruise recalls all its driverless cars after pedestrian hit and dragged," https://www.washingtonpost.com/technology/2023/11/08/cruise-crash-driverless-recall/, Nov. 2023, The Washington Post.

[4] B. Moye, "AAA: Fear in Self-Driving Vehicles Persists," https://newsroom.aaa.com/2025/02/aaa-fear-in-self-driving-vehicles-persists/, Feb. 2025, AAA Newsroom.

[5] S. Singh, "Critical reasons for crashes investigated in the national motor vehicle crash causation survey," Tech. Rep., 2015.

[6] H. X. Liu and S. Feng, "Curse of rarity for autonomous vehicles," *nature communications*, vol. 15, no. 1, p. 4808, 2024.

[7] M. Abdel-Aty and S. Ding, "A matched case-control analysis of autonomous vs human-driven vehicle accidents," *Nature communications*, vol. 15, no. 1, p. 4931, 2024.

[8] K. Bengler, K. Dietmayer, B. Farber, M. Maurer, C. Stiller, and H. Winner, "Three decades of driver assistance systems: Review and future perspectives," *IEEE Intelligent transportation systems magazine*, vol. 6, no. 4, pp. 6–22, 2014.

[9] S. E. Shladover, "Path at 20—history and major milestones," *IEEE Transactions on intelligent transportation systems*, vol. 8, no. 4, pp. 584–592, 2007.

[10] M. Lu, K. Wevers, and R. Van Der Heijden, "Technical feasibility of advanced driver assistance systems (adas) for road traffic safety," *Transportation Planning and Technology*, vol. 28, no. 3, pp. 167–187, 2005.

[11] A. Khodayari, A. Ghaffari, S. Ameli, and J. Flahatgar, "A historical review on lateral and longitudinal control of autonomous vehicle motions," in *2010 International Conference on Mechanical and Electrical Technology*. IEEE, 2010, pp. 421–429.

[12] D. A. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network," *Advances in neural information processing systems*, vol. 1, 1988.

[13] M. Buehler, K. Iagnemma, and S. Singh, *The DARPA urban challenge: autonomous vehicles in city traffic*. Springer Science & Business Media, 2009, vol. 56.

[14] M. Montemerlo, J. Becker, S. Bhat, H. Dahlkamp, D. Dolgov, S. Ettinger, D. Haehnel, T. Hilden, G. Hoffmann, B. Huhnke *et al.*, "Junior: The stanford entry in the urban challenge," *Journal of field Robotics*, vol. 25, no. 9, pp. 569–597, 2008.

[15] Y. Kuwata, G. A. Fiore, J. Teo, E. Frazzoli, and J. P. How, "Motion planning for urban driving using rrt," in *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2008, pp. 1681–1686.

[16] C. Urmson, J. Anhalt, D. Bagnell, C. Baker, R. Bittner, M. Clark, J. Dolan, D. Duggins, T. Galatali, C. Geyer *et al.*, "Autonomous driving in urban environments: Boss and the urban challenge," *Journal of field Robotics*, vol. 25, no. 8, pp. 425–466, 2008.

[17] D. Yang, K. Jiang, D. Zhao, C. Yu, Z. Cao, S. Xie, Z. Xiao, X. Jiao, S. Wang, and K. Zhang, "Intelligent and connected vehicles: Current status and future perspectives," *Science China Technological Sciences*, vol. 61, pp. 1446–1471, 2018.

[18] Z. Zhang, "Introduction to machine learning: k-nearest neighbors," *Annals of translational medicine*, vol. 4, no. 11, 2016.

[19] R. Schnabel, R. Wahl, and R. Klein, "Efficient ransac for point-cloud shape detection," in *Computer graphics forum*, vol. 26, no. 2. Wiley Online Library, 2007, pp. 214–226.

[20] Z. Cao, D. Yang, K. Jiang, S. Xu, S. Wang, M. Zhu, and Z. Xiao, "A geometry-driven car-following distance estimation algorithm robust to road slopes," *Transportation research part C: emerging technologies*, vol. 102, pp. 274–288, 2019.

[21] A. M. Muad, A. Hussain, S. A. Samad, M. M. Mustaffa, and B. Y. Majlis, "Implementation of inverse perspective mapping algorithm for the development of an automatic lane tracking system," in *2004 IEEE Region 10 Conference TENCON 2004.* IEEE, 2004, pp. 207–210.

[22] S. S. Beauchemin and J. L. Barron, "The computation of optical flow," *ACM computing surveys (CSUR)*, vol. 27, no. 3, pp. 433–466, 1995.

[23] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*. Ieee, 2011, pp. 2564–2571.

[24] S. Sivaraman and M. M. Trivedi, "Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis," *IEEE transactions on intelligent transportation systems*, vol. 14, no. 4, pp. 1773–1795, 2013.

[25] S. Yenikaya, G. Yenikaya, and E. Düven, "Keeping the vehicle on the road: A survey on on-road lane detection systems," *ACM Computing Surveys (Csur)*, vol. 46, no. 1, pp. 1–43, 2013.

[26] M. Yang, K. Jiang, B. Wijaya, T. Wen, J. Miao, J. Huang, C. Zhong, W. Zhang, H. Chen, and D. Yang, "Review and challenge: High definition map technology for intelligent connected vehicle," *Fundamental Research*, 2024.

[27] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2110–2118.

[28] E. D. Kaplan and C. Hegarty, *Understanding GPS/GNSS: principles and applications*. Artech house, 2017.

[29] X. Luo, S. Li, and H. Xu, "Results of real-time kinematic positioning based on real gps l5 data," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 8, pp. 1193–1197, 2016.

[30] N. Ahmad, R. A. R. Ghazilla, N. M. Khairi, and V. Kasi, "Reviews on various inertial measurement unit (imu) sensor applications," *International Journal of Signal Processing Systems*, vol. 1, no. 2, pp. 256–262, 2013.

[31] I. Skog and P. Handel, "In-car positioning and navigation technologies—a survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, no. 1, pp. 4–21, 2009.

[32] O. Souissi, R. Benatitallah, D. Duvivier, A. Artiba, N. Belanger, and P. Feyzeau, "Path planning: A 2013 survey," in *Proceedings of 2013 international conference on industrial engineering and systems management (IESM)*. IEEE, 2013, pp. 1–8.

[33] S. Feng, Y. Feng, C. Yu, Y. Zhang, and H. X. Liu, "Testing scenario library generation for connected and automated vehicles, part i: Methodology," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1573–1582, 2020.

[34] T. Toledo, H. N. Koutsopoulos, and M. Ben-Akiva, "Integrated driving behavior modeling," *Transportation Research Part C: Emerging Technologies*, vol. 15, no. 2, pp. 96–112, 2007.

[35] A. Kesting, M. Treiber, and D. Helbing, "General lane-changing model mobil for car-following models," *Transportation Research Record*, vol. 1999, no. 1, pp. 86–94, 2007.

[36] M. Pivtoraiko, R. A. Knepper, and A. Kelly, "Differentially constrained mobile robot motion planning in state lattices," *Journal of Field Robotics*, vol. 26, no. 3, pp. 308–333, 2009.

[37] M. McNaughton, C. Urmson, J. M. Dolan, and J.-W. Lee, "Motion planning for autonomous driving with a conformal spatiotemporal lattice," in *2011 IEEE International Conference on Robotics and Automation*. IEEE, 2011, pp. 4889–4895.

[38] D. González, J. Pérez, V. Milanés, and F. Nashashibi, "A review of motion planning techniques for automated vehicles," *IEEE Transactions on intelligent transportation systems*, vol. 17, no. 4, pp. 1135–1145, 2015.

[39] M. Werling, J. Ziegler, S. Kammel, and S. Thrun, "Optimal trajectory generation for dynamic street scenarios in a frenet frame," in *2010 IEEE international conference on robotics and automation*. IEEE, 2010, pp. 987–993.

[40] "Pid controllers: theory, design, and tuning," *The international society of measurement and control*, 1995.

[41] A. E. Bryson, *Applied optimal control: optimization, estimation and control*. Routledge, 2018.

[42] E. F. Camacho and C. Bordons, Eds., *Model predictive control*. Berlin Heidelberg: Springer-Verlag, 1999.

[43] S. Xu and H. Peng, "Design, analysis, and experiments of preview path tracking control for autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 1, pp. 48–58, 2019.

[44] B. Paden, M. Čáp, S. Z. Yong, D. Yershov, and E. Frazzoli, "A survey of motion planning and control techniques for self-driving urban vehicles," *IEEE Transactions on intelligent vehicles*, vol. 1, no. 1, pp. 33–55, 2016.

[45] N. Lu, N. Cheng, N. Zhang, X. Shen, and J. W. Mark, "Connected vehicles: Solutions and challenges," *IEEE internet of things journal*, vol. 1, no. 4, pp. 289–299, 2014.

[46] Q. Li, Z. Chen, and X. Li, "A review of connected and automated vehicle platoon merging and splitting operations," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 12, pp. 22 790–22 806, 2022.

[47] G. Cui, W. Zhang, Y. Xiao, L. Yao, and Z. Fang, "Cooperative perception technology of autonomous driving in the internet of vehicles environment: A review," *Sensors*, vol. 22, no. 15, p. 5535, 2022.

[48] P. Lv, J. Han, J. Nie, Y. Zhang, J. Xu, C. Cai, and Z. Chen, "Cooperative decision-making of connected and autonomous vehicles in an emergency," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 2, pp. 1464–1477, 2022.

[49] Z. Bao, S. Hossain, H. Lang, and X. Lin, "A review of high-definition map creation methods for autonomous driving," *Engineering Applications of Artificial Intelligence*, vol. 122, p. 106125, 2023.

[50] M. Yang, X. Jiao, K. Jiang, Q. Cheng, Y. Yang, M. Yang, and D. Yang, "Online quantitative analysis of perception uncertainty based on high-definition map," *Sensors*, vol. 23, no. 24, p. 9876, 2023.

[51] T. Wen, K. Jiang, B. Wijaya, H. Li, M. Yang, and D. Yang, "Tm 3 loc: Tightly-coupled monocular map matching for high precision vehicle localization," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 20 268–20 281, 2022.

[52] K. Kim, S. Cho, and W. Chung, "Hd map update for autonomous driving with crowdsourced data," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1895–1901, 2021.

[53] X. Tang, K. Jiang, M. Yang, Z. Liu, P. Jia, B. Wijaya, T. Wen, L. Cui, and D. Yang, "High-definition maps construction based on visual sensor: A comprehensive survey," *IEEE Transactions on Intelligent Vehicles*, 2023.

[54] G. Bresson, Z. Alsayed, L. Yu, and S. Glaser, "Simultaneous localization and mapping: A survey of current trends in autonomous driving," *IEEE Transactions on Intelligent Vehicles*, vol. 2, no. 3, pp. 194–220, 2017.

[55] Autoware Foundation, "Autoware," 2024, accessed: 2025-01-03. [Online]. Available: https://www.autoware.org

[56] Baidu Apollo, "Apollo open platform," 2024, accessed: 2025-01-03. [Online]. Available: https://developer.apollo.auto/

[57] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[58] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[59] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[60] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[61] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.

[62] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. R. Qi, Y. Zhou *et al.*, "Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9710–9719.

[63] P. Sermanet, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[64] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[65] R. Girshick, "Fast r-cnn," *arXiv preprint arXiv:1504.08083*, 2015.

[66] S. Ren, "Faster r-cnn: Towards real-time object detection with region proposal networks," *arXiv preprint arXiv:1506.01497*, 2015.

[67] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[68] J. Redmon, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[69] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[70] Z. Ge, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[71] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[72] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[73] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.

[74] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6569–6578.

[75] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 3645–3649.

[76] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[77] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[78] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[79] J. Janai, F. Güney, A. Behl, A. Geiger *et al.*, "Computer vision for autonomous vehicles: Problems, datasets and state of the art," *Foundations and Trends® in Computer Graphics and Vision*, vol. 12, no. 1–3, pp. 1–308, 2020.

[80] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, vol. 111, no. 3, pp. 257–276, 2023.

[81] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[82] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[83] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4490–4499.

[84] B. Yang, W. Luo, and R. Urtasun, "Pixor: Real-time 3d object detection from point clouds," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 7652–7660.

[85] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 697–12 705.

[86] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 1907–1915.

[87] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 918–927.

[88] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep continuous fusion for multi-sensor 3d object detection," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 641–656.

[89] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, "Joint 3d proposal generation and object detection from view aggregation," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–8.

[90] Y. Li and J. Ibanez-Guzman, "Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems," *IEEE Signal Processing Magazine*, vol. 37, no. 4, pp. 50–61, 2020.

[91] C. Xiang, C. Feng, X. Xie, B. Shi, H. Lu, Y. Lv, M. Yang, and Z. Niu, "Multi-sensor fusion and cooperative perception for autonomous driving: A review," *IEEE Intelligent Transportation Systems Magazine*, 2023.

[92] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[93] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[94] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[95] D. Meng, X. Chen, Z. Fan, G. Zeng, H. Li, Y. Yuan, L. Sun, and J. Wang, "Conditional detr for fast training convergence," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3651–3660.

[96] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu *et al.*, "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 87–110, 2022.

[97] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[98] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[99] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, "Twins: Revisiting the design of spatial attention in vision transformers," *Advances in neural information processing systems*, vol. 34, pp. 9355–9366, 2021.

[100] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *Advances in neural information processing systems*, vol. 34, pp. 15 908–15 919, 2021.

[101] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.

[102] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022.

[103] B. Yang, M. Liang, and R. Urtasun, "Hdnet: Exploiting hd maps for 3d object detection," in *Conference on Robot Learning*. PMLR, 2018, pp. 146–155.

[104] Tesla, "Tesla ai day 2021," https://www.youtube.com/watch?v=j0z4FweCy4M&ab_channel=Tesla, 2024, accessed: 2024-12-20.

[105] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *European conference on computer vision*. Springer, 2022, pp. 1–18.

[106] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2023, pp. 2774–2781.

[107] C. Yang, Y. Chen, H. Tian, C. Tao, X. Zhu, Z. Zhang, G. Huang, H. Li, Y. Qiao, L. Lu *et al.*, "Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 830–17 839.

[108] H. Liu, Y. Teng, T. Lu, H. Wang, and L. Wang, "Sparsebev: High-performance sparse 3d object detection from multi-camera videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 580–18 590.

[109] T. Tang, D. Wei, Z. Jia, T. Gao, C. Cai, C. Hou, P. Jia, K. Zhan, H. Sun, J. Fan *et al.*, "Bev-tsr: Text-scene retrieval in bev space for autonomous driving," *arXiv e-prints*, pp. arXiv–2401, 2024.

[110] C. Pan, B. Yaman, S. Velipasalar, and L. Ren, "Clip-bevformer: Enhancing multi-view image-based bev detector with ground truth flow," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 216–15 225.

[111] H. Li, C. Sima, J. Dai, W. Wang, L. Lu, H. Wang, J. Zeng, Z. Li, J. Yang, H. Deng *et al.*, "Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[112] S. Thrun, "Probabilistic robotics," *Communications of the ACM*, vol. 45, no. 3, pp. 52–57, 2002.

[113] A.-Q. Cao and R. De Charette, "Monoscene: Monocular 3d semantic scene completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3991–4001.

[114] Tesla, "A look at tesla's occupancy networks," https://www.thinkautonomous.ai/blog/occupancy-networks/, 2024, accessed: 2024-11-23.

[115] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "Tri-perspective view for vision-based 3d semantic occupancy prediction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 9223–9232.

[116] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, J. Zhou, and J. Lu, "Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 729–21 740.

[117] Y. Zhang, Z. Zhu, and D. Du, "Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9433–9443.

[118] Z. Li, Z. Yu, D. Austin, M. Fang, S. Lan, J. Kautz, and J. M. Alvarez, "Fb-occ: 3d occupancy prediction based on forward-backward view transformation," *arXiv preprint arXiv:2307.01492*, 2023.

[119] J. Ma, X. Chen, J. Huang, J. Xu, Z. Luo, J. Xu, W. Gu, R. Ai, and H. Wang, "Cam4docc: Benchmark for camera-only 4d occupancy forecasting in autonomous driving applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 486–21 495.

[120] X. Tian, T. Jiang, L. Yun, Y. Mao, H. Yang, Y. Wang, Y. Wang, and H. Zhao, "Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[121] X. Wang, Z. Zhu, W. Xu, Y. Zhang, Y. Wei, X. Chi, Y. Ye, D. Du, J. Lu, and X. Wang, "Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 850–17 859.

[122] Y. Huang, W. Zheng, B. Zhang, J. Zhou, and J. Lu, "Selfocc: Self-supervised vision-based 3d occupancy prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 946–19 956.

[123] C. Zhang, J. Yan, Y. Wei, J. Li, L. Liu, Y. Tang, Y. Duan, and J. Lu, "Occnerf: Self-supervised multi-camera occupancy prediction with neural radiance fields," *arXiv e-prints*, pp. arXiv–2312, 2023.

[124] M. Pan, J. Liu, R. Zhang, P. Huang, X. Li, H. Xie, B. Wang, L. Liu, and S. Zhang, "Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 12 404–12 411.

[125] S. Boeder, F. Gigengack, and B. Risse, "Occflownet: Towards self-supervised occupancy estimation via differentiable rendering and occupancy flow," *arXiv preprint arXiv:2402.12792*, 2024.

[126] H. Xu, J. Chen, S. Meng, Y. Wang, and L.-P. Chau, "A survey on occupancy perception for autonomous driving: The information fusion perspective," *Information Fusion*, vol. 114, p. 102671, 2025.

[127] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 961–971.

[128] Y. Chai, B. Sapp, M. Bansal, and D. Anguelov, "Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction," *arXiv preprint arXiv:1910.05449*, 2019.

[129] B. Varadarajan, A. Hefny, A. Srivastava, K. S. Refaat, N. Nayakanti, A. Cornman, K. Chen, B. Douillard, C. P. Lam, D. Anguelov *et al.*, "Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 7814–7821.

[130] J. Gao, C. Sun, H. Zhao, Y. Shen, D. Anguelov, C. Li, and C. Schmid, "Vectornet: Encoding hd maps and agent dynamics from vectorized representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 525–11 533.

[131] H. Zhao, J. Gao, T. Lan, C. Sun, B. Sapp, B. Varadarajan, Y. Shen, Y. Shen, Y. Chai, C. Schmid *et al.*, "Tnt: Target-driven trajectory prediction," in *Conference on Robot Learning*. PMLR, 2021, pp. 895–904.

[132] J. Gu, C. Sun, and H. Zhao, "Densetnt: End-to-end trajectory prediction from dense goal sets," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 303–15 312.

[133] A. Seff, B. Cera, D. Chen, M. Ng, A. Zhou, N. Nayakanti, K. S. Refaat, R. Al-Rfou, and B. Sapp, "Motionlm: Multi-agent motion forecasting as language modeling," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8579–8590.

[134] S. Shi, L. Jiang, D. Dai, and B. Schiele, "Motion transformer with global intention localization and local movement refinement," *Advances in Neural Information Processing Systems*, vol. 35, pp. 6531–6543, 2022.

[135] S. Shi, L. Jiang, D. Dai, and B.Schiele, "Mtr++: Multi-agent motion prediction with symmetric scene modeling and guided intention querying," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[136] Z. Zhou, J. Wang, Y.-H. Li, and Y.-K. Huang, "Query-centric trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 863–17 873.

[137] M. Bansal, A. Krizhevsky, and A. Ogale, "Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst," *arXiv preprint arXiv:1812.03079*, 2018.

[138] A. Cui, S. Casas, A. Sadat, R. Liao, and R. Urtasun, "Lookout: Diverse multi-future prediction and planning for self-driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 107–16 116.

[139] H. Song, W. Ding, Y. Chen, S. Shen, M. Y. Wang, and Q. Chen, "Pip: Planning-informed trajectory prediction for autonomous driving," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer, 2020, pp. 598–614.

[140] S. Casas, A. Sadat, and R. Urtasun, "Mp3: A unified model to map, perceive, predict and plan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 403–14 412.

[141] Z. Huang, H. Liu, J. Wu, and C. Lv, "Differentiable integrated motion prediction and planning with learnable cost function for autonomous driving," *IEEE transactions on neural networks and learning systems*, 2023.

[142] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 627–635.

[143] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[144] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel *et al.*, "A general reinforcement learning algorithm that masters chess, shogi, and go through self-play," *Science*, vol. 362, no. 6419, pp. 1140–1144, 2018.

[145] S. Shalev-Shwartz, S. Shammah, and A. Shashua, "Safe, multi-agent, reinforcement learning for autonomous driving," *arXiv preprint arXiv:1610.03295*, 2016.

[146] A. E. Sallab, M. Abdou, E. Perot, and S. Yogamani, "Deep reinforcement learning framework for autonomous driving," *arXiv preprint arXiv:1704.02532*, 2017.

[147] A. Kendall, J. Hawke, D. Janz, P. Mazur, D. Reda, J.-M. Allen, V.-D. Lam, A. Bewley, and A. Shah, "Learning to drive in a day," in *2019 international conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 8248–8254.

[148] J. Chen, B. Yuan, and M. Tomizuka, "Model-free deep reinforcement learning for urban autonomous driving," in *2019 IEEE intelligent transportation systems conference (ITSC)*. IEEE, 2019, pp. 2765–2771.

[149] J. Duan, S. Eben Li, Y. Guan, Q. Sun, and B. Cheng, "Hierarchical reinforcement learning for self-driving decision-making without reliance on labelled driving data," *IET Intelligent Transport Systems*, vol. 14, no. 5, pp. 297–305, 2020.

[150] J. Chen, S. E. Li, and M. Tomizuka, "Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 5068–5078, 2021.

[151] Z. Cao, S. Xu, X. Jiao, H. Peng, and D. Yang, "Trustworthy safety improvement for autonomous driving using reinforcement learning," *Transportation research part C: emerging technologies*, vol. 138, p. 103656, 2022.

[152] S. E. Li, *Reinforcement learning for sequential decision and optimal control*. Springer, 2023.

[153] S. Gu, L. Yang, Y. Du, G. Chen, F. Walter, J. Wang, and A. Knoll, "A review of safe reinforcement learning: Methods, theories and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[154] Z. Cao, K. Jiang, W. Zhou, S. Xu, H. Peng, and D. Yang, "Continuous improvement of self-driving cars using dynamic confidence-aware reinforcement learning," *Nature Machine Intelligence*, vol. 5, no. 2, pp. 145–158, 2023.

[155] B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. Al Sallab, S. Yogamani, and P. Pérez, "Deep reinforcement learning for autonomous driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 4909–4926, 2021.

[156] C. Pek, S. Manzinger, M. Koschi, and M. Althoff, "Using online verification to prevent autonomous vehicles from causing accidents," *Nature Machine Intelligence*, vol. 2, no. 9, pp. 518–528, 2020.

[157] S. Shalev-Shwartz, "On a formal model of safe and scalable self-driving cars," *arXiv preprint arXiv:1708.06374*, 2017.

[158] Z. Cao, D. Yang, S. Xu, H. Peng, B. Li, S. Feng, and D. Zhao, "Highway exiting planner for automated vehicles using reinforcement learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 2, pp. 990–1000, 2020.

[159] T. Hester, M. Vecerik, O. Pietquin, M. Lanctot, T. Schaul, B. Piot, D. Horgan, J. Quan, A. Sendonaris, I. Osband *et al.*, "Deep q-learning from demonstrations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[160] Y. Lu, J. Fu, G. Tucker, X. Pan, E. Bronstein, R. Roelofs, B. Sapp, B. White, A. Faust, S. Whiteson *et al.*, "Imitation is not enough: Robustifying imitation with reinforcement learning for challenging driving scenarios," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 7553–7560.

[161] A. Kuefler, J. Morton, T. Wheeler, and M. Kochenderfer, "Imitating driver behavior with generative adversarial networks," in *2017 IEEE intelligent vehicles symposium (IV)*. IEEE, 2017, pp. 204–211.

[162] Y. Li, J. Song, and S. Ermon, "Infogail: Interpretable imitation learning from visual demonstrations," *Advances in neural information processing systems*, vol. 30, 2017.

[163] R. Bhattacharyya, B. Wulfe, D. J. Phillips, A. Kuefler, J. Morton, R. Senanayake, and M. J. Kochenderfer, "Modeling human driving behavior through generative adversarial imitation learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 3, pp. 2874–2887, 2022.

[164] X. Yan, Z. Zou, S. Feng, H. Zhu, H. Sun, and H. X. Liu, "Learning naturalistic driving environment with statistical realism," *Nature communications*, vol. 14, no. 1, p. 2037, 2023.

[165] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[166] H. Wang, P. Cai, Y. Sun, L. Wang, and M. Liu, "Learning interpretable end-to-end vision-based motion planning for autonomous driving with optical flow distillation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 731–13 737.

[167] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 563–578.

[168] Y. Feng, W. Hua, and Y. Sun, "Nle-dm: Natural-language explanations for decision making of autonomous driving based on semantic scene understanding," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 9, pp. 9780–9791, 2023.

[169] M. Bojarski, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[170] A. E. Sallab, M. Abdou, E. Perot, and S. Yogamani, "End-to-end deep reinforcement learning for lane keeping assist," *arXiv preprint arXiv:1612.04340*, 2016.

[171] H. Xu, Y. Gao, F. Yu, and T. Darrell, "End-to-end learning of driving models from large-scale video datasets," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2174–2182.

[172] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang *et al.*, "Planning-oriented autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 853–17 862.

[173] B. Jiang, S. Chen, Q. Xu, B. Liao, J. Chen, H. Zhou, Q. Zhang, W. Liu, C. Huang, and X. Wang, "Vad: Vectorized scene representation for efficient autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8340–8350.

[174] W. Tong, C. Sima, T. Wang, L. Chen, S. Wu, H. Deng, Y. Gu, L. Lu, P. Luo, D. Lin *et al.*, "Scene as occupancy," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8406–8415.

[175] S. Chen, B. Jiang, H. Gao, B. Liao, Q. Xu, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Vadv2: End-to-end vectorized autonomous driving via probabilistic planning," *arXiv preprint arXiv:2402.13243*, 2024.

[176] W. Sun, X. Lin, Y. Shi, C. Zhang, H. Wu, and S. Zheng, "Sparsedrive: End-to-end autonomous driving via sparse scene representation," *arXiv preprint arXiv:2405.19620*, 2024.

[177] D. McCandless. (2021) Million lines of code visualization. Accessed: 2024-12-18. [Online]. Available: https://informationisbeautiful.net/visualizations/million-lines-of-code/

[178] G. Lou, Y. Deng, X. Zheng, M. Zhang, and T. Zhang, "Testing of autonomous driving systems: where are we and where should we go?" in *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2022, pp. 31–43.

[179] N. Rajabli, F. Flammini, R. Nardone, and V. Vittorini, "Software verification and validation of safe autonomous cars: A systematic literature review," *IEEE Access*, vol. 9, pp. 4797–4819, 2020.

[180] National Highway Traffic Safety Administration (NHTSA), "New Car Assessment Program (NCAP)," 2024, accessed: 2025-01-03. [Online]. Available: https://www.transportation.gov/bipartisan-infrastructure-law/regulations/2022-04894

[181] International, "ISO 26262: Road Vehicles – Functional Safety," 2018, accessed: 2025-01-03. [Online]. Available: https://www.iso.org/standard/68383.html

[182] International Organization for Standardization (ISO), "ISO 21448: Road Vehicles – Safety of the Intended Functionality," 2022, accessed: 2025-01-03. [Online]. Available: https://www.iso.org/standard/77490.html

[183] Underwriters Laboratories (UL), "UL 4600: Standard for Safety for the Evaluation of Autonomous Products," 2022, accessed: 2025-01-03. [Online]. Available: https://users.ece.cmu.edu/∼koopman/ul4600/index.html

[184] International Organization for Standardization (ISO), "ISO 34502: Road vehicles — Test scenarios for automated driving systems — Scenario based safety evaluation framework," Year, accessed: 2025-01-03. [Online]. Available: https://www.iso.org/standard/78951.html

[185] J. Zhou and L. del Re, "Reduced complexity safety testing for adas & adf," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 5985–5990, 2017.

[186] G. E. Mullins, P. G. Stankiewicz, R. C. Hawthorne, and S. K. Gupta, "Adaptive generation of challenging scenarios for testing and evaluation of autonomous vehicles," *Journal of Systems and Software*, vol. 137, pp. 197–215, 2018.

[187] D. Jung, D. Jung, C. Jeong, Y. Kou, and H. Peng, "Worst case scenarios generation and its application on driving," SAE Technical Paper, Tech. Rep., 2007.

[188] D. Zhao, H. Lam, H. Peng, S. Bao, D. J. LeBlanc, K. Nobukawa, and C. S. Pan, "Accelerated evaluation of automated vehicles safety in lane-change scenarios based on importance sampling techniques," *IEEE transactions on intelligent transportation systems*, vol. 18, no. 3, pp. 595–607, 2016.

[189] S. Feng, Y. Feng, H. Sun, S. Bao, Y. Zhang, and H. X. Liu, "Testing scenario library generation for connected and automated vehicles, part ii: Case studies," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 9, pp. 5635–5647, 2020.

[190] S. Feng, Y. Feng, X. Yan, S. Shen, S. Xu, and H. X. Liu, "Safety assessment of highly automated driving systems in test tracks: A new framework," *Accident Analysis & Prevention*, vol. 144, p. 105664, 2020.

[191] N. Fulton and A. Platzer, "Safe ai for cps," in *2018 IEEE International Test Conference (ITC)*. IEEE, 2018, pp. 1–7.

[192] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[193] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.

[194] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, "The apolloscape dataset for autonomous driving," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 954–960.

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

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

[197] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[198] H. Caesar, J. Kabzan, K. S. Tan, W. K. Fong, E. Wolff, A. Lang, L. Fletcher, O. Beijbom, and S. Omari, "nuplan: A closed-loop ml-

based planning benchmark for autonomous vehicles," *arXiv preprint arXiv:2106.11810*, 2021.

[199] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.

[200] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. R. Qi, Y. Zhou *et al.*, "Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9710–9719.

[201] J. Houston, G. Zuidhof, L. Bergamini, Y. Ye, L. Chen, A. Jain, S. Omari, V. Iglovikov, and P. Ondruska, "One thousand and one hours: Self-driving motion prediction dataset," in *Conference on Robot Learning*. PMLR, 2021, pp. 409–418.

[202] F. H. A. F. U.S. Department of Transportation, "Next generation simulation (ngsim) vehicle trajectories and supporting data," 2005, accessed: 2025-01-03. [Online]. Available: https://data.transportation.gov/Automobiles/Next-Generation-Simulation-NGSIM-Vehicle-Trajector/8ect-6jqj/about_data

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

[204] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, "The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems," in *2018 21st international conference on intelligent transportation systems (ITSC)*. IEEE, 2018, pp. 2118–2125.

[205] R. Krajewski, T. Moers, J. Bock, L. Vater, and L. Eckstein, "The round dataset: A drone dataset of road user trajectories at roundabouts in germany," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2020, pp. 1–6.

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

[207] R. Zhang, D. Meng, S. Shen, Z. Zou, H. Li, and H. X. Liu, "Msight: An edge-cloud infrastructure-based perception system for connected automated vehicles," *arXiv preprint arXiv:2310.05290*, 2023.

[208] O. Zheng, M. Abdel-Aty, L. Yue, A. Abdelraouf, Z. Wang, and N. Mahmoud, "Citysim: a drone-based vehicle trajectory dataset for safety-oriented research and digital twins," *Transportation research record*, vol. 2678, no. 4, pp. 606–621, 2024.

[209] D. Gloudemans, Y. Wang, J. Ji, G. Zachar, W. Barbour, E. Hall, M. Cebelak, L. Smith, and D. B. Work, "I-24 motion: An instrument for freeway traffic science," *Transportation Research Part C: Emerging Technologies*, vol. 155, p. 104311, 2023.

[210] E. Barmpounakis and N. Geroliminis, "On the new era of urban traffic monitoring with massive drone data: The pneuma large-scale field experiment," *Transportation research part C: emerging technologies*, vol. 111, pp. 50–71, 2020.

[211] J. Janai, F. Güney, A. Behl, A. Geiger *et al.*, "Computer vision for autonomous vehicles: Problems, datasets and state of the art," *Foundations and Trends® in Computer Graphics and Vision*, vol. 12, no. 1–3, pp. 1–308, 2020.

[212] Y. Wang, Z. Han, Y. Xing, S. Xu, and J. Wang, "A survey on datasets for the decision making of autonomous vehicles," *IEEE Intelligent Transportation Systems Magazine*, 2024.

[213] W. LLC, "Waymo open challenge," 2020, accessed: 2025-01-03. [Online]. Available: https://waymo.com/open/challenges/

[214] N. Montali, J. Lambert, P. Mougin, A. Kuefler, N. Rhinehart, M. Li, C. Gulino, T. Emrich, Z. Yang, S. Whiteson *et al.*, "The waymo open sim agents challenge," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[215] N. Karnchanachari, D. Geromichalos, K. S. Tan, N. Li, C. Eriksen, S. Yaghoubi, N. Mehdipour, G. Bernasconi, W. K. Fong, Y. Guo *et al.*, "Towards learning-based planning: The nuplan benchmark for real-world autonomous driving," *arXiv preprint arXiv:2403.04133*, 2024.

[216] C. Team, "Carla challenge," 2020, accessed: 2025-01-03. [Online]. Available: https://leaderboard.carla.org/challenge/

[217] O. Lab, "Autonomous grand challenge," 2024, accessed: 2025-01-03. [Online]. Available: https://opendrivelab.com/challenge2024/

[218] Mcity, "Mcity autonomous vehicle challenge," 2024, accessed: 2025-01-03. [Online]. Available: https://mcity.umich.edu/av-challenge/

[219] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.

[220] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and Service Robotics: Results of the 11th International Conference*. Springer, 2018, pp. 621–635.

[221] G. Rong, B. H. Shin, H. Tabatabaee, Q. Lu, S. Lemke, M. Možeiko, E. Boise, G. Uhm, M. Gerow, S. Mehta *et al.*, "Lgsvl simulator: A high fidelity simulator for autonomous driving," in *2020 IEEE 23rd International conference on intelligent transportation systems (ITSC)*. IEEE, 2020, pp. 1–6.

[222] B. Wymann, E. Espié, C. Guionneau, C. Dimitrakakis, R. Coulom, and A. Sumner, "Torcs, the open racing car simulator," *Software available at http://torcs. sourceforge. net*, vol. 4, no. 6, p. 2, 2000.

[223] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wießner, "Microscopic traffic simulation using sumo," in *2018 21st international conference on intelligent transportation systems (ITSC)*. IEEE, 2018, pp. 2575–2582.

[224] E. Leurent, "An environment for autonomous driving decision-making," https://github.com/eleurent/highway-env, 2018.

[225] "Gazebo simulator," 2024, accessed: 2025-01-03. [Online]. Available: https://gazebosim.org/home

[226] M. S. Corporation, "Carsim vehicle simulation software," 2024, accessed: 2025-01-03. [Online]. Available: https://www.carsim.com/products/carsim/index.php

[227] Q. Li, Z. Peng, L. Feng, Q. Zhang, Z. Xue, and B. Zhou, "Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 3, pp. 3461–3475, 2022.

[228] W. Li, C. Pan, R. Zhang, J. Ren, Y. Ma, J. Fang, F. Yan, Q. Geng, X. Huang, H. Gong *et al.*, "Aads: Augmented autonomous driving simulation using data-driven algorithms," *Science robotics*, vol. 4, no. 28, p. eaaw0863, 2019.

[229] S. Manivasagam, S. Wang, K. Wong, W. Zeng, M. Sazanovich, S. Tan, B. Yang, W.-C. Ma, and R. Urtasun, "Lidarsim: Realistic lidar simulation by leveraging the real world," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 167–11 176.

[230] Z. Yang, Y. Chen, J. Wang, S. Manivasagam, W.-C. Ma, A. J. Yang, and R. Urtasun, "Unisim: A neural closed-loop sensor simulator," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1389–1399.

[231] X. Yan, S. Feng, H. Sun, and H. X. Liu, "Distributionally consistent simulation of naturalistic driving environment for autonomous vehicle testing," *arXiv preprint arXiv:2101.02828*, 2021.

[232] S. Suo, S. Regalado, S. Casas, and R. Urtasun, "Trafficsim: Learning to simulate realistic multi-agent behaviors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 400–10 409.

[233] Z. Zhang, A. Liniger, D. Dai, F. Yu, and L. Van Gool, "Trafficbots: Towards world models for autonomous driving simulation and motion prediction," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 1522–1529.

[234] J. Yang, S. Gao, Y. Qiu, L. Chen, T. Li, B. Dai, K. Chitta, P. Wu, J. Zeng, P. Luo *et al.*, "Generalized predictive model for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 662–14 672.

[235] J. You, H. Shi, Z. Jiang, Z. Huang, R. Gan, K. Wu, X. Cheng, X. Li, and B. Ran, "V2x-vlm: End-to-end v2x cooperative autonomous driving through large vision-language models," *arXiv preprint arXiv:2408.09251*, 2024.

[236] B. Li, Y. Wang, J. Mao, B. Ivanovic, S. Veer, K. Leung, and M. Pavone, "Driving everywhere with large language model policy adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2024, pp. 14 948–14 957.

[237] T. Choudhary, V. Dewangan, S. Chandhok, S. Priyadarshan, A. Jain, A. K. Singh, S. Srivastava, K. M. Jatavallabhula, and K. M. Krishna, "Talk2bev: Language-enhanced bird's-eye view maps for autonomous driving," *arXiv preprint arXiv:2310.02251*, 2023.

[238] S. Yang, J. Liu, R. Zhang, M. Pan, Z. Guo, X. Li, Z. Chen, P. Gao, Y. Guo, and S. Zhang, "Lidar-llm: Exploring the potential of large language models for 3d lidar understanding," *arXiv preprint arXiv:2312.14074*, 2023.

[239] T.-H. Wang, A. Maalouf, W. Xiao, Y. Ban, A. Amini, G. Rosman, S. Karaman, and D. Rus, "Drive anywhere: Generalizable end-to-end autonomous driving with multi-modal foundation models," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 6687–6694.

[240] B. Jiang, S. Chen, B. Liao, X. Zhang, W. Yin, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Senna: Bridging large vision-language models and end-to-end autonomous driving," *arXiv preprint arXiv:2410.22313*, 2024.

[241] C. Cui, Y. Ma, X. Cao, W. Ye, and Z. Wang, "Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 902–909.

[242] Z. Li, K. Li, S. Wang, S. Lan, Z. Yu, Y. Ji, Z. Li, Z. Zhu, J. Kautz, Z. Wu *et al.*, "Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation," *arXiv preprint arXiv:2406.06978*, 2024.

[243] H. Lu, X. Jia, Y. Xie, W. Liao, X. Yang, and J. Yan, "Activead: Planning-oriented active learning for end-to-end autonomous driving," *arXiv preprint arXiv:2403.02877*, 2024.

[244] M. Guo, Z. Zhang, Y. He, K. Wang, and L. Jing, "End-to-end autonomous driving without costly modularization and 3d manual annotation," *arXiv preprint arXiv:2406.17680*, 2024.

[245] Z. Qiao, H. Li, Z. Cao, and H. X. Liu, "Lightemma: Lightweight end-to-end multimodal model for autonomous driving," *arXiv preprint arXiv:2505.00284*, 2025.

[246] Z. Xu, Y. Zhang, E. Xie, Z. Zhao, Y. Guo, K.-Y. K. Wong, Z. Li, and H. Zhao, "Drivegpt4: Interpretable end-to-end autonomous driving via large language model," *IEEE Robotics and Automation Letters*, 2024.

[247] W. Wang, J. Xie, C. Hu, H. Zou, J. Fan, W. Tong, Y. Wen, S. Wu, H. Deng, Z. Li *et al.*, "Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving," *arXiv preprint arXiv:2312.09245*, 2023.

[248] H. Sha, Y. Mu, Y. Jiang, L. Chen, C. Xu, P. Luo, S. E. Li, M. Tomizuka, W. Zhan, and M. Ding, "Languagempc: Large language models as decision makers for autonomous driving," *arXiv preprint arXiv:2310.03026*, 2023.

[249] J. Liu, P. Hang, X. Qi, J. Wang, and J. Sun, "Mtd-gpt: A multi-task decision-making gpt model for autonomous driving at unsignalized intersections," in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2023, pp. 5154–5161.

[250] A.-M. Marcu, L. Chen, J. Hünermann, A. Karnsund, B. Hanotte, P. Chidananda, S. Nair, V. Badrinarayanan, A. Kendall, J. Shotton *et al.*, "Lingoqa: Video question answering for autonomous driving," *arXiv preprint arXiv:2312.14115*, 2023.

[251] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, J. Beißwenger, P. Luo, A. Geiger, and H. Li, "Drivelm: Driving with graph visual question answering," *arXiv preprint arXiv:2312.14150*, 2023.

[252] J. Mao, Y. Qian, J. Ye, H. Zhao, and Y. Wang, "Gpt-driver: Learning to drive with gpt," *arXiv preprint arXiv:2310.01415*, 2023.

[253] L. Chen, O. Sinavski, J. Hünermann, A. Karnsund, A. J. Willmott, D. Birch, D. Maund, and J. Shotton, "Driving with llms: Fusing object-level vector modality for explainable autonomous driving," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 14 093–14 100.

[254] X. Tian, J. Gu, B. Li, Y. Liu, Y. Wang, Z. Zhao, K. Zhan, P. Jia, X. Lang, and H. Zhao, "Drivevlm: The convergence of autonomous driving and large vision-language models," *arXiv preprint arXiv:2402.12289*, 2024.

[255] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A frontier large vision-language model with versatile abilities," *arXiv preprint arXiv:2308.12966*, 2023.

[256] J.-J. Hwang, R. Xu, H. Lin, W.-C. Hung, J. Ji, K. Choi, D. Huang, T. He, P. Covington, B. Sapp *et al.*, "Emma: End-to-end multimodal model for autonomous driving," *arXiv preprint arXiv:2410.23262*, 2024.

[257] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[258] S. Xing, C. Qian, Y. Wang, H. Hua, K. Tian, Y. Zhou, and Z. Tu, "Openemma: Open-source multimodal model for end-to-end autonomous driving," in *Proceedings of the Winter Conference on Applications of Computer Vision*, 2025, pp. 1001–1009.

[259] H. Shao, Y. Hu, L. Wang, G. Song, S. L. Waslander, Y. Liu, and H. Li, "Lmdrive: Closed-loop end-to-end driving with large language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 120–15 130.

[260] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[261] M. Nie, R. Peng, C. Wang, X. Cai, J. Han, H. Xu, and L. Zhang, "Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving," in *European Conference on Computer Vision*. Springer, 2025, pp. 292–308.

[262] Y. Jin, X. Shen, H. Peng, X. Liu, J. Qin, J. Li, J. Xie, P. Gao, G. Zhou, and J. Gong, "Surrealdriver: Designing generative driver agent simulation framework in urban contexts based on large language model," *arXiv preprint arXiv:2309.13193*, 2023.

[263] S. Ishida, G. Corrado, G. Fedoseev, H. Yeo, L. Russell, J. Shotton, J. F. Henriques, and A. Hu, "Langprop: A code optimization framework using large language models applied to driving," in *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.

[264] Y. LeCun, "A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27," *Open Review*, vol. 62, no. 1, pp. 1–62, 2022.

[265] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, "Mastering diverse control tasks through world models," *Nature*, pp. 1–7, 2025.

[266] N. Hansen, H. Su, and X. Wang, "Td-mpc2: Scalable, robust world models for continuous control," *arXiv preprint arXiv:2310.16828*, 2023.

[267] OpenAI, "Video generation models as world simulators," https://openai.com/index/video-generation-models-as-world-simulators/, 2024, accessed: 2024-11-23.

[268] Y. Wang, J. He, L. Fan, H. Li, Y. Chen, and Z. Zhang, "Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 749–14 759.

[269] G. Zhao, C. Ni, X. Wang, Z. Zhu, X. Zhang, Y. Wang, G. Huang, X. Chen, B. Wang, Y. Zhang *et al.*, "Drivedreamer4d: World models are effective data machines for 4d driving scene representation," *arXiv preprint arXiv:2410.13571*, 2024.

[270] L. Zhang, Y. Xiong, Z. Yang, S. Casas, R. Hu, and R. Urtasun, "Learning unsupervised world models for autonomous driving via discrete diffusion," *arXiv preprint arXiv:2311.01017*, 2023.

[271] C. Min, D. Zhao, L. Xiao, Y. Nie, and B. Dai, "Uniworld: Autonomous driving pre-training via world models," *arXiv preprint arXiv:2308.07234*, 2023.

[272] W. Zheng, W. Chen, Y. Huang, B. Zhang, Y. Duan, and J. Lu, "Occworld: Learning a 3d occupancy world model for autonomous driving," in *European Conference on Computer Vision*. Springer, 2025, pp. 55–72.

[273] A. Popov, A. Degirmenci, D. Wehr, S. Hegde, R. Oldja, A. Kamenev, B. Douillard, D. Nistér, U. Muller, R. Bhargava *et al.*, "Mitigating covariate shift in imitation learning for autonomous vehicles using latent space generative world models," *arXiv preprint arXiv:2409.16663*, 2024.

[274] J. Wei, S. Yuan, P. Li, Q. Hu, Z. Gan, and W. Ding, "Occllama: An occupancy-language-action generative world model for autonomous driving," *arXiv preprint arXiv:2409.03272*, 2024.

[275] G. Zhao, X. Wang, Z. Zhu, X. Chen, G. Huang, X. Bao, and X. Wang, "Drivedreamer-2: Llm-enhanced world models for diverse driving video generation," *arXiv preprint arXiv:2403.06845*, 2024.

[276] A. Bar, G. Zhou, D. Tran, T. Darrell, and Y. LeCun, "Navigation world models," *arXiv preprint arXiv:2412.03572*, 2024.

[277] Y. Yang, J. Mei, Y. Ma, S. Du, W. Chen, Y. Qian, Y. Feng, and Y. Liu, "Driving in the occupancy world: Vision-centric 4d occupancy forecasting and planning via world models for autonomous driving," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 9, 2025, pp. 9327–9335.

[278] V. Voleti, A. Jolicoeur-Martineau, and C. Pal, "Mcvd-masked conditional video diffusion for prediction, generation, and interpolation," *Advances in neural information processing systems*, vol. 35, pp. 23 371–23 385, 2022.

[279] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu, "Drivedreamer: Towards real-world-driven world models for autonomous driving," *arXiv preprint arXiv:2309.09777*, 2023.

[280] X. Li, Y. Zhang, and X. Ye, "Drivingdiffusion: Layout-guided multi-view driving scene video generation with latent diffusion model," *arXiv preprint arXiv:2310.07771*, 2023.

[281] R. Gao, K. Chen, E. Xie, L. Hong, Z. Li, D.-Y. Yeung, and Q. Xu, "Magicdrive: Street view generation with diverse 3d geometry control," *arXiv preprint arXiv:2310.02601*, 2023.

[282] F. Jia, W. Mao, Y. Liu, Y. Zhao, Y. Wen, C. Zhang, X. Zhang, and T. Wang, "Adriver-i: A general world model for autonomous driving," *arXiv preprint arXiv:2311.13549*, 2023.

[283] V. Zyrianov, H. Che, Z. Liu, and S. Wang, "Lidardm: Generative lidar simulation in a generated world," *arXiv preprint arXiv:2404.02903*, 2024.

[284] J. Lu, Z. Huang, Z. Yang, J. Zhang, and L. Zhang, "Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation," in *European Conference on Computer Vision*. Springer, 2025, pp. 329–345.

[285] X. Wang, Z. Zhu, G. Huang, B. Wang, X. Chen, and J. Lu, "Worlddreamer: Towards general world models for video generation via predicting masked tokens," *arXiv preprint arXiv:2401.09985*, 2024.

[286] A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, and G. Corrado, "Gaia-1: A generative world model for autonomous driving," *arXiv preprint arXiv:2309.17080*, 2023.

[287] J. Zhang, C. Xu, and B. Li, "Chatscene: Knowledge-enabled safety-critical scenario generation for autonomous vehicles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 459–15 469.

[288] W. Wu, X. Feng, Z. Gao, and Y. Kan, "Smart: Scalable multi-agent real-time simulation via next-token prediction," *arXiv preprint arXiv:2405.15677*, 2024.

[289] C. Jiang, A. Cornman, C. Park, B. Sapp, Y. Zhou, D. Anguelov *et al.*, "Motiondiffuser: Controllable multi-agent motion prediction using diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9644–9653.

[290] Z. Zhong, D. Rempe, Y. Chen, B. Ivanovic, Y. Cao, D. Xu, M. Pavone, and B. Ray, "Language-guided traffic simulation via scene-level diffusion," in *Conference on Robot Learning*. PMLR, 2023, pp. 144–177.

[291] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[292] Z. Xie, J. Zhang, W. Li, F. Zhang, and L. Zhang, "S-nerf: Neural radiance fields for street views," *arXiv preprint arXiv:2303.00749*, 2023.

[293] A. Tonderski, C. Lindström, G. Hess, W. Ljungbergh, L. Svensson, and C. Petersson, "Neurad: Neural rendering for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 895–14 904.

[294] T. Yuan, Y. Mao, J. Yang, Y. Liu, Y. Wang, and H. Zhao, "Presight: Enhancing autonomous vehicle perception with city-scale nerf priors," *arXiv preprint arXiv:2403.09079*, 2024.

[295] H. Wu, X. Zuo, S. Leutenegger, O. Litany, K. Schindler, and S. Huang, "Dynamic lidar re-simulation using compositional neural fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 988–19 998.

[296] Y. Wei, Z. Wang, Y. Lu, C. Xu, C. Liu, H. Zhao, S. Chen, and Y. Wang, "Editable scene simulation for autonomous driving via collaborative llm-agents," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 077–15 087.

[297] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[298] X. Zhou, Z. Lin, X. Shan, Y. Wang, D. Sun, and M.-H. Yang, "Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 634–21 643.

[299] Y. Yan, H. Lin, C. Zhou, W. Wang, H. Sun, K. Zhan, X. Lang, X. Zhou, and S. Peng, "Street gaussians for modeling dynamic urban scenes," *arXiv preprint arXiv:2401.01339*, 2024.

[300] T. Fischer, J. Kulhanek, S. R. Bulò, L. Porzi, M. Pollefeys, and P. Kontschieder, "Dynamic 3d gaussian fields for urban areas," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[301] Connected Places Catapult (CPC) and WMG, University of Warwick, "Certicav assurance paper," https://cp-catapult.s3.amazonaws.com/uploads/2021/06/CertiCAV-Assurance-Paper-v1-4.pdf, 2021.

[302] H. Sun, X. Yan, Z. Qiao, H. Zhu, Y. Sun, J. Wang, S. Shen, D. Hogue, R. Ananta, D. Johnson *et al.*, "Terasim: Uncovering unknown unsafe events for autonomous vehicles through generative simulation," *arXiv preprint arXiv:2503.03629*, 2025.

[303] L. Russell, A. Hu, L. Bertoni, G. Fedoseev, J. Shotton, E. Arani, and G. Corrado, "Gaia-2: A controllable multi-view generative world model for autonomous driving," *arXiv preprint arXiv:2503.20523*, 2025.

[304] Z. Zhang, P. Karkus, M. Igl, W. Ding, Y. Chen, B. Ivanovic, and M. Pavone, "Closed-loop supervised fine-tuning of tokenized traffic models," *arXiv preprint arXiv:2412.05334*, 2024.

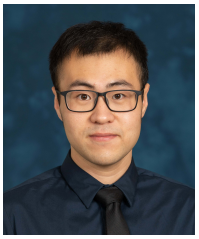
[305] NVIDIA Corporation, "NVIDIA Omniverse." [Online]. Available: https://developer.nvidia.com/omniverse

[306] N. Kalra and S. M. Paddock, "Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability?" *Transportation Research Part A: Policy and Practice*, vol. 94, pp. 182–193, 2016.

[307] S. Feng, X. Yan, H. Sun, Y. Feng, and H. X. Liu, "Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment," *Nature communications*, vol. 12, no. 1, p. 748, 2021.

[308] S. Feng, H. Sun, X. Yan, H. Zhu, Z. Zou, S. Shen, and H. X. Liu, "Dense reinforcement learning for safety validation of autonomous vehicles," *Nature*, vol. 615, no. 7953, pp. 620–627, 2023.

Henry X. Liu (Senior Member, IEEE) is the Director of the University of Michigan Transportation Research Institute (UMTRI) and the Bruce D. Greenshields Collegiate Professor Professor in Civil and Environmental Engineering at the University of Michigan, Ann Arbor. He also directs the Center for Connected and Automated Transportation, a USDOT funded Region 5 university transportation center. Dr. Liu conducts interdisciplinary research at the interface of transportation engineering, automotive engineering, and artificial intelligence. He is recognized for his foundational work in cyber-physical transportation systems, particularly on the development of smart traffic signal systems and testing/evaluation of autonomous vehicles. Prof. Liu is the managing editor of Journal of Intelligent Transportation Systems and a board member for the ITS America. He is also the VP for Educational Activities for IEEE ITS Society for 2025-2026.



Zhong Cao (Member, IEEE) is an assistant research scientist in Civil and Environmental Engineering at the University of Michigan. He received his Ph.D. in Mechanical Engineering from Tsinghua University in 2020. His research interests include autonomous vehicle and intelligent transportation systems.

Xintao Yan (Member, IEEE) received his bachelor's degree from the School of Vehicle and Mobility at Tsinghua University, China, in 2018, and his Ph.D. degree from the Department of Civil and Environmental Engineering at the University of Michigan, Ann Arbor, in 2023. He is currently a Postdoctoral Research Fellow at the University of Michigan, Ann Arbor. His research focuses on enhancing the safety performance of connected and automated vehicles, with an emphasis on naturalistic driving environment modeling and automated driving system evaluation. He has received several honors, including the Exceptional Paper Award from the Traffic Signal Systems Committee at the 2019 TRB Annual Meeting, the ITS Best Paper Award from the INFORMS TSL Society in 2021, and recognition as a finalist for the 2024 IEEE ITSS Best Ph.D. Dissertation Award.

Shuo Feng (Member, IEEE) received the bachelor's and Ph.D. degrees in the Department of Automation at Tsinghua University, China, in 2014 and 2019, respectively. He was a postdoctoral research fellow in the Department of Civil and Environmental Engineering and also an Assistant Research Scientist at the University of Michigan Transportation Research Institute (UMTRI) at the University of Michigan, Ann Arbor. He is currently an Assistant Professor in the Department of Automation at Tsinghua University. His research interests lie in the development and validation of safety-critical machine learning, particularly for connected and automated vehicles. He was a recipient of the Best Ph.D. Dissertation Award from the IEEE Intelligent Transportation Systems Society in 2020 and the ITS Best Paper Award from the INFORMS TSL society in 2021. He is an Associate Editor of the IEEE TRANSACTIONS ON INTELLIGENT VEHICLES and an Academic Editor of the Automotive Innovation.

Qiujing Lu received the B.S. degree in automation from Tsinghua University, Beijing, China, in 2015, and the Ph.D. degree in electrical and computer engineering from the University of California at Los Angeles, Los Angeles, CA, USA, in 2022. She is currently a postdoctoral research fellow in the Department of Automation at Tsinghua University. Her research focuses on generative models and safe autonomous driving.