IEEE Open Journal of Intelligent Transportation Systems

Received XX Month, XXXX; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version 27 June, 2025.

Digital Object Identifier 10.1109/OJITS.2022.1234567

Foundation Models in Autonomous Driving: A Survey on Scenario Generation and Scenario Analysis

Yuan Gao[1], Mattia Piccinini[1], Yuchen Zhang[1], Dingrui Wang[1], Korbinian Moller[1], Roberto Brusnicki[1], Baha Zarrouki[1], Alessio Gambi[2], Jan Frederik Totz[3], Kai Storms[4], Steven Peters[4], Andrea Stocco[5], Bassam Alrifaee[6], Marco Pavone[7], Johannes Betz[1]

[1]Y. Gao, M. Piccinini, Y. Zhang, D. Wang, K. Moller, R. Brusnicki, B. Zarrouki, and J. Betz are with the Professorship of Autonomous Vehicle Systems and Munich Institute of Robotics and Machine Intelligence (MIRMI), Technical University of Munich, 85748 Garching, Germany (e-mail: johannes.betz@tum.de)
[2]A. Gambi is with the Austrian Institute of Technology (AIT), Vienna, Austria (e-mail: alessio.gambi@ait.ac.at)
[3]J.F. Totz is with AUDI AG (e-mail: jan.frederik.totz@audi.de)
[4]K. Storms, S. Peters are with the Institute of Automotive Engineering (FZD) at TU Darmstadt; Department of Mechanical Engineering, 64289 Darmstadt, Germany (e-mail: steven.peters@tu-darmstadt.de)
[5]A. Stocco is with the Technical University of Munich and fortiss GmbH, Munich, Germany (e-mail: andrea.stocco@tum.de—stocco@fortiss.org)
[6]B. Alrifaee is with the Department of Aerospace Engineering, University of the Bundeswehr Munich, Germany, (e-mail: bassam.alrifaee@unibw.de)
[7]M. Pavone is with Stanford University, Stanford, CA 94305, USA (e-mail: pavone@stanford.edu)

CORRESPONDING AUTHOR: Yuan Gao (e-mail: yuan_avs.gao@tum.de).

ABSTRACT Ensuring the safety of autonomous vehicles in real-world environments requires handling a wide spectrum of diverse and rare driving scenarios. Scenario-based testing addresses this need by offering a scalable and controlled approach to develop and validate autonomous driving systems. However, traditional scenario generation methods relying on rule-based logic, knowledge-driven models, or data-driven synthesis often yield limited diversity and unrealistic cases. With the emergence of foundation models, which represent a new generation of pre-trained, general-purpose Artificial Intelligence (AI) models, developers can process heterogeneous inputs (e.g., natural language, sensor data, maps, and control actions), enabling the synthesis, interpretation, analysis of complex driving scenarios. In this paper, we review the use of foundation models for scenario generation and scenario analysis in autonomous driving. Our survey presents a unified taxonomy that includes large language models, vision language models, multimodal large language models, diffusion models, and world models for the generation and analysis of autonomous driving scenarios, outlining their fundamental principles, applications, and corresponding evaluation metrics. In addition, we review the methodologies, open-source datasets, simulation platforms, and benchmark challenges. Finally, the survey concludes by highlighting the open challenges, research questions and promising future directions in applying foundation models to scenario generation and analysis in autonomous driving. All reviewed papers are listed in a continuously maintained repository, which is publicly available and updated with new research: GitHub.com/TUM-AVS/FM-for-Scenario-Generation-Analysis.

INDEX TERMS Autonomous vehicles, foundation model, scenario generation, scenario analysis, scenario based testing

I. Introduction

AUTONOMOUS DRIVING has seen rapid advancements in recent years, reaching a stage where human intervention is minimal or entirely unnecessary within specific Operational Design Domains (ODDs) [1]. Companies such as Waymo have successfully deployed fully autonomous robotaxi services [2] operating at SAE Level 4 [3] since 2018, demonstrating the feasibility of driverless mobility in specific urban environments. As of 2025, Waymo serves $250,000$ commercial rides per week [4]. These advancements have been driven by the development and rigorous validation of highly reliable modular Autonomous Driving (AD) software

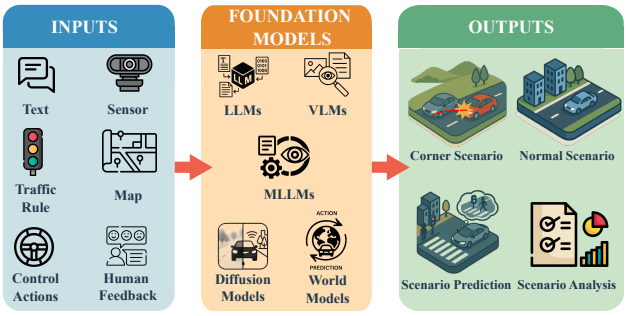

FIGURE 1. This survey critically analyzes existing FMs across LLMs, VLMs, MLLMs, DMs, and WMs for scenario generation and scenario analysis in autonomous driving.

functions, including perception, prediction, planning, and control [5]. In addition to the traditional modular architecture, end-to-end learning-based approaches [6], [7] have emerged, leveraging deep learning to process raw sensor data and directly generate trajectories or control actions [8].

Scenario-based testing in simulations is a key element for evaluating and validating the safety and performance of AD systems [9]. As a cost-efficient alternative to physical testing, it enables the simulation of realistic, reproducible, and controllable driving environments [10], and is particularly effective in replicating safety-critical situations, including rare corner cases that are often absent in real-world datasets [11], [12]. Therefore, the ability to systematically generate and analyze driving scenarios is crucial to scenario-based testing [13]. More specifically, generation focuses on creating diverse, safety-critical, and controllable driving situations for AD testing, while analysis concerns the evaluation of these scenarios in terms of safety, risk, and behavioral understanding to classify or select scenarios for testing, or to enhance overall AD performance.

Recent advances in Machine Learning (ML), especially the emergence of large-scale Foundation Models (FMs), offer new opportunities to enhance the realism, diversity, and scalability of scenario-based testing in autonomous driving. FMs were introduced by the Stanford Institute for Human-Centered Artificial Intelligence (HAI) in August 2021 [14] to describe a class of models trained on large-scale, diverse datasets typically using self-supervised learning. Unlike traditional ML models, which are often trained for specific, narrowly defined tasks, FMs can be adapted to a wide range of tasks through techniques such as prompting or fine-tuning. These models have demonstrated strong performance across various domains, including Natural Language Processing (NLP) [15], visual understanding [16], and code generation [17]. In the context of autonomous driving, FMs have recently garnered significant attention, as they combine general knowledge learned through large-scale pre-training with efficient adaptability to specific AD tasks like perception, palnning, control [18]–[20].

A. Scope of the Considered Literature

In this survey, we focus on publications addressing *scenario generation* and *scenario analysis* in the context of autonomous driving with Foundation Models (see Figure 1). Our survey is based on keyword searches in Google Scholar. The full list of keywords, as well as an overview of all referenced papers, is available in the GitHub repository of this paper[1].

To ensure both breadth and relevance, we included peer-reviewed conference and journal papers as well as preprints from arXiv. Although arXiv publications are not formally peer-reviewed, they often present timely and impactful research, especially in rapidly developing areas such as FM applications. Our survey covers papers published between October 2022 and May 2025, with a primary focus on venues in autonomous driving, computer vision, machine learning, and robotics. Figure 2 shows monthly trends in publication counts and their distribution by the thematic focus of the publication venues, e.g., conferences, journals, or preprint platforms.

B. Structure of the Survey

The structure of this survey is outlined in Figure 3. Section II provides an introduction to Foundation Models and a critical review of related surveys on scenario generation and analysis, encompassing both classical approaches and recent advances with Foundation Models. Sections III, IV, and V systematically examine language-based Foundation Models, beginning with fundamental concepts and followed by an in-depth discussion of recent applications of Large Language Models (LLMs), Vision Language Models (VLMs), and Multimodal Large Language Models (MLLMs) in scenario generation and analysis. Sections VI and VII address vision-centric Foundation Models, detailing the principles of Diffusion Models (DMs) and World Models (WMs) and their relevance to scenario generation. Section VIII surveys commen evaluation metrics, publicly available datasets, and simulation benchmarks pertinent to scenario generation and analysis in autonomous driving, and highlights major competition challenges in the field. Finally, Section IX and Section X identify open research questions and outline prospective research directions, while Section XI summarizes the key findings of this survey.

II. Related Work & Contributions

A. Development of Foundation Models

The term *Foundation Models*, introduced in 2021 [14], refers to general-purpose models trained on large-scale unlabeled data, designed to operate and generalize across a wide range of applications. Their cross-modal adaptability has the potential to enable tasks like Question Answering (QA), image captioning, sentiment analysis, information extraction, object recognition, and instruction following, combining generative abilities with deep contextual understanding.

[1] https://github.com/TUM-AVS/FM-for-Scenario-Generation-Analysis

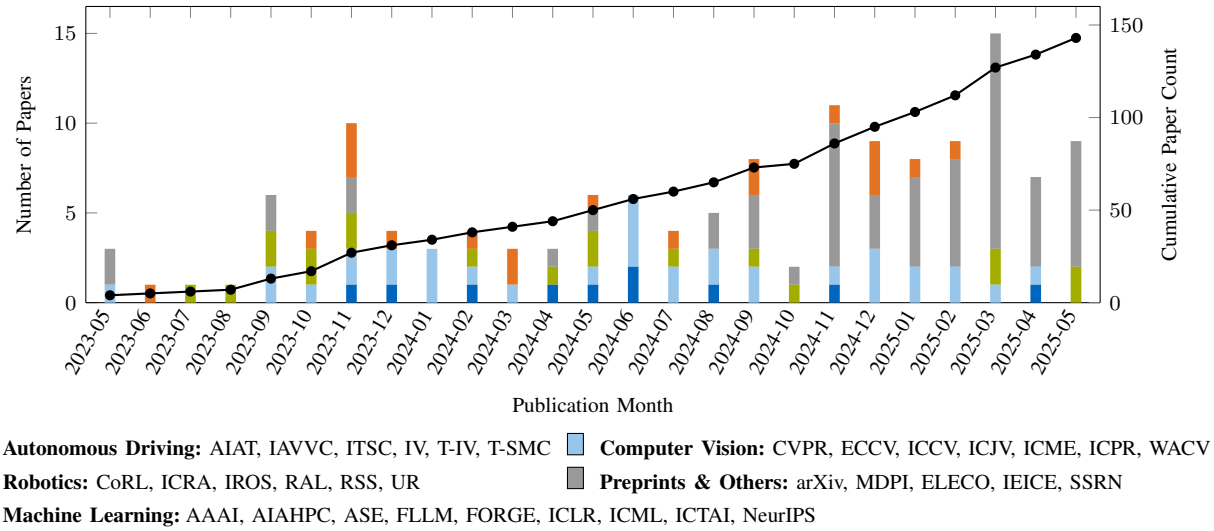

FIGURE 2. Timeline of foundation model-related publications in scenario generation and analysis, across selected journals, conferences and platforms between May 2023 and May 2025. Each bar represents the monthly count of papers, grouped by thematic category. The black line shows the cumulative number of papers over time (right axis). Note that the grouping refers to the general scope of the conference or journal, not to the content of the individual papers. For example, preprints listed on *arXiv* are categorized under *Preprints & Others*, although they may address topics from other categories.

Although FMs and generative AI are related, they represent distinct concepts: FMs are broad, adaptable systems, whereas generative AI focuses specifically on content creation.

In 2020, OpenAI released GPT-3 [21], a major milestone that popularized **LLMs**. The success of GPT-3 built upon the Transformer Architecture [22], whose self-attention mechanism facilitated efficient parallel training on long sequences. Subsequent models further refined this design, including BERT [15] (encoder-only for masked language modeling), GPT [23] (decoder-only for autoregressive generation), and T5 [24] (encoder-decoder for text-to-text transfer). Each employs self-supervised pre-training and serves as a backbone for downstream adaptation.

The principles of the transformer architecture were quickly extended beyond NLP, and enabled visual understanding [16], speech [25], tabular [26], and multimodal learning [27]. The extension of transformer architectures across different domains led to the development of **VLMs** such as Contrastive Language–Image Pre-training (CLIP) [28] and **MLLMs** such as LLaVA [29], that combine linguistic reasoning from LLMs with visual representations to produce cross-modal alignment and grounded understanding. More specifically, compact VLMs typically use an LLM as a backbone, augmented with a text tokenizer and a dedicated vision encoder to extract visual features. MLLMs further extend this paradigm by incorporating additional modality-specific encoders, such as audio, depth, or sensory inputs, and employ alignment modules to fuse these heterogeneous representations with the LLM backbone. The strong reasoning capability inherited from LLMs enables these models to perform complex multimodal inference, while the added visual and sensory encoders allow VLMs and MLLMs to operate effectively in real-world settings, where understanding both linguistic instructions and perceptual inputs is essential.

At the same time, advances in generative modeling developed **DMs**, and specifically Denoising Diffusion Probabilistic Models (DDPMs) [30] that generate high-quality samples using learned denoising processes. Subsequent variants, including improved DDPMs [31], Latent Diffusion Models (LDMs) [32], and Video Diffusion Models (VDMs) [33], further enhanced generation efficiency, controllability, and temporal coherence. Extending beyond image synthesis, DMs naturally support multimodal conditioning, enabling text-, audio-, and video-guided generation. Their high generation fidelity and flexible conditioning mechanisms make them powerful complements to transformer-based architectures, particularly in multimodal learning and world-modeling applications.

Finally, **WMs** [34] were developed to learn compact representations of interactive environments. Classical WMs combine vision encoders (e.g., Variational Autoencoders (VAEs)) with recurrent memory modules (e.g., Recurrent Neural Networks (RNNs)) and lightweight controllers (e.g., Evolution Strategies) to enable *future prediction*, such as forecasting video frames or rolling out latent state trajectories. Recent WMs designs integrate FMs into their components, e.g., by replacing encoders or memory modules with DMs [35] or LLMs [36], thus potentially unifying perception, reasoning, and generation in the same framework. Overall, this inter-connected evolution represents a progression from modality-specific FMs to general multimodal systems for holistic environmental understanding and interaction.

B. Foundation Models in Autonomous Driving

Recent studies have explored the integration of FMs into AD systems, exploiting their adaptability and multimodality across both modular and end-to-end architectures. Comprehensive surveys such as [18], [19]

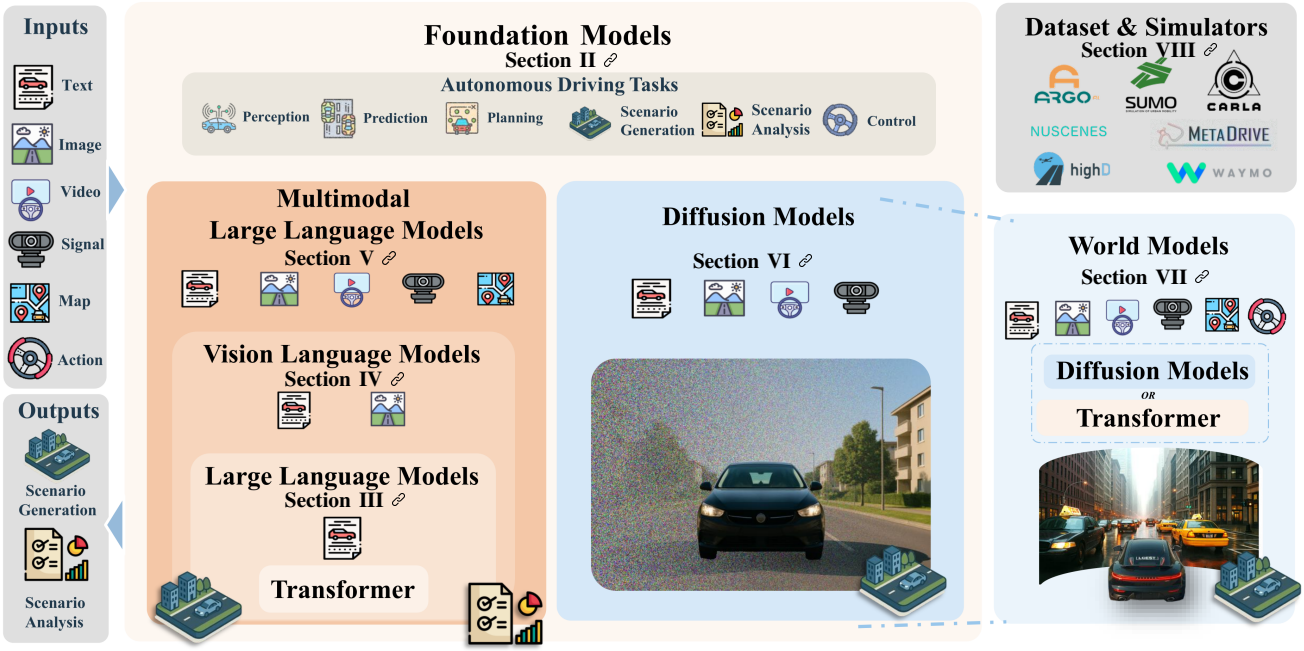

FIGURE 3. **Overview of FMs applied to scenario generation and analysis for autonomous driving, and the corresponding structure of this survey.**

offer a broad overview of the current landscape, covering LLMs, VLMs, MLLMs, DMs, and WMs.

LLMs in Autonomous Driving: The survey by Zhu et al. [37] reviews the integration of LLMs into modular autonomous driving systems, and focuses on perception, decision-making, control, and end-to-end approaches. Similarly, Wu et al. [38] investigate the use of LLMs for multi-agent perception, decision-making, and simulation. Finally, Li et al. [39] review the role of LLMs in enabling human-like reasoning in both modular and end-to-end AD systems and also emphasize training and integration strategies, which is not relevant to our tasks.

VLMs in Autonomous Driving: The survey [40] explores the use of VLMs across a range of AD tasks, where diffusion and world models are involved in scene understanding, visual reasoning, and dataset generation.

MLLMs in Autonomous Driving: Cui et al.'s survey [41] focuses on MLLMs architectures, modality fusion, and their applications across AD tasks. Fourati et al.'s [42] survey introduces XLMs as a combination of LLMs, VLMs, and MLLMs, providing a review of their use in AD that covers concepts, workflows, and techniques. Finally, Li et al's [43] survey explores LLM and MLLM applications across different AD modules, covering integration and training techniques.

DMs and WMs in Autonomous Driving: Guan et al. [44] provide an overview of world models in AD, focusing on their applications in scenario generation, motion prediction, and control. Driving WMs are further explored by Tu et al. [45], which categorize them into 2D scene evolution, 3D occupancy prediction, and scene-free paradigms.

Regarding the overlap with **generative AI**, Wang et al. [46] review generative models across the AD stack. While broad in scope, their survey adopts a model-centric perspective and lacks a focused discussion on scenario generation.

In summary, although the above works cover perception, planning, decision-making, simulation, and testing in AD, they do not explicitly address the roles of FMs in scenario generation or analysis, as this was not their primary focus. Our review aims to fill this gap.

C. Scenario Generation in Autonomous Driving

Scenario formats in AD range from annotated sensor data and multi-camera streams to map-based layouts, simulated urban scenes, and traffic-level environments, e.g., OpenScenario[2]. Figure 4 shows examples of driving scenarios in different formats. In the following, we review the existing surveys on scenario generation with classical and FM-based methods.

Surveys with Classical Approaches: Most of the existing reviews deal with classical methods (i.e., not FM-based) for scenario generation. Nalic et al [53] introduce knowledge-driven and data-driven generation approaches, and discuss safety metrics for scenario assessment. They also propose a six-layer model, which captures all essential components of a scenario. Ding et al. [54] categorize scenario generation methods into data-driven, adversarial, and knowledge-based approaches, providing detailed insights into the mechanisms underlying each. They also highlight the role of deep generative models for synthesizing image- and video-based scenarios with several papers across different models. In alignment with the ISO/WD PAS 21448 standard, Safety of the Intended Functionality (SOTIF), Schutt et al. [55] examine scenario generation

[2]https://www.asam.net/standards/detail/openscenario/

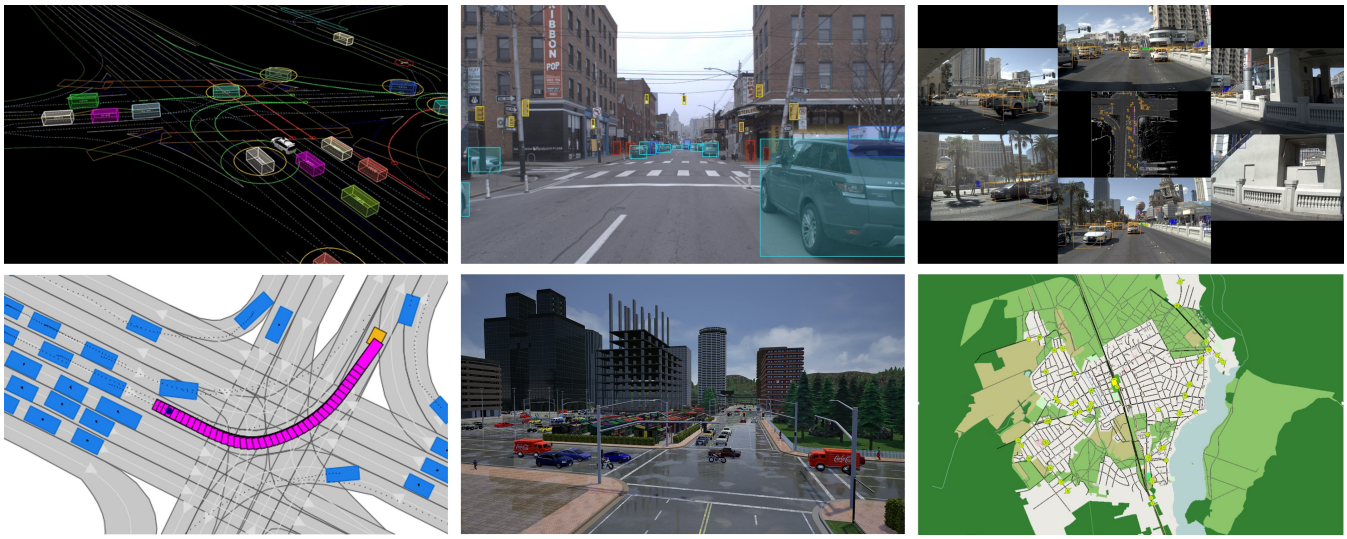

FIGURE 4. **Examples of driving scenarios in autonomous driving: datasets and simulations used for scenario-based testing. Sensor data such as camera images, videos, and LiDAR point clouds derived from these scenarios can be used to evaluate perception algorithms. Concurrently, simulator-specific scenario formats support rigorous testing of planning and control modules.**
Top row (left to right): Waymo Open motion [47] dataset; Argoverse2 [48] dash camera video; NuPlan [49] multi-camera views with map overlays.
Bottom row (left to right): CommonRoad [50] motion planning scenario; CARLA [51] simulated urban scenario; SUMO [52] large-scale traffic scenario.

across functional, logical, and concrete levels of abstraction. Their review includes machine learning-based generation, optimization-driven scenario exploration, scenario extraction from driving data, and manual scenario design.

Survey with FMs: Huang et al. [18] provide an overview of various types of foundation models and briefly discuss scenario generation. However, their analysis is limited to input modalities and model types, without addressing specific techniques or evaluation strategies.

Surveys with VLMs: Yang et al. [56] examine the use of LLMs and VLMs in tasks such as perception, question answering, and generation. They mention scenario generation using VLMs, DMs, and WMs but provide no clear distinction between these model types. While several evaluation metrics are cited, these are not organized by task or application. Tian et al. [57] present a more structured review of VLMs in autonomous driving across LLMs, VLMs, and WMs, focusing particularly on traffic simulation via VLM-guided generation and their integration with diffusion models. However, the survey lacks information about input modalities, scenario generation techniques, and the distinction of model types.

Surveys with DMs and WMs: Fu et al. [58] review video generation and WMs, covering diffusion-based, autoregressive, and reinforcement learning approaches. Feng et al. [59] focus on WMs, categorizing outputs into images, bird's-eye views, and 3D point clouds, and discuss evaluation metrics such as semantic segmentation and occupancy prediction. However, neither survey directly connects model outputs to scenario generation tasks. They also fail to distinguish between standalone DMs and WMs, and lack discussion of concrete techniques and evaluation strategies.

D. Scenario Analysis in Autonomous Driving

Scenario analysis involves the systematic evaluation of driving scenarios (Figure 4). It encompasses tasks such as scenario evaluation, scene understanding, risk assessment, anomaly detection, and accident prediction. Further, it includes identifying safety-critical situations, evaluating system robustness, and supporting informed decision-making in both simulation and real-world environments.

Surveys with Classical Approaches: Riedmaier et al. [10] propose a taxonomy of scenario-based safety assessment methods, including knowledge-based, data-driven, and falsification-based approaches. They emphasize the use of key performance indicators (e.g., time-to-collision, required deceleration) as proxies for accident risk and advocate for the integration of formal methods into safety validation.

Mahmud et al. [60] review proximal surrogate indicators such as time-to-collision, post-encroachment time, and deceleration rate to avoid a crash. They categorize these metrics and identify key research challenges, including metric standardization, real-world validation, and integration into simulation-based scenario analysis frameworks.

Survey with FMs: Huang et al. [18] briefly mention scenario analysis under the term "perception data annotation", but do not categorize tasks based on their goals. Additionally, they neither associate datasets with individual studies nor discuss modality transformations; as such, their review does not focus on pre-trained FMs.

Surveys with VLMs: Yang et al. [56] address scenario analysis in the context of question answering, focusing mainly on dataset descriptions. Their analysis remains limited, as it lacks discussion of input modalities, methodological approaches, model taxonomy, and evaluation metrics. Similarly, Tian et al. [57] consider Visual Question Answering

TABLE 1. Comparison of surveys on FMs for scenario generation and analysis in AD. Our survey is the first to cover all FM types, scenario categories, input modalities, datasets, model types, techniques, and evaluation metrics.

Survey	LLM	VLM	MLLM	DM	WM	Scenario Generation[1]								Scenario Analysis[2]						
						Scenario Category	Input Modality	Dataset	Scenario Controllability	Model	Technique	Metric	[n/m]	Scenaio Category	Input Modality	Dataset	Model	Technique	Metric	[n/m]
2023 Huang. [18]	✓	✓	✓	✓	✓		✓			✓			13/261		✓		✓			5/261
2024 Yang. [56]	✓	✓					✓					✓	11/155	✓			✓			13/155
2024 Fu. [58]					✓		✓			✓			11/114							
2024 Tian. [57]	✓	✓			✓	✓		✓		✓			15/124			✓	✓			9/124
2025 Feng. [59]					✓					✓	✓		13/166							
Our Work	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**93/348**	✓	✓	✓	✓	✓	✓	**56/348***

[1] **Scenario Category**: e.g., safety-critical scenario; **Input Modality**: e.g., text, image, sensor signal; **Dataset**: e.g., nuScenes; **Scenario Controllability**: full (script), partial (trajectory); **Model**: e.g., GPT, latent diffusion; **Technique**: e.g., prompting; **Metric**: e.g., realism.

[2] **Scenario Category**: e.g., evaluation, risk assessment; **Input Modality**: e.g., text, image, sensor signal; **Dataset**: e.g., nuScenes; **Model**: e.g., GPT; **Technique**: e.g., zero-shot, adapter layer; **Metric**: e.g., accuracy, language generation quality.

[n/m] = number of papers using FMs (large pre-trained models) / total papers reviewed in the survey.

* The 348 papers are categorized as follows: 93 on scenario generation, 56 on scenario analysis, 58 on datasets, 21 on simulators, 25 on benchmark challenges, and 95 on other related topics (e.g., FMs' implementation).

(VQA) as a form of scenario analysis using VLMs, but cover a small number of resources and provide minimal discussion.

E. Critical Summary

To the best of our knowledge, existing surveys on FM-based scenario generation and/or anaysis in autonomous driving are limited by the following aspects (summary in Table 1):

- **Lack of focus on scenario generation:** None of the reviewed surveys explicitly focuses on scenario generation using FMs. When addressed, scenario generation is either mentioned briefly (e.g., [18], [43]) or discussed without in-depth analysis of generation techniques, scenario controllability, or evaluation metrics (e.g., [56]–[58]).
- **Incomplete coverage of scenario analysis:** Tasks such as scenario understanding, evaluation, and risk assessment are overlooked. When addressed (e.g., [56] and [57]), the analysis is typically reduced to question answering, with little attention paid to task-specific models, methods, or evaluation strategies.
- **Limited connections among modalities and tasks:** While several surveys consider the input modalities of FMs, they do not establish clear links between these modalities and the techniques, models, and datasets used for scenario generation and analysis.
- **Absence of a structured classification:** No prior work presents a structured classification of FMs that spans both scenario generation and scenario analysis, considering pre-trained model types, adaptation methods (e.g., prompting, fine-tuning), input modalities, datasets, and evaluation metrics.

F. Contributions

To address the limitations of the existing literature reviews, this survey evaluates the landscape of FMs in the fields of scenario generation and scenario analysis (Table 1). In summary, this work provides the following key contributions:

1) We present the first review on the use of **FMs for scenario generation and analysis** in AD.
2) **Structured classification of existing methods**: Our work offers a structured classification covering all FM types (i.e., LLMs, VLMs, MLLMs, DMs, WMs), scenario categories, input modalities, model types, datasets, techniques, and evaluation metrics.
3) **Review of datasets, simulation platforms and existing benchmarking competitions**: We review the openly-accessible datasets and simulators used for scenario generation and analysis. Meanwhile, we provide the first review on benchmarking competitions for FMs in AD.
4) **Identification of open challenges and future directions**: We identify open research challenges in applying FMs to scenario-based testing. By leveraging our analysis, we propose future research directions to improve the adaptability and robustness of FM-driven approaches in scenario generation and analysis.

III. Large Language Models (LLMs)

This section introduces the foundation and evolution of LLMs, presents key technological advancements, and reports on common adaptation techniques (e.g., prompt engineering and fine-tuning strategies). We then explore how LLMs support scenario generation, safety-critical cases, real-world scene synthesis, driving policy evaluation, closed-loop simulation, and Advanced Driver Assistance Systems (ADAS) testing. The section concludes with a discussion on scenario analysis,

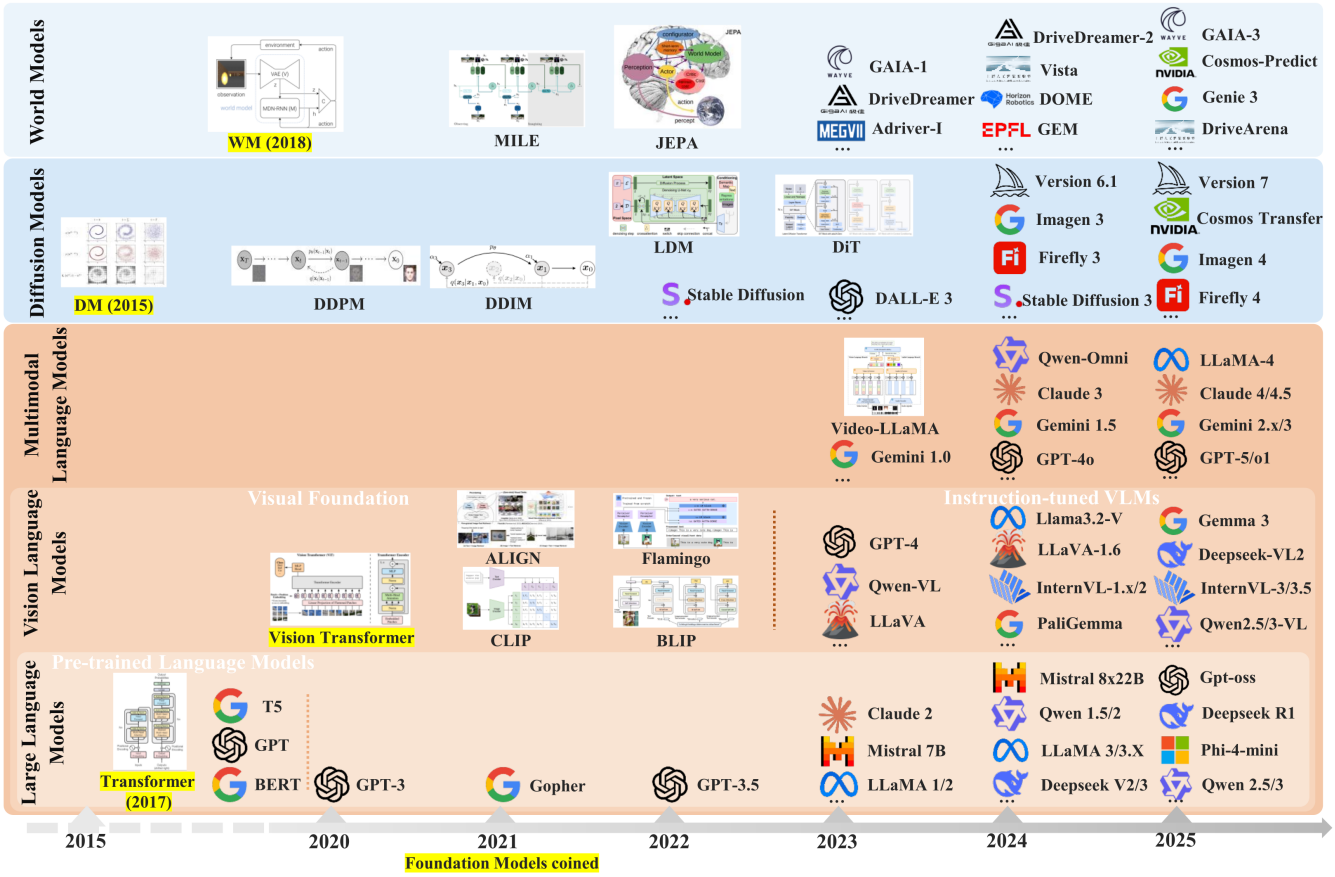

FIGURE 5. Timeline of the development of FMs. LLMs, including transformer-based architectures (e.g., BERT), are shown at the bottom. VLMs built upon visual FMs (e.g., ViT and CLIP) are illustrated in the middle layer, together with instruction-tuned VLMs that enable interactive, chat-style vision–language reasoning. MLLMs are shown above. In parallel, the evolution of visual FMs is highlighted through DMs (e.g., the first diffusion model [61]) and WMs (e.g., the world model architecture [34]). Highlighted entries indicate key conceptual milestones.

including question answering, scenario understanding, and scenario evaluation.

A. Development of LLMs

The most prominent category of FMs are LLMs, which focus on the text modality and are built upon the transformer architecture [22]. A defining characteristic of these models is their use of self-supervised learning, where they learn language representations by predicting masked or missing parts of text from large-scale unlabeled corpora. This paradigm has enabled models to capture rich contextual and semantic information without the need for manual annotation. The foundation for this approach was laid by static word embeddings [62], which evolved into pre-trained language models such as GPT [23], BERT [15], and T5 [24]. These models replaced static embeddings with dynamic, context-aware representations learned directly from text. The release of GPT-3, with 175B parameters [21], marked a major milestone in self-supervised language modeling by demonstrating strong generalization and few-shot learning capabilities, significantly reducing the need for task-specific fine-tuning compared to earlier generations. The historical evolution of text-only pre-trained language models and LLMs is summarized in Figure 5.

OpenAI's 2020 scaling laws [63] showed that LLMs performance improves predictably with increased model size, data, and compute, fueling the trend toward ever-larger foundation models. However, DeepSeek [64] challenged this assumption by demonstrating that data quality and alignment matter as much as scale. Through supervised fine-tuning on synthetic expert data and reinforcement learning via Group Relative Policy Optimization (GRPO), they trained smaller models that achieved competitive performance.

Since LLMs are pre-trained on large-scale unlabeled data, various adaptation techniques such as prompt engineering and fine-tuning have been developed to tailor LLM's behavior to specific tasks. Sahoo et al. [65] provide an overview of these techniques in their recent survey. Here, we focus on the adaptation techniques that were applied in the context of driving scenario generation and analysis.

Prompt Engineering: This refers to designing and structuring input prompts to guide a pre-trained language model toward producing desired outputs, without modifying its internal parameters. The different techniques are:

1) Contextual Prompting (CP): Augments prompt with task-specific context or background information, thereby helping the model align more closely with the intended application domain.

(2) Chain-of-Thought (CoT): Encourages the model to generate intermediate reasoning steps before producing a final answer. This structured reasoning enhances the logical consistency of the model's reasoning chain, which is particularly beneficial for complex, multi-step tasks.

(3) In-Context Learning (ICL): Involves task demonstrations (e.g., *one-shot*, or *few-shot*) in the prompt to guide the model towards the correct task behavior.

(4) Self-Consistency (SC): A decoding strategy that samples multiple outputs for a given prompt and selects the most frequent or consistent one, improving answer robustness and reliability.

(5) Retrieval-Augmented Generation (RAG): Enhances the performance on specific tasks by retrieving external knowledge from a database at inference time. A retrieval component identifies the relevant documents to condition the model's response, thereby improving its accuracy.

Fine-Tuning: These techniques train the model on datasets to improve its ability for specific tasks. The different fine-tuning techniques are Full Fine-Tuning (FFT) and Parameter-Efficient Fine-Tuning (PEFT). FFT updates all model parameters using domain-specific data. While effective, it requires significant computational resources and has limited scalability. PEFT updates only a small portion of the model's parameters, while keeping most of the model frozen. A specific PEFT method is *Low-Rank Adaptation (LoRA)* , which injects trainable low-rank matrices into the attention modules of the model, enabling adaptation with minimal parameter updates and reducing computational cost. For instance, full fine-tuning of GPT-3 requires updating all 175B parameters, whereas LoRA can achieve comparable performance by training only around 37.7M parameters [66].

Additionally, more advanced techniques to adapt LLMs exist in the reviewed papers. These include *multi-stage prompting* [67], [68] and *Multi-LLM Agent Systems (MLAs)* [69], which coordinate multiple interacting LLMs to solve complex tasks collaboratively. Tooling frameworks such as LangChain [70] facilitate the construction of modular, agent-based architectures that extend beyond traditional single-prompt interactions.

B. LLM-Based Scenario Generation

Advances in LLMs have triggered the development of LLM-driven scenario generation to test intelligent vehicle systems. Based on their individual objectives, we classify the existing works into six categories and list representative works within each category in Table 2:

Safety-Critical Scenario Generation: A key application of LLM-based scenario generation lies in the creation of safety-critical scenarios. Often termed "corner cases", "long-tail", or "Out-of-Distribution (OOD)" situations, these scenarios involve high collision risk, abnormal agents' behavior, or reduced safety margins [11]. Recent LLM-based approaches can synthesize rare trajectories and scene configurations beyond nominal driving conditions, to stress-test the robustness of AD systems and uncover residual safety risks. Unlike ADAS test scenarios, this category neither targets specific assistance functions nor follows predefined regulatory test protocols.

The work LLMScenario [71] focuses on safety-critical scenario generation based on the HighD dataset [72] with GPT-4. They use ICL, incorporating critical examples, which are evaluated based on reality and rarity, to guide the generation. Using CoT and SC prompting, the framework generates safety-critical trajectories step-by-step in MetaScenario [73]. ChatScene [74] uses GPT-4 with RAG to translate textual safety-critical descriptions into domain-specific language (DSL) scripts such as Scenic [75] for CARLA [51]. Its retrieval database is built using Sentence-T5 embeddings that map behaviors and geometric patterns to code snippets. These snippets are then retrieved through RAG and assembled into complete Scenic scripts. Building on structured generation, Aasi et al. [69] propose a multi-agent pipeline that constructs a branching tree of OOD scenarios using CoT prompting. Their Augmenter-LLM, based on GPT-4o, translates descriptions into CARLA scene configurations, which contain maps, weather, objects, and behaviors via API calls. A VLM then classifies the simulated scenes by OOD type to identify the safety-critical scenarios.

The methods discussed above operate open loop. In contrast, Mei et al. [76] focus on online interactive scenario generation using Waymo Open Motion Dataset [47]. Their retrieval-augmented framework uses DeepSeek-V3 and DeepSeek-R1 to infer risky behaviors of a vehicle in real time and synthesize adversarial trajectories for it to collide with the ego vehicle. A memory module stores and retrieves intent–planner pairs, allowing continuous refinement and adaptation of the generated scenarios.

Despite promising advances, current works often operate offline or focus on limited risk types, limiting their generalizability to complex, multi-agent contexts. Future work could integrate interactive generation, enhance safety verification in simulation, and develop evaluation pipelines by leveraging VLMs to assess the plausibility and criticality of the generated scenarios.

Real-World Scenario Replication: Creating realistic driving scenarios is challenging due to the difficulty of accurately reproducing real-world conditions. A common strategy involves replaying recorded driving data in simulation environments or leveraging real crash reports to reconstruct the corresponding events. Realistic traffic scenes can also be replicated by grounding them on real-world maps, thereby preserving authentic infrastructure, road layouts, and environmental features.

LCTGen [77] leverages GPT-4 with ICL and CoT prompting to convert crash report into structured YAML-like

descriptions. Then a retriever module matches these structured descriptions with relevant maps from the Waymo Open Dataset [78]. These "map-grounded" inputs are then processed by a generative model using multi-layer perceptrons and learned masks to produce realistic driving scenarios. In Chat2Scenario [79], recorded datasets from HighD with user-defined criticality and textual descriptions are used as input. They use a templated contextual prompting scheme with GPT-4 and retrieve relevant scenarios that match the user's input with ASAM OpenScenario [80] format. For microscopic simulation, ChatSUMO [81] utilizes Llama 3.1 [82] with template-based prompts to extract user requirements for traffic volume, city, and network type. Then, ChatSUMO translates these parameters into SUMO [52] configurations, with osm [83] maps retrieved through RAG. Simulation outputs, including traffic density, travel time, and emissions, are visualized and summarized via a Streamlit[3]. SeGPT [84] synthesizes diverse and challenging test data from recorded trajectories. Their framework supports large-scale scenario synthesis and compares zero-shot prompting with CoT to evaluate LLM-guided generation performance on the dataset from [85].

The reviewed papers generate scenarios from recorded data and crash reports. One of the future directions is to first generate scenarios from recorded datasets, and then incorporate natural language descriptions as feedback. This hybrid approach could significantly enhance both the realism and diversity of the resulting scenarios by aligning data-driven generation with human-intuitive requirements.

Driving Policy Test Scenario Generation: Driving policy test scenario generation focuses on evaluating automated driving policies, such as motion planners or controllers, under systematically constructed traffic situations. Recent LLM-based methods generate executable scenarios that are deployed in simulations to assess planning behavior, policy robustness during algorithm development.

In LCTGen [77], generated real-world crash scenarios are used to assess the performance of a motion planner within the MetaDrive [86] simulator. In TTSG [87], GPT-4o-generated scenarios are used for multi-agent planning validation in critical scenarios. Specifically, they constructed a road and agent database using RAG with LLMs and proposed ranking strategies to select the best-fitting road based on the agent's behavior. In contrast, AutoSceneGen [88] uses a code-designed filter to select the valuable parts of the scenario description based on simulation documents and ICL, and adds scenario examples to the prompt. A code-based validator then transfers and verifies the GPT-4 output, which directly generates DSL-style configuration code compatible with CARLA. The resulting scenarios are subsequently used to evaluate the performance of a motion planner.

Closed-loop Scenario Generation: Recent works address the limitations of static datasets by introducing closed-loop scenario generation with LLMs. Closed-loop scenarios enable the validation of multi-agent interactions and ego-reactive behaviors.

ProSim [89] presents a promptable closed-loop simulation framework, where prompts such as goal points, route sketches, action tags, and natural language instructions are used to guide an agent's behavior. Llama3.1-8B is fine-tuned with LoRA to generate policy tokens, and a lightweight policy module rolls out the agent's trajectories in a closed loop within the Waymo simulator. In LLM-attacker [90], an adversarial scenario generation is proposed. It employs three coordinated modules based on Llama3.1-8B, for initialization, reflection, and modification, to identify and refine adversarial vehicle behaviors using CoT. These modules iteratively generate and adjust the attacker's trajectories to collide with the ego vehicle. Their framework is trained with reinforcement learning in a closed-loop setting using the Waymo Open Dataset. In contrast, CRITICAL [91] focuses on ego-agent policy learning without adversarial agents. It integrates Mistral-7B via LangChain [70] into a standard reinforcement learning loop in the HighwayEnv environments [92]. Their LLM is used to generate diverse scenario configurations, e.g., vehicle density, number of cars, and to shape safety-related rewards, enabling robust policy learning under different conditions.

Together, these works demonstrate complementary strategies: ProSim enables fine-grained control and interactivity, LLM-Attacker focuses on adversarial testing, and CRITICAL supports LLM-guided training environments. Future research could benefit from unifying these paradigms into a single framework that supports diverse behavior modeling, adversarial robustness, and controllable training environments.

Image Datasets Generation: Real-world camera datasets are widely used in autonomous driving research, but often lack the diversity and editability required for generating specific test cases. To address this, recent work explores language-guided editing of recorded images.

ChatSim [93] introduces a collaborative multi-agent framework with GPT-4, where each LLM agent handles a specialized scene editing task, such as viewpoint changes, vehicle manipulation, asset insertion, and motion planning, based on natural language instructions. ChatSim leverages neural rendering and lighting estimation to achieve photorealistic, multi-camera scene synthesis with external digital assets.

ADAS Test Scenario Generation: ADAS test scenario generation focuses on translating functional descriptions derived from regulations, standards, and test protocols into executable and reproducible test scenarios for function-level evaluation. Recent LLM-based approaches parse regulatory text, specifications, or reports into structured representations, and generate logical or concrete scenarios in domain-specific languages for simulation-based testing of ADAS stacks, such as Apollo [94]. This category emphasizes standardization, reproducibility, and coverage, and is closely aligned

[3]https://streamlit.io/

TABLE 2. Summary of Scenario Generation Studies Using Large Language Models.

Category	Input				Model	Technique[1]	Simulator	Output[2]	Paper
	Trajectory	DSL	Dataset	Database					
Safety-critical Scenario	✓		HighD [72]		GPT-4	CoT, ICL, SC	Metascenario [73]		LLMScenario [71]
		✓		Map Position Behaviors	GPT-4	CoT, ICL, RAG	CARLA [51]		ChatScene [74]
			CARLA		GPT-4o Claude 3.5 Sonnet Gemini 1.5 Pro	CoT, MLAs	CARLA		Aasi et al. [69]
	✓		WOMD [47]	Trajectory Behaviors	DeepSeek V3 DeepSeek R1	CoT, CP, ICL, RAG	WOMD		Mei et al. [76]
Real-world Scenario Replication			Waymo Open [78]	Map NHTSA [95]	GPT-4	CoT, ICL, RAG	MetaDrive [86]		LCTGen [77]
			HighD		GPT-4	CoT	Esmini CarMaker		Chat2Scenario [79]
			SUMO, OSM [83]	Map	Llama 3.1-8b	CoT, RAG	SUMO [52]		ChatSUMO [81]
	✓		Interaction [85]		GPT-4	CoT	Interaction		SeGPT [84]
Driving policy Testing Scenario			CARLA	Map Position Behaviors	GPT-4o	CoT, RAG	CARLA		TTSG [87]
		✓	CARLA		GPT-4	CoT, CP	CARLA		AutoSceneGe [88]
Closed-loop Scenario			Waymo Open	Map States	Llama 3.1-8b	LoRA	Waymo Sim		ProSim [89]
	✓		Waymo Open MetaDrive [86]		Llama 3.1-8b	CoT, RL	MetaDrive		LLM-attacker [90]
	✓		HighD		Mistral-7B	CoT	HighwayEnv [92]		CRITICAL [91]
Image Dataset			Waymo Open	Images	GPT-4	MLAs			Chatsim [93]
ADAS Testing Scenario		✓		Regulation	GPT-4	ICL	SUMO		Guzay et al. [96]
		✓		Traffic rule	GPT-4	Multi-stage, ICL, CoT, CP	CARLA		TARGET [67]
		✓		Standard Regulation Test Specification	GPT-4 Llama 3	CP, Multi-stage	CARLA		Petrovic et al. [97]
			LGSVL [98]	NHTSA	GPT-4	CP, ICL	LGSVL		SoVAR [99]
		✓		NHTSA	GPT-4	CP, ICL, Multi-stage	LGSVL		LeGEND [68]
		✓		NHTSA OpenXOntology	GPT-4	CoT, ICL, SC, Multi-stage	CARLA		Text2Scenario [100]
		✓	OpenDRIVE	UNECE R157 OpenX Ontology maps	Llama 3.1	CP	VTD		Zhou et al. [101]

[1] Techniques: CoT = Chain-of-Thought prompting; ICL = In-Context Learning; SC = Self-Consistency; CP = Contextual Prompting; RAG = Retrieval-Augmented Generation; RL = Reinforcement Learning; LoRA = Low-Rank Adaptation; MLAs = Multi-LLM Agent Systems.

[2] Output: Video, Trajectory, Scenario script.

with regulatory and assessment frameworks such as OpenXOntology and UN R157.

One of the pioneering papers on ADAS test scenario generation using LLMs is Guzay et al. [96], who use GPT-4 with ICL to convert regulatory descriptions into SUMO-compatible XML files. Expanding on this, TARGET [67] introduces a multi-stage prompting by using GPT-4 that parses traffic rules into a DSL of CARLA using CoT and ICL to evaluate multiple ADAS software. A rule-to-script generator then produces a scenario script. Petrovic et al. [97] extend this direction by processing ADAS test topologies and standardization documents from UNECE R157 [4]. The test topology is converted into a metamodel that includes elements such as the environment, sensor, and actuator configurations. Standardization documents are parsed into Object Constraint Language (OCL) using LLMs. Based on the combined metamodel, OCL constraints, and a specific test description, an LLM (e.g., GPT-4 or Llama 3) is used to generate DSL test scenarios, which are then simulated

[4] https://unece.org/transport/documents/2021/03/standards/un-regulation-no-157-automated-lane-keeping-systems-alks

in CARLA. In a more data-driven approach, SoVAR [99] reconstructs crash scenarios from NHTSA [95] reports by extracting structured attributes with GPT-4 and generating trajectories and simulation scripts via constraint solving, producing LGSVL [98]-compatible test scenarios via API calls focused on realism. In contrast, LeGEND [68] follows a top-down approach: it abstracts reports into functional scenarios, transforms them into logical DSL representations via a two-stage GPT-4 pipeline, and applies multi-objective search to generate diverse and critical scenarios to evaluate Apollo ADAS stack.

Text2Scenario [100] introduces a standardized hierarchical scenario repository based on the SOTIFs framework and applies multi-stage prompting (CoT, ICL, SC) with GPT-4 to generate logical scenarios from free-form descriptions. The resulting logical scenario is then converted into the OpenScenario format through code and simulated in CARLA to evaluate the performance of multiple ADAS stacks. Finally, Zhou et al. [101] focus on lane-keeping systems by using Llama 3.1 and prompt templates to extract scene elements from UNECE R157 aligned descriptions. These descriptions are structured and converted into OpenScenario DSL files using OpenXOntology[5] and OpenDRIVE[6], then simulated in the VTD[7] simulation environment.

While LLM-based frameworks effectively generate ADAS test scenarios from crash reports and regulations, they often overemphasize rare edge cases [68], [99], neglecting common driving scenarios that are essential for broader testing. Currently, we are missing the incorporation of routine test cases and utilizing real-world maps from OpenStreetMap (OSM) or SUMO to enhance the scenario diversity and fidelity.

C. LLM-based Scenario Analysis

Recent research has explored the use of LLMs as a scenario analysis tool and method. A key challenge is that LLMs are primarily designed to process natural language input, whereas driving scenarios are typically defined using structured data formats, such as scripts in DSL or sensor outputs with predefined syntax. This creates a mismatch between how the scenario's information is represented and how LLMs operate. Bridging this gap is critical to enable effective interpretation of driving scenarios using language models. We classify the existing works into three key areas and list representative works in Table 3.

Question Answering (QA): Applying LLMs to scenario analysis for AD requires domain-specific knowledge, which general-purpose pre-trained models may lack. To bridge this gap, fine-tuning with tailored datasets is essential. QA datasets describing driving scenarios help LLMs interpret structured driving contexts, and support downstream tasks like trajectory planning and decision-making.

A notable example is [102], where the authors automate the generation of QA datasets with driving scenarios using GPT-3.5. With a structured language generator, they convert vectorized scenario data from their in-house dataset, including agents' positions, speed, and distance, into natural language. With ICL and pre-defined driving rules, their model generates diverse, context-aware QA pairs to reflect realistic driving situations. The QA dataset of [102] focuses primarily on perception and prediction.

Scenario Understanding Here, the LLM processes structured sensor or simulator data, such as agent states, road layouts, and traffic signals, to support tasks like scenario captioning (concise descriptions) and reasoning (coherent narratives capturing intent and context).

The SenseRAG [104] introduces a RAG-based framework from the DLR urban traffic dataset [103] for scenario understanding. They use a VLM to generate traffic condition descriptions into textual descriptions, which are then mapped to a structured database, including additional structured information with weather, city, and traffic participants. Using CoT prompting and Structured Query Language (SQL) query generation, GPT-4 retrieves and reasons over the data to refine perception and enhance trajectory prediction.

Scenario Evaluation Recent work demonstrates how LLMs can support the evaluation of driving scenarios, by reasoning over structured simulation data or scenario images converted into natural language. This includes the evaluation of anomaly detection, scenario realism, safety-criticality, and driving behavior. Elhafsi et al. [105] detect semantic scenario anomalies using LLMs. Their scenarios are evaluated using OpenAI's text-davinci-003, which is prompted with CoT and ICL. Reality Bites [107] is one of the first works to evaluate the reasoning ability of LLMs in assessing scenario realism. It transforms XML-formatted DeepScenario [106] data into natural language and uses ICL prompting with models like GPT-3.5, Llama2-13B, and Mistral-7B to judge the alignment with realistic driving conditions. Gao et al. [108] propose a framework to analyze safety-criticality in driving scenarios from the CommonRoad [50] environment. They convert structured scenario data into natural language and prompt LLMs via CP, CoT, and ICL to evaluate the safety-criticality of the scenario and infer the risk level of the agent. Also, they generate safety-critical scenarios by modifying the trajectories of identified adversarial vehicles. Meanwhile, You et al. [109] focus on holistic driving assessment, converting interview and simulation data into a structured knowledge database for RAG. In their framework, GPT-4o classifies driving styles (cautious, aggressive) and performance levels based on aggregated context, including scenario-level information like weather, ego vehicle data, and surrounding traffic participants.

Overall, LLM-based scenario evaluation still depends on token-heavy prompting and handcrafted prompts. Emerging reasoning models, such as OpenAI GPT-5 and Gemini 2.5 Pro, may enable more efficient, zero-shot approaches.

[5]https://www.asam.net/standards/asam-openxontology/

[6]https://www.asam.net/standards/detail/opendrive/maps

[7]https://hexagon.com/de/products/virtual-test-drive

TABLE 3. Summary of Scenario-Analysis Studies Using Large Language Models.

Category	Input				Model	Technique[1]	Focus	Paper
	Scenario Elements	Elements Narrator	Dataset	Database				
Question Answering (QA)	Road, Ego, NPC Vehicles, Pedestrians	Language Generator	In-house		GPT-3.5	ICL	Driving QA	Chen et al. [102]
Scenario Understanding	Road, Weather, Ego, Traffic Light, NPC Vehicles	LLaVA, Language Parsing	DLR UT [103]	Traffic Condition Structured Data	GPT-4	CoT RAG	Reasoning	SenseRAG [104]
Scenario Evaluation	Road, Weather, Traffic Sign, NPC Vehicles	OWL-ViT, Language Parsing	CARLA		text-davinci-003	CoT ICL	Anomaly Detection	Elhafsi et al. [105]
	Road, Weather, Ego, NPC Vehicles	Vector Parse	DeepScenario [106]		GPT-3.5 LLaMA2-13B Mistral-7B	CP ICL	Realism	Reality Bites [107]
	Road, Ego, NPC Vehicles	Cartesian Parser, Ego Parser	CommonRoad [50]		GPT-4o Gemini-1.5Pro Deepseek-V3	CP CoT ICL	Safety-Criticality	Gao et al. [108]
	Road, Ego, Traffic Light, NPC Vehicles	Not Specified	CARLA	Interview Data	GPT-4o	CoT RAG	Driving Styles	You et al. [109]

[1] Techniques: CoT = Chain-of-Thought prompting; ICL = In-Context Learning; CP = Contextual Prompting; RAG = Retrieval-Augmented Generation.

D. Limitations and Future Directions

Our review of LLM-based scenario generation and analysis reveals that many existing approaches rely heavily on prompting strategies, as summarized in Table 2 and Table 3. The effectiveness of the corresponding frameworks often depends on manually crafted prompts. To mitigate this dependency, recent tools such as DSPy [110] provide AI-driven prompt optimization frameworks that automatically generate task-aligned prompts based on user-defined evaluation metrics. Another promising direction involves leveraging advanced reasoning models, such as OpenAI's GPT-o1 and DeepSeek-R1, which offer stronger zero-shot reasoning capabilities and may reduce the reliance on handcrafted prompts.

Furthermore, future research should explore moving beyond single-turn prompting by adopting interactive, dialogue-based generation. Structuring LLMs as chatbot-style agents would allow users to iteratively define scenario requirements, enabling the synthesis of customized, constraint-compliant scenarios rather than relying on static outputs.

LLM-based Scenario Generation: A significant gap persists between simulation-based scenario generation and real-world validation. Bridging this gap requires the development of ADAS test scenarios aligned with practical safety standards such as SOTIF. By leveraging the reasoning capabilities of LLMs, future systems could generate functional and logical scenarios directly from textual descriptions and test specifications. This would facilitate the creation of challenging, safety-critical corner cases and enhance the applicability of generated scenarios to real-world testing and system validation.

LLM-based Scenario Analysis: LLMs are also increasingly used to understand and analyze driving scenarios. While many innovative frameworks have emerged, a major limitation lies in computational inefficiency. Since most LLMs operate on textual inputs, sensor data from modalities such as LiDAR, images, and radar must first be converted into natural language descriptions using narrators or intermediate modules, as shown in Table 3. This pre-processing step adds latency and increases the input complexity. Moreover, improving the quality of the analysis often requires complex prompting strategies such as chain-of-thought reasoning, further complicating real-time deployment. To address these challenges, one promising approach is to fine-tune LLMs for scenario understanding tasks, avoiding reliance on elaborate prompting. However, this direction is currently hindered by the lack of large-scale, high-quality scenario question-answering datasets and evaluation benchmarks: most works focus on framework validation rather than dataset creation.

IV. Vision Language Models (VLMs)

This section introduces VLMs, summarizes their key adaptation techniques, and reviews VLM-based scenario generation for safety-critical, real-world, and ADAS testing applications, and image datasets generation. Additionally, it explores how VLMs support scenario analysis tasks such as VQAs, scene understanding, benchmarking, and risk assessment.

A. Development of VLMs

In 2020, the ViT [16] extended the transformer architecture from NLP to computer vision by splitting an image into fixed-size patches. This enabled embedding an image as

a sequence of tokens and processing the sequence of tokens with a standard transformer encoder. This success inspired researchers to combine visual and textual modalities, leading to the development of VLMs, which now can jointly process images and text at the same time. A milestone was the development of CLIP [28], which was trained on hundreds of millions of image–text pairs using a contrastive learning objective, enabling effective zero-shot performance without task-specific supervision. ALIGN [111] scaled this approach to billions of noisy web-crawled pairs. BLIP [112] unified multiple tasks with captioning and retrieval into a single training framework. Flamingo [113] introduced few-shot vision-language prompting with frozen backbones and cross-attention layers for rapid adaptation.

Building upon these visual foundations, a significant shift occurred with the introduction of visual instruction tuning, which aims to align vision–language inputs with human intent through instruction-following behavior. Representative models such as MiniGPT-4 [114] and LLaVA [29] align pretrained vision encoders (e.g., CLIP) with large language models such as LLaMA [82] via lightweight projection modules and apply instruction tuning, enabling chat-style vision–language reasoning. An overview of the evolution from visual foundation models to instruction-tuned VLMs is illustrated in Figure 5.

By leveraging the VLMs's ability to jointly reason over images and text, researchers have explored new concepts for autonomous driving taks. As summarized in recent surveys [56], [57], VLMs enable interpretable and adaptable systems that support open-ended interaction, improve generalization to unseen scenarios, and facilitate multimodal reasoning. These advancements mark a shift toward more intelligent and explainable autonomous vehicles, laying the groundwork for safer and more human-aligned driving agents.

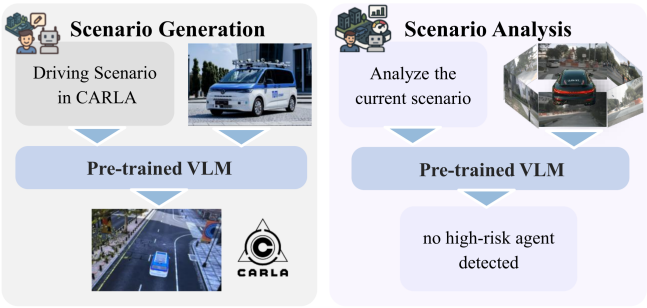

FIGURE 6. Pre-trained VLMs use both text descriptions and visual inputs for two tasks: (1) scenario generation using text prompts and scene images, and (2) scenario analysis using image understanding and textual reasoning for risk assessment.

VLMs provide three core capabilities: Multimodal understanding jointly processes images and text, such as image captioning and VQA (e.g., Flamingo, BLIP); Image–text matching involves assessing semantic alignment between an image and a caption (e.g., ALIGN, CLIP); Text-to-image generation involves synthesizing novel visuals from natural language prompts, pioneered by DALL-E [115]. Building on these foundations, VLMs can be adapted to support individual AD modules (perception, prediction, planning) and even end-to-end vision–language–action (VLA) frameworks that directly map visual and linguistic inputs to driving behaviors. In this survey, we focus specifically on VLM-based driving scenario generation and scenario analysis, as illustrated conceptually in Figure 6.

Adaptation Techniques for VLMs: Current compact VLMs use LLMs as backbones, by adding text tokenizers and vision encoders. Like LLMs, VLMs are pre-trained and then adapted for downstream tasks. Beyond the standard prompt engineering techniques of LLMs, the following adaptation strategies are commonly employed in the context of scenario generation and analysis in AD.

Modality alignment modules are additional trainable modules that transform visual inputs into formats compatible with language models. Common approaches include:
(I) *Query Transformer (Q-Former):* A transformer with learnable queries that aligns image features with the language model input space via cross-attention (e.g., BLIP-2 [116]).
(II) *Cross-attention:* Used to resample variable-length image or video tokens into a fixed-size latent representation, enabling consistent language interaction (e.g., Flamingo [113]).
(III) *Multi-Layer Perceptron (MLP) mapping:* A linear or MLPs projects vision encoder outputs to match the dimensionality required by the language model [117], [118].
(IV) *Structure-aware encoder (Prior tokenizer):* A perception-aware module that encodes structured detection outputs, such as semantic attributes, into token embeddings for downstream reasoning. For example, Reason2Drive [119] introduces a module called Prior Tokenizer to fuse region features with object-level semantics.

Fine-tuning techniques train VLMs on datasets of instruction–response pairs involving both visual and textual inputs to improve their ability to follow multimodal instructions. Two main strategies are used:
(I) *FFT:* All model weights are updated on the target dataset. It typically yields the highest task performance but incurs high computational costs and risks of overfitting. Several reviewed works adopt full fine-tuning for smaller VLMs, striking a practical balance between effectiveness and efficiency [117], [120].
(II) *PEFT:* These methods enable adaptation by updating only a small number of additional parameters. One of the most common methods are LoRA matrices, which are injected into attention or feed-forward layers, to enable efficient adaptation with minimal parameter overhead [66]. An extension of this idea is QLoRA, which further reduces memory usage by applying quantization to the base model during adapter training.

TABLE 4. Summary of Scenario-Generation Studies Using Vision Language Models.

Category	Input					Model	Technique[1]	Simulator	Output[2]	Paper
	Text	Image	View Type	Dataset	Database					
Safety-critical Scenario	✓	✓	BEV of Metadrive	Waymo Open		GPT-4o LLaVA	CoT	Metadrive		CurricuVLM [121]
Real-world Replication	✓	✓		SUMO	Road Network	GPT-4 GPT-4V	CoT RAG	SUMO		OmniTester [122]
		✓	Real FPV	CCD [123]		GPT-4o	ICL	CARLA		Miao et al. [124]
Image Dataset	✓			In-house		DALL-E2	ICL			WEDGE [125]
ADAS Testing Scenario	✓	✓	BEV of Sketch	nuScenes [126]	NHTSA	GPT-4o	CoT ICL	Metadrive BeamNG		TRACE [127]

[1] Techniques: CoT = Chain-of-Thought prompting; ICL = In-Context Learning; RAG = Retrieval-Augmented Generation.
[2] Output: Image, Trajectory, Scenario script.

B. VLM-based Scenario Generation

This subsection reviews how VLMs are used to generate driving scenarios by leveraging their understanding of visual and textual inputs. For consistency, the category definitions in this subsection follow the conceptual distinctions introduced in Section III-B. We group recent works into four categories and display them in Table 4:

Safety-critical Scenario Generation: Safety-critical scenario generation is a rapidly advancing application of VLMs in autonomous driving. It enables the synthesis of rare but relevant situations that are essential for evaluating system robustness. By combining visual perception with semantic reasoning, VLMs have the potential to identify abnormal behaviors or near-failure conditions and generate targeted, interpretable scenarios.

Recent frameworks such as CurricuVLM [121] illustrate the potential of VLMs. CurricuVLM integrates VLMs such as LLaVA into an online curriculum-learning loop. The VLM analyzes bird's-eye view (BEV) images and task descriptions to detect safety-critical events, while GPT-4o performs batch-level pattern analysis to reveal behavioral weaknesses. These insights guide a pre-trained DenseTNT model to generate tailored agent trajectories, and reinforcement learning adaptively selects the next scenarios.

However, CurricuVLM employs pre-trained VLMs, thus its performance in identifying safety-critical agents is limited. Future work could explore combining these frameworks with safety-aware, fine-tuned VLMs and incorporating temporal and multi-sensor contexts to improve reliability.

Real-World Scenario Replication: VLMs offers new opportunities for realistic driving scenario replication, by combining language understanding with visual modalities such as scenario images, enabling the creation of realistic traffic scenes based on real-world recorded dataset or maps.

OmniTester [122] proposes a framework with LLM and VLM to create realistic and diverse traffic scenarios in SUMO. User inputs and context from RAG with external knowledge and OSM map library are processed via GPT-4 [128] to generate SUMO scenario scripts. A GPT-4V analyzes the generated scenario using images and code, providing feedback in natural language. Then, the GPT-4 evaluator compares this feedback against the intended description to enhance scenario generation. Beyond the real-world map, the authors from [124] present a fully automated pipeline that transforms sample frames of dashcam crash video from the Car Crash Dataset (CCD) [123] into simulation scenarios for ADAS testing. Using GPT-4o with ICL, the system generates SCENIC scripts for CARLA, while a second GPT-4o compares real and simulated video frames based on predefined behavior features, enabling iterative refinement through visual feedback.

Current approaches using maps and recorded videos lack the use of real-world log replays, which is expected to enhance realism.

Dataset Generation: A key application of VLMs is text-to-image generation to build tailored driving datasets, particularly to improve perception systems under diverse conditions.

WEDGE [125] showcases the use of VLMs, specifically DALL-E 2, to synthesize images depicting 16 diverse and extreme weather conditions relevant to autonomous driving. Their dataset includes manually annotated 2D bounding boxes and is used to fine-tune object detectors. When evaluated on the real-world dataset, object detectors trained on WEDGE exhibit improved detection performance, highlighting the potential of VLM-generated data for enhancing perception robustness in adverse conditions.

Currently, hybrid training that combines real and synthetic data is underexplored. This approach is crucial because real-world datasets often contain very few safety-critical corner cases, whereas synthetic data enables the controlled generation of rare events, such as crashes, occlusions, and anomalies—thereby improving long-tail coverage and strengthening model robustness in high-risk scenarios.

Generation of ADAS Test Scenarios: VLMs extend ADAS scenario generation by grounding language in visual content, enabling semantically rich and visually faithful reconstructions of complex driving events. This facilitates the transformation of regulatory descriptions, test specifications,

or crash reports into executable and reproducible scenarios for simulation-based evaluation of ADAS performance.

TRACE [127] reconstructs ADAS test scenarios from unstructured multimodal crash reports, including textual summaries and visual sketches. It uses GPT-4o with ICL and CoT to extract road types and environmental details from sketches. An LLM, built on GPT and augmented with trajectory data from nuScenes [126], generates realistic vehicle paths. These components are transformed into a DSL-based scenario compatible with simulators like MetaDrive using a rule-based encoder these scenarios are further utilized to test multiple ADAS algorithms.

TRACE lack online interactive scenario editing, where users could modify scenes by sketching or annotating video frames, and VLMs could dynamically update the simulation code. This would enable human-in-the-loop control and more flexible scenario refinement.

C. VLM-based Scenario Analysis

The current progress in scenario generation with VLMs is quite at the beginning, but VLMs have already shown big promises for scenario analysis in AD. Examples include NuScenes-QA [134] for VQAs, where a VLM answers natural language questions grounded in driving scenes to support scenario analysis; NuPrompt [165] for language-guided tracking and prediction, and Refer-KITTI [166] for multi-object referring tracking tasks. However, these models are not considered foundation models because they do not utilize fully pre-trained foundation architectures. Rather, they construct task-oriented frameworks based on LLM backbone components.

In this section, we focus on foundation VLMs pre-trained on large-scale, diverse image–text datasets with cross-domain generalization. We examine their potential to improve transferability, explainability, and efficiency in analyzing complex AD scenarios. We structure our discussion around four key application areas, and show their techniques and applications in Table 5.

Visual Question Answering (VQA): VQA datasets for autonomous driving pair visual inputs with natural language queries, to evaluate scene understanding across tasks such as perception, prediction, and planning. While recent works have proposed VQA datasets, some QA remain conceptual or require human reasoning in their creation, whereas others make use of LLMs for automated generation. This section focuses on VQA-based scenario analysis methods that involve VLM execution.

Early efforts began by enriching existing scene representations with the *perception* task. Talk2BEV [129] uses a perception stack to generate BEV maps by fusing multi-view images and LiDAR, then applies BLIP-2 to augment these maps with object-level language descriptions. These descriptions are passed to GPT-4 with CoT prompting to answer spatial and semantic queries, enabling zero-shot VQA with annotated QA pairs focusing on perception and prediction. Similarly, NuScenes-MQA [117] uses GPT-4 to automatically generate diverse question templates within the Markup-QA scheme. The authors fully fine-tune a VLM that combines a CLIP-pre-trained ViT as a visual encoder, and OPT as a language model, using an MLP to align multi-camera visual features with text. This setup enables joint evaluation of caption generation and visual question answering in driving scenarios for perception.

Later works moved toward more advanced *reasoning* tasks. OmniDrive [118] introduces the first 3D VQA dataset for counterfactual reasoning in autonomous driving, evaluating VLMs with frozen EVA-02-L and Llama2–7B backbones, and using either an MLP projector (Omni-L) or a Q-Former (Omni-Q) as trainable modality bridges. Reason2Drive [119] presents a video–text VQA dataset composed of sequential images from nuScenes, Waymo, and ONCE [130], covering tasks in perception, prediction, and reasoning. The authors fine-tune a VLM consisting of FlanT5-XL and Vicuna-7B by using LoRA, leveraging a prior tokenizer and an instructed vision decoder. A Q-Former module is employed to jointly predict answers and perceptual cues.

Recent works in VQA-based scenario analysis focus on advancing multimodal reasoning and evaluation across *perception, prediction*, and *planning* tasks in autonomous driving. DriveLMM-o1 [131] introduces a step-by-step reasoning dataset based on nuScenes, incorporating both images and LiDAR points into the QA context. Their QA pairs are initially generated using GPT-4 and subsequently refined through human annotation. The authors fine-tune InternVL2.5-8B using LoRA, demonstrating improved performance on reasoning and final answer accuracy across perception, prediction, and planning. AutoDrive-QA [135] converts open-ended QA pairs from DriveLM [132], LingoQA [133], and NuScenes-QA [134] into multiple-choice questions using GPT-4o, adding distractors, which are plausible but incorrect answer choices designed to reflect realistic domain-specific errors, to simulate realistic errors. This forms a standardized benchmark to evaluate pre-trained VLMs across key scenario analysis tasks across perception, prediction, and planning.

Despite these advances, most of the current VQAs overlook traffic rules and real-world driving conventions. Future work should incorporate traffic rule-aware QA, grounded in road semantics (e.g., right-of-way rules and road signal compliance), to enable more realistic and safety-relevant scenario reasoning.

Scene Understanding: VLMs are heavily used to interpret complex driving scenarios. Recent works have leveraged VLMs for *scene tagging*, which represents the most basic level of scene understanding, involving binary or categorical assignments. Scene tagging assigns predefined labels at either the scene level (e.g., to analyze the weather conditions), or at pixel level (semantic segmentation) to characterize visual content for downstream tasks. Najibi et al. [136] leverage a pre-trained CLIP to perform zero-shot scene

TABLE 5. Summary of Scenario-Analysis Studies Using Vision Language Models.

Category	Input			Model			Technique[2]	Focus	Paper
	Context	Image[1]	Dataset	VLM	LLM	Role			
Visual Question Answering (VQA)		Multi-view	nuScenes	BLIP2 InstructBLIP2 MiniGPT4	GPT-4	VLM: BEV Feature Extraction LLM: QA Execution	Zero-shot	Perception Prediction	Talk2BEV [129]
	✓	Multi-view	nuScenes	ViT+OPT	GPT-4	VLM: VQA Execution LLM: QA Generation	MLP Fft	Perception	NuScenes-MQA [117]
	✓	Multi-view	nuScenes	EVA-02-L + Llama2-7B	GPT-4	VLM: VQA Execution LLM: QA Augumentation	MLP Q-Former Fft	Counterfactual Reasoning	OmniDrive [118]
	✓	FPV	nuScenes Waymo Open Once [130]	FlanT5-XL + Vicuna-7B	GPT-4	VLM: VQA Execution LLM: QA Augumentation	Q-Former Tokenizer LoRA	Perception Prediction Reasoning	Reason2Drive [119]
	✓	Multi-view	nuScenes	InterVL2.5-8B	GPT-4o	VLM: VQA Execution LLM: QA Generation	LoRA	Perception Prediction Planning	DriveLMM-o1 [131]
	✓	FPV	DriveLM [132] LingoQA [133] NuScenes-QA [134]	Qwen2-VL-7B Qwen2-VL-72B GPT-4o	GPT-4o	VLM: VQA Execution LLM: Multi-Choice QA	Zero-shot	Perception Prediction Planning	AutoDrive-QA [135]
Scene Understanding	✓	Multi-view	Waymo Open	CLIP			Zero-shot	Tagging	Najibi et al. [136]
	✓	Multi-view	SemanticKITTI [137]	Grounding DINO + SAM	GPT-3.5	VLM: Object Grounding LLM: Narrative Generation	Zero-shot	Tagging	OpenAnnotate3D [138]
	✓	FPV	Cityscapes [139] CamVid [140] CARLA	ImageGPT			Zero-shot	Tagging	Kou et al. [141]
	✓	Multi-view	DriveLM	ViT-B/32 + T5-Base/Large			MLP Fft/LoRA	Tagging	EM-VLM4AD [120]
	✓	FPV	FARS	GPT-4V LLaVA-13B	Llama2-13B Zephyr-7b-α	VLM: Scene Captioning LLM: Risk Assessment	Zero-shot	Captioning	Zarzà et al. [142]
	✓	Roadcamera	RDD [143]	GPT4			Zero-shot	Captioning	ConnectGPT [144]
	✓	BEV	WOMD [47]	GPT-4V			Zero-shot	Captioning	Zheng et al. [145]
	✓	FPV	BDD100K [146]	Multiple VLMs			Zero-shot	Tagging Reasoning	Rivera et al. [147]
	✓	FPV	nuScenes BDD-X [148] CDD [123]	GPT-4V			Zero-shot	Reasoning	Wen et al. [149]
	✓	Multi-view	MAPLM-QA [150]	ViLA			Zero-shot	Reasoning	Keskar et al. [151]
Benchmark & Dataset	✓	FPV	DriveLM	Multiple VLMs	GPT-4o	VLM: VQA Execution LLM: Answer Evaluation	Zero-shot	Robustness	DriveBench [152]
	✓	Multi-view	nuScenes	ViT/V2-99 + LLaVA-1.5-7B			Tokenizer MLP LoRA	3D Grounding	NuGrounding [153]
	✓	FPV	nuScenes CARLA	BLIP2			LoRA	GVQA	DriveLM [132]
	✓	FPV	CODA [154]	LLaVA-llama-3-8B GPT-4o	GPT-4	VLM: Scene Captioning LLM: Caption Evaluation	LoRA	Corner Cases	CODA-LM [155]
	✓	FPV	In-house	GPT-4o			CP, ICL, CoT	ADAS-LKA	OpenLKA [156]
Risk Assessment	✓	Multi-view	In-house	GPT-4V			CP CoT	Risk Scoring	Hwang et.al [157]
	✓	FPV	DAD [158]	Flamingo	GPT-3.5		Zero-shot	Hazard Explanation	Latte [159]
	✓	FPV	CARLA	DINOv2 OWLV2+SAM2 GPT-4o			CP	Anomaly Detection	Ronecker et.al [160]
	✓	Multi-view	CARLA	InternViT	Interlm2-chat	VLM: Video Extraction LLM: Narrative Generation	MLP QLoRA CP, CoT	Situation Awareness Reasoning	Think-Driver [161]
	✓	Partially occluded BEV	CARLA	Llama3.2-11B LLaVA-1.6-7B Qwen2-VL-7B			LoRA CP	Uncertainty Scoring	Lee et.al [162]
	✓	FPV	BDD100K	Qwen2-VL7B			LoRA	Hazard Detection	INSIGHT [163]
	✓	FPV	OpenLKA [156]	Qwen2.5-VL-3B Qwen2.5-VL-7B			LoRA	LKA Failures Prediction	LKAlert [164]

[1] Image: FPV = First Person View; BEV = Bird Eye View.
[2] Techniques: Fft = Full fine-tuning; CP = Contextual Prompting; ICL = In-Context Learning; CoT = Chain-of-Thought prompting; Tokenizer = Prior Tokenizer; MLP = Multi-Layer Perceptron mapping.

tagging on camera images, assigning semantic labels that are projected onto LiDAR points. These labels guide the generation of 3D pseudo-labels, which are then used to train a 3D object detector without human annotations. OpenAnnotate3D [138] introduces an auto-labeling system for multi-modal 3D data, using GPT-3.5 for interpreting natural language scene descriptions and a VLM with Grounding DINO and SAM for generating dense 2D masks, which are fused spatio-temporally and projected into 3D annotations. Kou et al. [141] propose a framework to enhance VLMs for street scene semantic understanding. They use a pre-trained ImageGPT to extract semantic features from First Person View (FPV) images, and train a lightweight perception head that maps the semantic features to pixel-wise semantic segmentation masks. EM-VLM4AD [120] proposes a lightweight VLM trained on the dataset from DriveLM [132] with a primary focus on scenario tagging. It uses a ViT image encoder and explores two adaptation strategies: full fine-tuning of T5-base and LoRA-based tuning of T5-large. The model is benchmarked against baselines in terms of parameter count, Floating Point Operations Per Second (FLOPs), and memory usage, showcasing strong efficiency for deployment in resource-constrained settings.

Building on scene tagging, recent efforts have advanced toward the intermediate-level task of *scene captioning*, which bridges perception and language by generating open-form descriptions. Scene captioning generates concise natural language descriptions of visible elements. Zarzà et al. [142] propose a framework using structured inputs with principal component analysis, and adopt Llama2-13B with CoT and CP to assess the risks in a scenario, suggesting driving adaptations. They test their framework with the FARS dataset[8]. Additionally, they leverage a VLM, specifically LLaVA-13B with CP, to perform image-based scenario captioning, enhancing scene understanding through natural language descriptions. ConnectGPT [144] leverages VLMs to generate standardized Cooperative Intelligent Transport Systems (C-ITS) messages for Connected and Automated Vehicles. GPT-4 is used to interpret infrastructure camera images, generate C-ITS messages, with validation conducted on a small curated set of highway images, including samples originating from the Road Damage Dataset (RDD) [143].

Zheng et al. [145] introduce a context-aware motion prediction framework using VLMs. They employ GPT-4V to extract traffic context from a transportation context map. They combine vector map data and historical trajectories, and feed the generated scenario description into a motion transformer to improve trajectory prediction.

Several studies address the most advanced form of scene understanding: *scene reasoning*, which requires interpreting interactions, causality, and abstract situational context. Scene reasoning interprets relationships and interactions among agents while producing coherent narratives that capture intent, causality, and situational context. Rivera et al. [147] propose a scalable pipeline for traffic scene classification using off-the-shelf VLMs such as GPT-4V, LLaVA, and CogAgent-VQA [167]. These models are evaluated zero-shot to reason about predefined scenario elements, such as lane markings and vehicle maneuvers, using self-developed and the BDD100K [146] datasets. Wen et al. [149] explore GPT-4V's zero-shot capability for road scene interpretation from dashcam footages, evaluating the model on object detection, scene captioning, VQA, and causal reasoning, while highlighting its potential and limitations for autonomous driving. Keskar et al. [151] evaluate NVIDIA's ViLA on the MAPLM-QA [150] benchmark for traffic scene understanding. Using contextual prompting, they assess ViLA on multiple-choice VQA tasks, including lane counting, intersection detection, scene classification, and point cloud quality assessment. ViLA shows strong performance on high-level VQA tasks but struggles with fine-grained spatial reasoning.

Benchmarks & Datasets: To support the development and evaluation of VLMs in autonomous driving, recent efforts have introduced specialized *benchmarks* and curated *datasets* covering key tasks such as perception, prediction, planning, and scenario reasoning under real-world and safety-critical conditions.

Aiming for a standardized evaluation, several works present benchmarks aligned with diverse driving scenarios. DriveBench [152] introduces a benchmark for evaluating scenario reasoning across multiple driving tasks. It extends the VQA dataset from DriveLM [132] and adds diverse visual corruption categories to assess the model's robustness. Using this benchmark, the authors evaluate the robustness of a range of pre-trained and fine-tuned VLMs (e.g., GPT-4o, Qwen2-VL [168]) under clean, corrupted, and text-only conditions. GPT-4o is further employed as an automatic evaluator for open-ended answers. nuGrounding [153] proposes the first 3D visual grounding benchmark with human-annotated object grounding based on nuScenes. The authors fine-tune LLaVA-1.5 using LoRA, with ViT or V2-99 as the visual encoder. To incorporate 3D understanding, they extract BEV features via a BEV-based detector, map them into the LLM adapter, and fuse them with VLM outputs through a query fuser for accurate object detection and localization.

Complementing these benchmarks, other works provide high-quality datasets to train and adapt VLMs to complex driving environments. DriveLM [132] introduces a graph-structured visual question answering (GVQA) which leverages graph-based scene representations to answer structured perception, prediction, and planning questions in autonomous driving scenarios, using human-curated QA graphs from nuScenes and rule-based annotations from CARLA. A BLIP-2-based VLM is fine-tuned with LoRA and guided by graph-based question prompting to enable zero-shot interpretable scenario reasoning across perception, prediction, and planning. CODA-LM [155] introduces a corner-case image-text dataset derived from the CODA

[8]https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars

dataset [154]. The authors use GPT-4V to generate multi-task captions spanning perception, prediction, and planning for each image. These captions are then evaluated and refined using GPT-4. After constructing the dataset, they fine-tune a LLaVA-llama-3-8B model to enhance vision language understanding in corner-case driving scenarios. OpenLKA [156] introduces a large-scale, real-world dataset for Lane Keeping Assist under diverse driving conditions. GPT-4o is used in conjunction with CP, CoT, and ICL to generate structured scene annotations that describe lane quality, weather, and traffic context.

However, the existing benchmarks and datasets still lack realism and diversity. For example, DriveBench exposes the VLM's vulnerability to corruption, suggesting the need for more realistic disturbances (e.g., occlusions, night-time). CODA-LM relies on filtered GPT captions, underscoring the gap in real-world edge-case coverage.

Risk Assessment: VLMs are increasingly applied to autonomous driving risk assessment, addressing tasks like hazard detection, uncertainty estimation, and failure prediction. Recent approaches leverage both prompting and fine-tuning and use diverse visual inputs, including BEV maps, multi-view images, and segmentation masks. These methods aim to improve safety through interpretable reasoning and context-aware decision support.

Recent advances have explored prompting techniques for risk analysis. Hwang et al. [157] utilize GPT-4V in a zero-shot setting for risk scoring in street-crossing scenarios. The model receives structured visual inputs, including bounding boxes, segmentation masks, and optical flow, alongside contextual prompts formulated using CoT. Instead of directly processing raw images, GPT-4V reasons over augmented visual features to assess safety levels and provide natural language justifications. Similarly, LATTE [159] introduces a real-time hazard detection framework that utilizes off-the-shelf computer vision modules and three lightweight attention modules for spatial reasoning, temporal modeling, and risk prediction. Upon hazard detection, Flamingo and GPT-3.5 are triggered to generate scene captions and verbal explanations. The system operates in a zero-shot manner by leveraging contextual prompting for situational reasoning. For anomaly object detection, Ronecker et al. [160] proposed both patch-based and instance-based embedding methods using vision foundation models, evaluated on a CARLA-based dataset. They leverage the zero-shot capabilities of DINOv2 for visual embeddings and combine OWLv2 with SAM2 for object-level instance segmentation. Their instance-based approach achieves slightly better results than GPT-4o using contextual prompting.

Think-Driver [161] proposes a VLM that uses multi-view images to assess perceived traffic conditions and evaluate the risks of current driving maneuvers. It employs multi-view RGB inputs and ego state data, processed by InternViT and InterLM2-chat, respectively. The model is fine-tuned using Quantized Low-Rank Adaptation (QLoRA) and trained on CoT-style QA data that cover scene understanding, hazard reasoning, and action prediction. In consideration of occlusion-aware BEV representations, Lee et al. [162] first investigate the use of VLM for uncertainty prediction in autonomous driving. They construct a dataset from CARLA using BEV images that contain occlusion masks, paired with driving actions and uncertainty scores. Three VLMs are fine-tuned using LoRA to compare their performance under occluded conditions. For hazard detection and explanation, INSIGHT [163] fine-tunes Qwen2-VL-7B via LoRA. Using annotated hazard locations in BDD100K images, the model is trained to localize high-risk regions and generate natural language descriptions. It outperforms several pre-trained VLMs in both spatial localization and interpretability tasks. Finally, LKAlert [164] develops a VLM-based framework for predicting lane-keeping assist failures. It integrates RGB dashcam images, CAN bus signals, and lane segmentation masks from LaneNet. A Qwen2.5-VL model is fine-tuned via LoRA, with lane masks serving as spatial guidance. The model outputs binary alerts and interpretable explanations to enhance safety transparency.

To enable real-world deployment, the inference latency and resource demands need to be further reduced through model compression, efficient prompting, and lightweight VLM architectures optimized for onboard execution in autonomous vehicles.

D. Limitations and Future Directions

VLM-based Scenario Generation: Compared to LLM-based scenario generation (Section III B), VLMs remain underexplored in areas such as scenario synthesis for training driving policies and closed-loop scenario generation. With their ability to process both visual and textual inputs, VLMs offer a powerful extension to existing frameworks. A promising direction is to use them as auxiliary analysis modules to improve the interpretability and fidelity of the generated scenarios, while also providing feedback signals to iteratively enhance the scenario quality.

Moreover, there is strong potential to develop more sophisticated and interdisciplinary pipelines that fully leverage the multimodal reasoning capabilities of VLMs. For instance, in scenario-based testing, real-world traffic videos could be interpreted by VLMs to produce detailed scene captions. These captions could serve as structured conditions for DMs to regenerate photorealistic driving scenes or videos. Such a multi-stage pipeline, linking perception, semantic understanding, and simulation, represents a promising direction for building holistic and scalable scenario generation systems.

VLM-based Scenario Analysis: In the domain of scenario analysis, VLMs show advantages over text-only LLM-based frameworks. Current research follows two main trends.

The first trend centers on developing task-specific frameworks, often augmented with external computer vision modules (e.g., for 3D grounding or hazard detection).

Meanwhile, the rapid progress of general-purpose pre-trained VLMs raises a key research question: to what extent can these models handle scenario analysis effectively without relying on external tools like object detectors, depth estimators, or 3D grounders? Investigating the capabilities and limitations of such end-to-end VLMs could enable more streamlined, scalable solutions that reduce the system's complexity while preserving, or even enhancing, their analytical performance.

The other trend emphasizes VQA, designing tailored VQA tasks that fine-tune VLMs for improved task-oriented performance. Despite recent advances, several challenges persist. While large-scale pre-trained VLMs exhibit strong potential, the scenario analysis pipeline in autonomous driving remains highly complex and poorly standardized. Specifically, there is a lack of benchmark datasets, consistent annotation frameworks for VQA tasks, and unified evaluation metrics tailored to scenario analysis. Addressing these gaps is essential for developing more robust and task-specific VLMs capable of handling real-world autonomous driving scenarios.

V. Multimodal Large Language Models (MLLMs)

This section begins with the development of MLLMs, highlighting their architectural evolution and adaptation techniques, such as modality bridging and instruction tuning. Then, it covers scenario generation from multimodal input and scenario analysis tasks, including VQA, scene understanding, and risk assessment in AD contexts.

A. Development of MLLMs

MLLMs extend pre-trained LLMs by integrating three or more modalities, such as vision, audio, and video, enabling the system to reason over richer and more diverse sensory inputs beyond image–text pairs. Early VLMs such as BLIP-2 [116] use frozen vision backbones connected to LLMs via adapters such as Q-Former. These models primarily extend VLMs by connecting frozen perception encoders to LLMs, but do not yet constitute full MLLMs. In parallel, early multimodal extensions such as Video-LLaMA [169] incorporated additional modalities, including video and audio, enabling joint reasoning over text, visual frames, and acoustic signals. Although these models marked an initial step toward MLLMs, they typically relied on frozen backbones and lacked unified multimodal training, resulting in limited temporal and cross-modal reasoning capabilities.

More recent models, including GPT-4o, represent a further step toward fully unified MLLMs by integrating vision and audio within a single model. Similarly, Google Gemini [170] and Qwen-Omni [171] natively support multiple modalities, enabling open-ended reasoning over images, videos, audio, and structured visual content, such as charts and diagrams. While effective for general visual-language tasks, these models fall short in autonomous driving, which requires reasoning over structured inputs like temporal object tracks, BEV layouts, and interaction-aware motion patterns. The progression from LLMs to MLLMs is depicted in Figure 5.

To address the unique demands of autonomous driving, recent MLLM architectures have begun incorporating structured, domain-specific modalities such as multi-view video, LiDAR point clouds, and BEV layouts. These additions enable spatial and temporal grounding, allowing LLMs to reason more effectively over complex driving scenes and multi-agent dynamics [172].

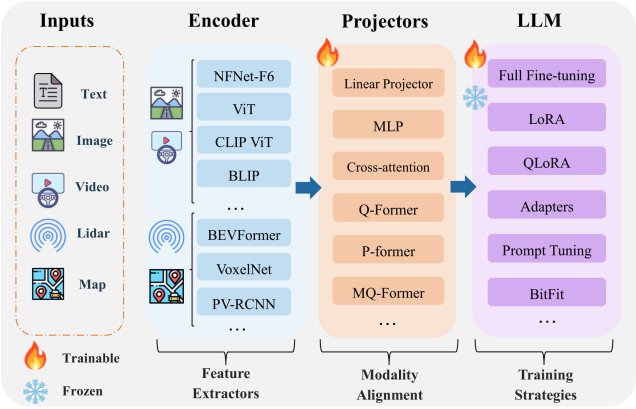

FIGURE 7. Overview of adaptation techniques for MLLMs in autonomous driving. Encoders extract features from modality-specific inputs. Projectors are trainable modules that map features into the LLM's embedding space to enable cross-modal alignment. The LLM serves as the reasoning core and can be frozen or trainable, depending on the available resources and the task, using fine-tuning techniques.

While building on techniques from VLMs, MLLMs are adapted to support a broader range of modalities essential for autonomous driving, such as video, LiDAR point clouds, BEV maps, and High-Definition (HD) semantic features. As illustrated in Figure 7, these systems typically consist of specialized modality encoders (e.g., BEVFormer [173], CLIP-ViT [174], VoxelNet [175]), projection modules to align multi-modal features (e.g., MLPs, Q-Former, cross-attention), and task-specific training strategies. MLLMs often keep both the perception and language backbones frozen, with adaptation focusing on lightweight bridging and instruction tuning for downstream driving-related tasks. The main adaptation strategies are discussed in the following.

Modality alignment modules: These modules serve as a bridge between non-text modalities and the LLM's token space. The main modality alignment modules are:

(1) Linear projector: a single linear (i.e., fully connected) neural layer used to project modality-specific features into the LLM's embedding space. It offers a lightweight mapping strategy and is often used in early-stage VLMs or in combination with pre-trained encoders [176], [177].

(2) MLP projection: Projects high-dimensional features from vision or spatial encoders (e.g., ViT, BEVFormer) into the LLM's token space. Used in models such as BLIP-2 [116] and driving-centric adapters like P-Adapter [178], which align BEV or LiDAR features for language-based reasoning.

(3) Spatio-Temporal (ST)-Adapter: A lightweight temporal adapter module is used to extend image-based MLLMs

to process sequential video inputs [173], [178]. It enables spatiotemporal modeling without modifying the core LLM weights.

(4) Cross-attention: Uses learnable queries to attend over image or point cloud tokens, enabling multimodal fusion for tasks such as spatial/temporal grounding, semantic alignment, and instruction following [179], [180].

(5) Q-Former: Transformer-based query modules that distill task-relevant embeddings from multi-modal inputs using cross-attention. These modules are applied in BLIP-2 [116], InternDrive [181], and NuInstruct [173] for structured fusion across video, LiDAR, and BEV inputs.

(6) Fusion transformers: Specialized attention blocks designed to integrate features across multiple streams such as BEV maps, multi-view video, or LiDAR point clouds. Modules like BEV-Injection [173] serve as fusion transformers by aligning and injecting multi-modal features (e.g., from images or LiDAR) into a unified BEV representation. These are commonly used in driving-centric MLLMs.

(7) Structure-aware encoder: A module that converts structured perception inputs, such as 3D bounding boxes [175], scene graphs, or motion trajectories, into token embeddings suitable for language-based reasoning.

Multimodal fine-tuning: Once modality alignment is achieved, an MLLM can be trained to follow task-specific prompts using paired instruction data like VQA. This stage teaches the model to reason over multimodal contexts and produce grounded outputs. Similarly to VLMs, two main strategies are commonly employed to achieve this adaptation:

(1) PEFT: PEFT strategies adapt MLLMs by updating only a small subset of the model's parameters, typically keeping the LLM frozen. While classical PEFT methods such as adapter layers, LoRA, and prompt tuning (discussed below) operate inside the LLM, recent works in autonomous driving often apply PEFT to modality alignment modules [173]. For example, components like ST-Adapters and Q-Formers are trained to bridge visual or spatial inputs to the LLM, enabling task adaptation without modifying the core language model.

Adapter layers: Lightweight trainable modules inserted between the layers of an LLM, typically using a down-projection and up-projection structure. They are used in LLaMA Adapter V2 [182] and InternDrive [181].

LoRA: Applies low-rank updates to attention and feed-forward modules. Frequently used in driving models like DriveGPT4 [176].

PEFT-Modality Alignment (MA): Fine-tuning only the MA modules (e.g., Q-Former, ST-Adapter), while keeping the LLM's weights frozen.

(2) FFT: Full fine-tuning updates all model parameters, including vision encoders, spatial encoders, and the LLM. While this approach typically yields the highest task-specific performance, it is computationally intensive. To reduce the computational cost, some works apply FFT to smaller models, for example using Qwen2-0.5B [183].

B. MLLM-based Scenario Generation

MLLMs can jointly process diverse visual inputs from vehicle sensors or human sources, enabling a comprehensive understanding of complex driving environments. Their ability to integrate multiple modalities also supports the generation of more realistic, context-aware scenarios. In the following, we categorize the use of MLLMs into two groups, summarized in Table 6. For consistency, the scenario categories follow the definitions introduced in Section III B.

Safety-critical Scenario Generation: MLLM-based safety-critical scenario generation focuses on synthesizing rare and high-risk driving situations by leveraging heterogeneous modalities such as videos, GPS traces, and crash reports. By reasoning jointly over spatial, temporal, and semantic cues, these methods reconstruct or generate corner cases that closely reflect real-world hazardous events.

AutoScenario [184] presents a pipeline to generate realistic corner cases using multimodal crash data from NHTSA, including text, images, videos, and semi-structured reports. They use GPT-4o with CoT to generate structured scenario descriptions, which are then used to produce road networks in SUMO and agent behaviors in CARLA. Their scenario refinement is guided by GPS traces and frame-level similarity between simulated and real scenes, to ensure good matching with the original crash event. A promising future direction is to incorporate spatial modalities, such as LiDAR point clouds or BEV maps, to achieve more accurate scene geometries and agents' localization, enhancing realism beyond what 2D video and depth sensing alone can offer.

ADAS Testing Scenario Generation: Scenario generation for ADAS testing with MLLMs aims to derive executable and reproducible test cases from real-world multimodal observations, primarily targeting function-level validation of ADAS stack under typical but diverse driving conditions. LEADE [186] generates ADAS test scenarios from real-traffic videos in the HDD dataset [185]. Key frames are used in multimodal ICL and CoT prompting with GPT-4V to create abstract scenarios, which are converted into executable programs for the LGSVL simulator [98]. The Apollo ADAS stack [94] runs an ego vehicle, and a dual-layer search identifies semantic-equivalent scenarios that expose behavioral differences between Apollo and human drivers. Future work could align scenario generation with ADAS test standards, enabling the synthesis of regulation-compliant scenarios. Incorporating traffic rules and structured priors would also improve controllability and test coverage.

C. MLLM-based Scenario Analysis

This section discusses the papers using MLLMs for scenario analysis in AD. We categorize the existing works into three key tasks, as reported in Table 7.

Visual Question Answering (VQA): In comparison with VQA-based scenario analysis with VLMs, MLLMs have extended capabilities to deal with multi-modal sensor data such as videos, LiDAR point clouds, and HD maps [199],

TABLE 6. Summary of Scenario-Generation Studies Using Multimodal Large Language Models.

Category	Input					Model	Technique	Simulator	Output[1]	Paper
	Text	Image	Video	Dataset	Database					
Safety-critical Scenario	✓	✓	Real FPV	SUMO CARLA OSM	NHTSA GPS	GPT-4o	CoT	CARLA		AutoScenario [184]
ADAS Testing Scenario	✓		Real FPV	HDD [185]		GPT-4V	CoT ICL	LGSVL		LEADE [186]

[1] Output icons: Image, Trajectory, Scenario script.

besides images and text. Based on their task and data modality, existing VQA datasets can be grouped into four categories:

(I) *General AD Tasks – Perception, Reasoning, and Control:* Several datasets target core autonomous driving tasks, including visual perception, reasoning, and decision-making. DriveGPT4 [176] introduces the first driving-specific, video QA-style instruction-following dataset, generated using GPT-4 with structured inputs including object detection bounding boxes, captions, and control signals formatted as text. It fine-tunes a MLLM combining CLIP, and LLaMA2 with LoRA adapters to produce both textual explanations and control outputs. Meanwhile, a mix-finetuning strategy merges general visual instruction data with driving-specific samples to improve reasoning and performance. VLAAD [177] introduces a multi-modal assistant for autonomous driving, trained on an instruction-following dataset derived from BDD-X and HDD videos, with QA pairs augmented using GPT-4. The model is built on Video-LLaMA, which combines a BLIP-2-based visual encoder, a Video Q-Former for temporal modeling, and a frozen LLaMA-2-7B language model. PEFT-MA is applied only to the Q-Former and projection layers, enabling the model to efficiently perform tasks such as VQA, free-form QA, ego-intention prediction, and scenario-level reasoning. LingoQA [133] presents a VQA dataset for autonomous driving, covering perception, reasoning and action. It includes an action set annotated with GPT-3.5 and a scenery set generated by GPT-4 using CoT. The baseline model processes video frames using CLIP and a Q-Former, with a linear projector to align features to Vicuna-1.5-7B's token space. For the fine-tuning, PEFT-MA is applied to the Q-Former and projector and the LLM remains frozen. Evaluation is conducted using the novel Lingo-Judge classifier, which is trained with LoRA.

(II) *Spatio-Temporal Reasoning:* Datasets in this group emphasize reasoning over agent motion, temporal dependencies, and event semantics in driving scenarios. TUMTraffic-VideoQA [183] introduces a multiple-choice video QA dataset for roadside traffic scenes, covering object captioning and spatio-temporal grounding, and facilitating fine-grained spatio-temporal reasoning in traffic scenarios. Visual metadata is extracted using standard detectors and captioned by off-the-shelf VLMs, while GPT-4o-mini generates QA pairs via template-augmented prompting. The baseline model (TUMTraffic-Qwen) uses SigLIP for visual encoding, an MLP projector for modality alignment, and Qwen2 (0.5B/7B) as the LLM, which is fully fine-tuned for instruction-following QA. NuPlanQA [172] introduces a video QA dataset built on nuPlan, using GPT-4o to generate free-form QA pairs for training and multiple-choice QA for evaluation. To leverage this data, the authors propose BEV-LLM, an MLLM that integrates multi-view images and BEV features through a BEV encoder, a BEV-Fusion module, and an MLP projector. The model uses LLaMA-3.2-Vision as a frozen backbone, while training is applied only to the BEV-Fusion module and projection layers, following a PEFT-MA strategy.

(III) *Risk-Aware Reasoning:* To address safety-critical understanding, several datasets focus on risk recognition, intention estimation, and planning-related queries. NuInstruct [173] introduces multi-view video QA datasets covering perception, prediction, risk, and planning tasks. QAs are generated via a structured SQL pipeline. The authors propose BEV-InMLLM, which extends MLLMs (e.g., Video-LLaMA) by utilizing ST-Adapters and a BEV-Injection module or a Fusion transformer that integrates spatial features from multi-view videos, resulting in improved performance on holistic autonomous driving tasks. HiLM-D [178] introduces DRAMA-ROLISP, a risk-aware VQA dataset for risk assessment that is enhanced using GPT-4. The model fine-tunes MiniGPT-4 with a ViT and a ST-Adapter for video input, a ResNet-based encoder, and a P-Adapter for spatial fusion. A Query-Aware Detector integrates outputs for risk object localization and intention reasoning. The LLM itself remains frozen, with only the adapters, fusion, and projector layers fine-tuned. DVBench [189] introduces a comprehensive video-based VQA benchmark for safety-critical autonomous driving, built on SHRP2 [188] dashcam data. Multiple-choice QA pairs are generated and refined using GPT-4o and Qwen2.5-72B, covering perception and reasoning tasks, and classifying into 11 subcategories. The benchmark evaluates 14 MLLMs using the self proposed metric, which rotates answer positions to assess robustness. The authors also compare the performance of Qwen2-VL-2B/7B with and without full fine-tuning on the DVBench dataset.

(IV) *Multi-Modal Extensions with LiDAR (with/without HD Maps:* To extend reasoning beyond RGB data, some datasets incorporate 3D point clouds or HD maps. LiDAR-LLM [175]

TABLE 7. Summary of Scenario-Analysis Studies Using Multimodal Large Language Models.

Category	Input						Model			Technique[3]	Focus[4]	Paper
	Image	Context	Lidar	Video[1]	Map	Dataset	MLLM	LLM	Role[2]			
Visual Question Answering (VQA)		✓		FPV		BDD-X	CLIP +Llama2	GPT-4	MLLM: VideoQA Exec. LLM: QA Gen.	Projector LoRA	Perception Reasoning Control	DriveGPT4 [176]
	✓	✓		FPV		BDD-X HDD	BLIP2 + Llama2-7B	GPT-4	MLLM: VideoQA Exec. LLM: QA Aug.	QueryTrans Projector PEFT-MA	Prediction Reasoning	VLAAD [177]
	✓	✓		FPV	✓	In-house	CLIP + Vicuna1.5-7B	GPT-4	MLLM: VideoQA Exec. LLM: QA Gen.	QueryTrans Projector	Prediction Reasoning Control	LingoQA [133]
	✓	✓		Roads		TUMTraffic VideoQA	SigLIP + Qwen2-0.5B/7B	GPT-4omini	MLLM: VideoQA Exec. LLM: QA Gen.	MLP PEFT-MA Fft	ST Reasoning	TUMTraffic VideoQA [183]
	✓	✓		Multi View	✓	NuPlan	BEV Encoder Llama3.2V-11B	GPT-4o	MLLM: VideoQA Exec. LLM: MC-QA Gen.	MLP FusionTrans PEFT-MA	Perception ST Reasoning	NuPlanQA [172]
		✓		Multi View		nuScenes	Video-Llama		MLLM: VideoQA Exec.	Cross-attention ST-Adapter QueryTrans FusionTrans PEFT-MA	Perception Prediction Reasoning Risk	NuInstruct [173]
	✓	✓		Multi View		DRAMA [187]	ViT + MiniGPT-4	GPT-4o	MLLM: VideoQA Exec. LLM: VQA Aug.	ST-Adapter MLP Cross-Attention PEFT-MA	Perception Prediction Reasoning Risk	HiLM-D [178]
	✓	✓		FPV		SHRP2 [188]	LLaMA-VID-7B 14 MLLMs	GPT-o1 Qwen2.5-72B	MLLM: VideoQA Exec. LLM: MC-QA Gen.	Fft ICL	Perception Reasoning Risk	DVBench [189]
		✓	✓			nuScenes	Voxel + Llama2-7B		MLLM: VQA Exec.	Encoder QueryTrans MLP PEFT-Adapter	Grouding Captioning	Lidar-llm [175]
	✓	✓	✓		✓	In-house	CLIP + Llama2-7B		MLLM: VQA Exec.	Projector LoRA	Perception	MAPLM [150]
	✓	✓	✓	Roads FPV		V2X-Real V2V4Real	PointPilars + LLaVA-v1.5-7b		MLLM: VQA Exec.	Projector LoRA	Perception Planning	V2V-LLM [190]
Scene/Scenario Understanding	✓	✓				nuScenes	InternVl-1.5	GPT-4o	MLLM: Scene Under. LLM: QA Gen.	LoRA	Scene perception prediction Reasoning	InterDrive [181]
	✓	✓	✓			KITTI [191] nuScenes	Video-LLaVA GPT-4o			CoT	Scene Reasoning	Jain et al. [192]
		✓		FPV		BDD-X	VideoMA +Ada-002 +OpenFlamingo			Cross-Attention LoRA	Scenario Reasoning	Dolphins [179]
	✓	✓	✓	Multi View		DriveLM	ViT-L/14 Llama-Adapter V2			QueryTrans PEFT-Adapter PEFT-MA	Scenario Reasoning	Ishaq et al. [174]
		✓		FPV		BDD100K	ViT-L/14 +Vicuna-7B	GPT-3.5	MLLM: Video Capt. LLM: Caption Eval.	Projector QueryTrans PEFT-MA	Scenario Captioning	WTS [193]
		✓		FPV		LingoQA	LLaVA-VL-7B Qwen-VL-7B	Qwen2.5-1.5B Qwen2.5-7B	MLLM: Scene Extract. LLM: Scene Under.	Zero-shot	Scenario Captioning	V3LMA [194]
Risk Assessment	✓	✓			✓	DeepAccident [195]	GPT-4V	GPT-4	MLLM: Image Extract. LLM: Narrative Gen.	Zero-shot	Accident Prevention	AccidentGPT [196]
	✓	✓				DRAMA -ROLISP, DRAMA -SRIS [178]	ResNet-101 + Swin-L +Llama2-7B			QueryTrans Cross-Attention MLP PEFT-Adapter	Safety Interaction	MLLM-SUL [180]
	✓	✓		FPV		nuScenes	VideoLlama2	Llama3.1-8B	MLLM: Video Extract. LLM: Narrative Gen.	Zero-shot	Safety Interaction	ScVLM [197]
	✓	✓		FPV		DRAMA	Gemini1.5V-Pro			ICL	Risk event Detection	Abu et al. [198]

[1] Video: FPV = First Person View; BEV = Bird Eye View;
[2] Role: Exec.= Execution; Aug. = Augmentation; Gen. = Generation; Under. = Understanding; Capt. = Captioning; Eval. = Evaluation; Extract. = Extraction;
[3] Techniques: Only focus on the techniques for MLLMs. Projector = Linear projector; MLP = MLP projection; QueryTrans = Query Transformer; FusionTrans = Fusion Transformer; PEFT-Adapter = Adapter layers; Fft = Full fine-tuning; PEFT-MA: Only trains modality alignment modules and LLMs are frozen; Encoder = Structure-aware encoder; CP = Contextual Prompting; ICL = In-Context Learning; CoT = Chain-of-Thought prompting;
[4] Focus: ST Reasoning = Spatio-temporal reasoning.

first tackles 3D captioning, grounding, and VQA from LiDAR point clouds. It extracts BEV features via a voxel encoder, embeds them using a View-Aware Transformer with learnable queries, which acts as a prior tokenizer, and projects them into the language space through an MLP. Adapter layers are fine-tuned within the LLM to support 3D scene understanding. MAPLM [150] introduces a large-scale multimodal benchmark and VQA dataset and focuses on perception and HD map understanding in autonomous driving. It includes panoramic 2D images, BEV projections from LiDAR point clouds, and text descriptions extracted from HD maps. The baseline model aligns visual features using pre-trained CLIP encoders and lightweight projection adapters, mapping them into the LLM's embedding space. Instruction tuning is performed via LoRA on Vicuna or LLaMA-2, enabling the model to perform effective scene-level reasoning across modalities. V2V-LLM [190] further extends LiDAR-based multimodal reasoning to cooperative driving by fusing point-cloud features from multiple connected vehicles. It constructs a Vehicle-to-Vehicle VQA dataset based on dataset V2X-Real [200], [201] for perception and planning. The model adapts LLaVA by replacing RGB encoders with a LiDAR detector, aligning scene-level and object-level features through an MLP projector and fine-tuned LoRA layers.

A key next step for VQA in autonomous driving is to evaluate model robustness under out-of-distribution conditions. Current datasets mostly feature common driving scenarios and well-structured questions, leaving models largely untested on rare events, unfamiliar objects, or challenging conditions such as night, snow, or construction zones. Developing benchmarks that explicitly include these edge cases, and assessing how well models generalize to them, is essential for deploying VQA systems in safety-critical, real-world environments.

Scene/Scenario Understanding: This subsection distinguishes between *Scene Understanding*, which focuses on static, image-based perception, and *Scenario Understanding*, which captures temporal dynamics, agent interactions, and evolving causal events.

(I) Scene Understanding: InternDrive [181] and Jain et al. [192] focus on static scene understanding using image-based inputs. InternDrive proposes a framework for driving scenario understanding, covering perception, prediction, and reasoning, using MLLM. It generates QA pairs from nuScenes using GPT-4o, followed by human correction, and fine-tunes the MLLM InternVL-1.5 via LoRA on these annotations. The resulting model analyzes driving scenes from FPV images through visual instruction tuning. Jain et al. [192] evaluate MLLM for safety-critical scene understanding using QA pairs from KITTI and nuScenes across five categories. They benchmark Video-LLaVA and GPT-4V using merged image frames and textual LiDAR summaries, applying a CoT prompting approach to enhance multimodal reasoning without requiring true temporal modeling.

(II) Scenario Understanding: In contrast, DOLPHINS [179], Ishaq et al. [174], WTS [193], and V3LMA [194] target scenario understanding, where temporal context, agent interaction, and causal reasoning are central. DOLPHINS [179] presents an MLLM-based system for human-like understanding of driving scenarios and behaviors. The model is built on OpenFlamingo and first instruction-tuned on image–instruction pairs using a Grounded Chain of Thought (GCoT), where each reasoning step is explicitly linked to visual evidence to ensure visually grounded scenario reasoning. It is then adapted to driving videos using in-context examples retrieved by VideoMAE and Ada-002. During training, only the perceiver resampler, gated cross-attention, and LoRA modules are updated, making the framework efficient while supporting multiple driving tasks. Ishaq et al. [174] propose a scenario-level spatial understanding framework that integrates short video clips, driving trajectories as text, and textual queries. They use a trajectory encoder and a Query Former to fuse the modalities, which are then passed into a frozen LLaMA-2 model with adapter layers. The model is fine-tuned by training both the Query Former and the adapters for efficient multimodal reasoning.

Specificly, WTS [193] and V3LMA [194] focus on the scenario captioning which emphasizes observable elements, reasoning targets spatial-temporal relations, intent inference, and causal analysis. WTS [193] uses GPT-3.5 externally to generate human-guided ground truth captions and evaluate model outputs via LLMScore, which assesses semantic and syntactic similarity. The proposed Instance-VideoLLM combines CLIP ViT-L/14, a Video Q-Former, and Vicuna-7B, with fine-tuning applied to the adapter and Q-Former. The model is trained on enhanced video inputs incorporating bounding boxes, gaze data, and scene context, and is compared against other off-the-shelf MLLMs. V3LMA [194] proposes a fusion method that combines pre-trained LLMs and VLMs to enhance zero-shot 3D scenario understanding. They use off-the-shelf tools for grounding, object detection, and depth estimation to generate structured scenario descriptions, which are fed into the LLM. Visual features from an MLLM are then fused at either the feature level or the classification head. Despite being zero-shot, the model achieves competitive performance, comparable to fine-tuned MLLMs.

Current MLLMs for scene and scenario understanding primarily focus on short-term temporal contexts and curated question-answering tasks, which lacks validation in realistic, real-world settings. To move toward more comprehensive scenario understanding, future work should explore long-range temporal modeling, causal inference across event sequences, and robust handling of out-of-distribution scenarios.

Risk Assessment: The goals for the MLLMs include risk detection and violation inference for anticipating hazards,

inference and scene-level safety scoring for analyzing incidents, and actionable advice generation.

One approach to risk assessment emphasizes proactive hazard mitigation through interpretable scenario understanding. For example, AccidentGPT [196] combines multi-modal perception, such as images, 3D detections, BEV features, and trajectories, with GPT-4V for zero-shot scenario captioning based on dataset DeepAccident [195] and GPT-4 for further safety evaluation using CoT and CP. It supports real-time accident prevention, post-accident analysis, and interactive safety decision-making through interpretable reasoning.

Other works focus on enabling interactive safety perception and feedback. MLLM-SUL [180] fuses multi-scale visual inputs using ResNet-101 and Swin-L for low- and high-resolution features, combined via Query Formers and Gate-Attention based on the dataset Drama-ROLISP and Drama-SRIS from HiLM-D [178]. It fine-tunes LLaMA2-7B with adapters and applies an MLP head for scene captioning and risk object localization. Similarly, ScVLM [197] proposes a multi-stage MLLM framework for risk assessment based on the nuScenes dataset, combining event type classification, conflict type identification, and narrative generation. It uses VideoLLaMA2 for zero-shot visual context extraction and LLaMA 3.1 8B to generate detailed descriptions of safety-critical events based on FPV driving videos.

A third direction emphasizes risk reasoning through structured question answering. Abu et al. [198] present a MLLM-based framework for safety-critical event detection using FPV videos from the DRAMA dataset. They compare Gemini-Pro-V1.5, Gemini-Pro-Video, and LLaVA using QA-based risk analysis with in-context learning, leveraging sliding window capture and textual context prompts to enhance risk event detection.

However, to ensure practical impact, it is critical to establish the reliability and determinism of MLLM-based risk assessments. This remains a key challenge, as MLLMs' behavior is inherently stochastic and may produce inconsistent outputs.

D. Limitations and Future Directions

MLLMs offer a unique potential to generate and analyze scenarios by leveraging their multimodal capabilities. However, there is currently no pretrained MLLM specifically devised for AD with complementary sensor modalities such as LiDAR, camera, and radar. As a future direction, this highlights the need for large-scale multi-modal datasets and pretrained MLLMs tailored to AD.

MLLM-based Scenario Generation: As reported in Table 6, only two studies have explored MLLM-based scenario generation: one targeting safety-critical scenarios and the other focused on ADAS testing. This highlights a significant research gap and suggests that the broader potential of MLLMs in this domain remains largely unexplored. Future work could extend to additional applications such as driving policy evaluation, closed-loop scenario generation, and the reconstruction of complex real-world driving events.

An emerging research direction is retrieval-augmented scenario generation. While existing retrieval-augmented generation frameworks are typically based on textual databases, MLLMs allow for the integration of multimodal knowledge bases containing maps, annotated traffic videos, and LiDAR point clouds. Such enriched context could support more diverse, realistic, and situation-aware scenario generation pipelines.

MLLM-based Scenario Analysis: As summarized in Table 7, current pre-trained MLLMs are not yet sufficient to address the complexity of driving scenario analysis. Existing models often struggle with specialized tasks that require aligning and processing diverse multimodal inputs. While fine-tuning strategies such as instruction tuning, adapter-based methods, and parameter-efficient techniques are being actively explored, these adaptations are often necessary because general-purpose pre-trained models lack sufficient domain-specific understanding. At the same time, several technical challenges must be tackled. Reliability remains a major concern, as MLLMs are prone to factual hallucinations and inconsistent output issues that are especially critical in safety-sensitive applications. Decreasing inference times is equally important. This may involve architectural innovations, model compression and distillation, or adaptation strategies that support interpretable, low-latency reasoning across multiple modalities.

From an application standpoint, promising directions include using MLLMs for high-fidelity sensor simulation and modeling complex interactions among diverse traffic participants, such as vehicles, pedestrians, and cyclists. Additionally, deploying MLLMs at the edge to support real-time situational awareness and collaborative human-machine interaction represents a valuable and unexplored opportunity for future research.

VI. Diffusion Models (DMs)

This section provides an overview of DMs, explaining their underlying generative process and tracing their conceptual evolution. Given their generative nature, DMs excel at synthesizing novel scenarios rather than analyzing existing ones. Accordingly, we survey their applications in scenario generation for AD, encompassing traffic flow synthesis, road layout design, image generation, and video generation.

A. Development of DMs

DMs are generative models inspired by non-equilibrium thermodynamics [61], mirroring natural processes like ink diffusing through water. At their core, they follow the simple yet powerful idea of systematically and gradually destroying structure in data through iterative noise addition and learning to reverse this process in a step-wise fashion. While introduced by Sohl-Dickstein et al. [61], the approach gained widespread adoption through Ho et al.'s DDPM [30].

The framework of DMs, as illustrated in Figure 8, involves two key phases: the forward process and the backward process.

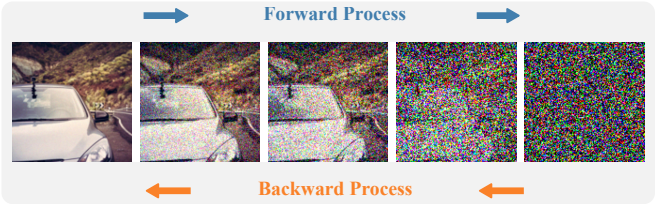

FIGURE 8. An illustration of how a DM transforms a clean image into noise through the forward process, and then reconstructs it in reverse during the backward process.

(1) Forward Process: The forward process refers to the act of gradually corrupting the original data x_0 by adding Gaussian noise over T steps, resulting in a sequence of noisy samples $x_1, x_2, \ldots, x_T$. By the final step, x_T, the sample is indistinguishable from pure noise.
(2) Backward Process: To generate realistic samples from pure Gaussian noise, a DM must learn to invert its forward corrupting process. This is achieved through an iterative denoising procedure, where the model progressively refines the noisy input to recover the underlying data distribution. At each step, the model estimates and removes the noise added during the forward process, gradually reconstructing the target sample. Denoising is typically parameterized by a neural network, such as a U-Net [30], which is trained to predict the noise component at each iteration.

Following the establishment of this paradigm, research advanced primarily along two key directions:

Controllability: Unlike the original DDPM [30], which is trained unconditionally and provides little control over the generated samples, subsequent research has developed methods to guide the diffusion process toward desired outputs. Conditioning the network on auxiliary signals, such as class labels, text embeddings, layout maps, or other modalities, enables structural constraints that guide the generative process. Classifier guidance [202] uses gradients from a separate classifier to steer sampling towards desired outputs. Classifier-free guidance [203] eliminates the need for a separate classifier by jointly training the model with and without conditioning signals, allowing adjustable control at inference. ControlNet [204] further expands controllability by incorporating spatial conditions such as edges, depth, or poses, enabling fine-grained user control.

Efficiency: The high computational cost of DDPM stems from many iterative steps at full resolution. LDMs [32] address this by operating in compressed latent spaces, reducing complexity while preserving quality. Diffusion Transformer (DiT) [205] builds on this by replacing the U-Net with a transformer backbone, improving scalability and global context modeling.

These previous innovations have enabled the use of DMs across a wide range of domains. These advances have also been adopted in large-scale commercial systems, such as Imagen, Stable Diffusion, and Adobe Firefly, which are illustrated in Figure 5 as part of the DMs' development timeline. AD is a particularly impactful area where DMs are used to generate realistic scenarios efficiently and controllably.

B. Scenario Generation

This section provides an overview of DMs for scenario generation in AD, organized by output type: dynamic traffic flow, static traffic elements, images, and videos.

Traffic Flow Generation: Traditional simulators [51], [52], [98], [206] typically rely on replaying driving logs or using heuristic-based controllers, which often do not accurately capture the complexity and adaptability of real human behavior. Recent advancements in generative models present an opportunity to create realistic and diverse traffic behavior of virtual agents directly. These models can generate the behavior (trajectories) of multiple agents over time. To serve as reliable simulation tools, such models must achieve both realism and controllability, reflecting human-like driving behaviour while adhering to customizable rules. To enhance realism, these models are typically trained on large-scale real-world driving datasets to learn the underlying dynamics and diversity of traffic behavior. In the following, we review different techniques to achieve controllability.

(I) *Gradient-Based Guidance* in DMs works by modifying the predicted mean at each denoising step using the gradient of a control objective. This perturbs the generation toward samples that better fulfill the objective while still following the underlying diffusion process. Depending on how the objective is defined, such guidance can either enforce safety constraints or, conversely, induce adversarial and safety-critical scenarios. CTG [207] incorporates Signal Temporal Logic (STL) to encode traffic rules, using the robustness score of STL as a measure of how well the rules are followed and leveraging its gradient to guide trajectory sampling. CCDiff [208] leverages the gradient of a constrained Markov Decision Process (MDP) to guide trajectory generation for multiple agents, with the MDP encoding specific control goals such as causing collisions. Before applying guidance, a causal reasoner ranks agents based on inter-agent influence and restricts guidance to the most impactful subset to improve efficiency and effectiveness. DiffScene [209] defines three differentiable objectives: safety-critical (maximizing collision risk), functional (hindering ego task completion), and constraint-based (enforcing realism rules). Lu et al. [210] extend DiffScene by encouraging adversarial agents to exhibit aggressive maneuvers (via acceleration/yaw rate variability) and manipulate traffic density around the ego vehicle. AdvDiffuser [211] trains a model to predict how likely a scenario causes failures for a given planner and uses this signal to guide the sampling process. SafeSim [212] and VBD [213] generate potential trajectories and identify those that would lead to collisions, then use guided diffusion

to denoise them. A different approach is proposed by Zhong et al. [214] and LD-Scene [215] , both of which leverage an LLM to translate natural language instructions (e.g., "aggressive lane change") into differentiable guidance functions, bridging high-level intent with low-level control.

(II) *Architecture Conditioning* embeds the control signal directly within the network's structure so that constraints are enforced throughout each iteration, rather than being injected afterwards as an external correction. DM achieve this by accepting extra conditioning inputs, such as tokens that carry agent attributes, scene statistics, language descriptions, or spatial masks. These additional inputs are processed by dedicated layers, for example, cross-attention blocks or inpainting modules, and are fused with the latent scene representation at each denoising iteration. Pronovost et al. [217] encode agent attributes (speed, heading) and global scene properties (agent density) as tokens processed by cross-attention layers. SceneDiffuser [218] frames trajectory generation as an inpainting task on a 3D tensor of shape $A \times T \times D$, each representing agents, timesteps, and features. Scene editing and agent injection are made possible by adjusting the scene tensor and the associated inpainting mask. DriveGen [216] uses a natural language description to generate road layouts and place vehicles via an LLM. A VLM is applied afterwards to analyze the BEV to identify potential future goals. Finally, a DM generates realistic trajectories from each vehicle's initial state to its predicted goal. DriveSceneGen [219] addresses two key problems: scene initialization and rollout. It first synthesizes a BEV image of road layouts and agent positions using a DM, then vectorizes the output for trajectory prediction with a Motion Transformer (MTR). SLEDGE [220] and ScenarioDreamer [221] address the same task but optimize the generation pipeline. Specifically, SLEDGE introduces a raster-to-vector autoencoder to compress scenes into latent maps for further diffusion, whereas ScenarioDreamer further advances this by operating the DM directly in vector space. Together, these methods reflect a progression from pixel-level (DriveSceneGen) to compressed-raster (SLEDGE) to fully vectorized (ScenarioDreamer) generation.

(III) *Preference Optimization (PO)* moves away from gradient-guidance and architecture-conditioning. Instead of explicit control signals or hand-crafted loss functions, Yu et al. [222] fine-tune the DM directly using PO. The model generates two candidate trajectories per scene, scores them via rule-based heuristics, and updates itself to favor the better one, thereby learning control preferences implicitly.

Despite recent advances, diffusion-based traffic flow generators still rely on manually crafted control inputs. Gradient-guided models require carefully tuned objective weights, while architecture-conditioned models depend on predefined token or mask schemas to encode rules. Adapting these approaches to new constraints often requires costly retraining or extensive fine-tuning.

Static Traffic Element: DMs have also been developed to generate various **AD** components beyond agents' trajectories.

DiffRoad [223] synthesizes 3D road layouts from structured text inputs (e.g., "two three-way intersections") and evaluates the outputs based on criteria such as smoothness and the presence of overlapping segments.

Pronovost et al. [224] and SceneControl [225] focus on generating initial agent placements for downstream traffic simulation. Pronovost et al. introduce a scene autoencoder that compresses rasterized agent layouts into latent embeddings. A DM, conditioned on a road map, is then trained over these embeddings, and a decoder reconstructs oriented bounding boxes for the agents. SceneControl offers additional flexibility through guided sampling, allowing fine-grained user control (e.g., enforcing speed constraints) and realism guarantees (e.g., collision avoidance and lane adherence) during the generation process. To assess how well the generated scenes match real-world data, both methods compare statistical distributions between real and synthetic datasets.

These static-scene generators still have notable gaps. When a DM is used to synthesize road layouts, fine-grained elements such as traffic signs, signals and lane markings are often omitted. As a result, the resulting maps lack the fidelity needed for high-realism driving simulation. Moreover, initial-scene generators are also highly map-specific: they absorb the spatial priors of the training corpus and can place agents unrealistically when applied to unseen road geometries or regions with different driving conventions.

Image Generation: Reliable AD perception depends on large annotated datasets. DMs offer an efficient alternative by generating realistic street-view images.

Text2Street [226] decomposes structured prompts, such as "a street view image with a crossing, 4 lanes, 3 cars, 2 persons, and 2 trucks on a sunny day", into three distinct components: road topology, object layout, and weather condition. Each of these components is handled by a dedicated DM. The first model processes the road topology to generate a BEV road layout. The second model takes this BEV layout and incorporates the object layout, producing a map that includes vehicles, pedestrians, and other foreground elements. The third model transforms this BEV representation into a realistic camera-view street scene. To handle geometric conditions more effectively, GeoDiffusion [227] converts bounding boxes into textual prompts that guide a pre-trained text-to-image DM. This involves translating continuous bounding box locations into discrete tokens and balancing the visual prominence of foreground objects with the often-dominant background regions during image generation. Baresi et al. [242] generate rare OOD driving scenarios (e.g., snow, desert) using three diffusion-based strategies: instruction editing, inpainting, and inpainting with refinement. Meanwhile, other works have focused on generating multi-view images. BEVControl [228] addresses the complexity of editing dense segmentation maps by using editable BEV sketches as input. It introduces

TABLE 8. Summary of Scenario Generation Studies Using DMs.

Category (Output)	Safety critical scenario?	Input				Controll -ability[1]	Controllable Factor[2]	Technique	Base Model	Dataset	Paper
		Road Topology	Initial State	Text Prompt	Bounding Boxes						
Traffic Flow	No	✓	✓			◑	Speed Goal Waypoint	STL as Guidance	DDPM	nuScenes	CTG [207]
				✓		●		LLM-Driven Scene Initialization	DiT	Argoverse 2	DriveGen [216]
		✓				◑	Traffic Density Agents' Position Agents' Speed Agents' Size	Architecture Conditioning	LDM	Argoverse 2	Pronovost et al. [217]
		✓				◑	Traffic Density	Architecture Conditioning	DiT	WOMD	SceneDiffuser [218]
						○		Map-Free Scene Generation	LDM	WOMD	DriveSceneGen [219]
						○		Raster-to-Vector Representation	DiT	nuPlan	Sledge [220]
				✓		◑	Traffic Density Road Layout	Vectorized Latent Diffusion	LDM	WOMD nuPlan	Rowe et al. [221]
		✓	✓			◑	Speed Goal Waypoint	Preference Optimization	DiT	nuScenes	Yu et al. [222]
	Yes	✓	✓			◑	Collision Type	MDP as Guidance	DDPM	nuScenes	CCDiff [208]
		✓	✓			◑	Speed	Gradient-Based Guidance	DDPM	CARLA	DiffScene [209]
		✓	✓			◑	Traffic Density Speed	Gradient-Based Guidance	DDPM	nuScenes	Lu et al. [210]
		✓	✓			○		Gradient-Based Guidance	LDM	nuScenes	AdvDiffuser [211]
		✓	✓			◑	Dirving Style Collision Type	Partial Diffusion	DDPM	nuPlan nuScenes	SafeSim [212]
		✓	✓			◑	Driving Style	Gradient-Based Guidance	DiT	WOMD	VBD [213]
		✓	✓	✓		●		LLM-Generated Loss Function	DiT	nuScenes	Zhong et al. [214]
				✓		●		LLM-Driven Scene Initialization	LDM	nuScenes	LD-Scene [215]
Static Traffic Element	No			✓		◑	Number of Lanes Type of Road	Road-UNet architecture	DDPM	OSM	DiffRoad [223]
		✓				○		End-to-End Differentiable	LDM	In-house	Pronovost et al. [224]
		✓				◑	Agents' Position Agents' Density Agent' Speed Agents' Size	Guided Agent Placement	DDPM	Argoverse 2	SceneControl [225]
Driving Image	No			✓		◑	Road Topology Traffic Density Weather	Structured Prompt	LDM DDPM	nuScenes	Text2Street [226]
					✓	◑	Camera Pose	Bounding Box Translation	LDM	nuSences	GeoDiffusion [227]
			BEV Sketch			◑	Weather Lighting Condition	Controller & Coordinator	LDM	nuScenes	BEVControl [228]
		✓			✓	◑	Camera Pose Weather Lighting Condition	Cross-View Attention	LDM	nuScenes	MagicDrive [229]
		✓		✓	✓	◑	Weather Lighting Condition Camera Pose	Dual-Branch Diffusion	LDM	nuScenes	DualDiff [230]
Driving Video	No		BEV Sequence			◑	Weather Lighting Condition Landscape	4D Attention	LDM	nuScenes	Panacea [231]
			3D Layout Sequence			◑	Weather Lighting Condition	Cascaded Video Synthesis	LDM	nuScenes	DrivingDiffusion [232]
		✓		✓	✓	●		Multi-Control Distillation	DiT	nuScenes	DiVE [233]
			Canny Edge Map Depth Map Text Prompt			◑	Weather Lighting Condition	Dual-Branch Diffusion	LDM	DriveScene -DDM [234]	DcTDM [234]
			Initial Frames			○		Frame Sampling Scheme	DDPM	WOMD	DriveGenVLM [235]
		✓		✓	✓	◑	Weather Lighting Condition	Dual-Branch Diffusion	LDM	nuScenes Waymo Open	DualDiff+ [236]
				✓		◑	Weather Traffic Density Landscape	Adapting Existing Methods	LDM	KITTI	GenDDS [237]
	Yes			✓		●		Temporal Shift Adapter	LDM	DoTA [238]	DrivingGen [239]
				✓		●		Adapting Existing Methods	DiT	MM-AU [240]	AVD2 [241]

[1] Controllability: ● Full control (users can fully customize scenes); ◑ Partial control (supports specific parameter adjustments); ○ No control.
[2] Only models with partial controllability are discussed here in this column. Fully controllable models can follow any input (typically via LLMs), and models without control fall outside the scope of this discussion.

a "controller and coordinator" mechanism to ensure that generated objects match the sketch accurately and maintain consistency across multiple viewpoints. MagicDrive [229] considers road layouts, bounding boxes, camera poses, and textual descriptions such as weather and time of day as input. It introduces a cross-view attention module that allows each camera view to access information from its immediate neighbors, ensuring visual consistency and coherence across all generated views. DualDiff [230] adopts a dual-branch architecture that separately generates foreground and background. It projects 3D occupancy data onto camera planes to form dense feature maps, fuses them with 3D bounding boxes and road maps, and then combines the branch outputs to synthesize the final image.

In spite of recent progress, fine details such as traffic signs, pole-mounted signals and lane markings are frequently simplified or omitted, resulting in generated images that fail to cover many visual corner cases that real perception stacks must handle. Photometric realism is also limited: simplified lighting models and the absence of camera artifacts such as rolling-shutter distortion, lens flare, and sensor noise create a noticeable domain gap when these synthetic frames are used to train or evaluate real-world detectors.

Video Generation: Recent work has also advanced DM-based driving video generation, improving temporal consistency, controllability, and diversity.

Several studies have introduced innovative architectures to ensure multi-view and temporal consistency in generated videos. Panacea [231] generates multi-view video sequences by first synthesizing images from BEV inputs and then expanding them along the temporal dimension. The method introduces a 4D attention mechanism that takes into account intra-view (within each camera), cross-view (between adjacent cameras) and cross-frame (between temporal patches). DrivingDiffusion [232] also employs a multi-stage approach: it first generates a consistent initial frame across all camera views from a layout, then uses a temporal model to produce short view-specific sequences, and finally refines long-term consistency via a sliding-window post-processing module. DiVE [233] focuses specifically on efficient multi-view driving scene generation. It introduces Multi-Control Auxiliary Branch Distillation (MAD) to streamline multi-condition classifier-free guidance, significantly reducing inference time. DiVE also proposes view-inflated attention, a lightweight mechanism enforcing cross-view consistency without adding parameters.

Another strategy for video generation is adapting image DMs with temporal expansion. DrivingGen [239] extends a text-to-image DM by incorporating a temporal shift adapter that efficiently propagates information across frames using modified 2D convolutions instead of costly 3D operations. Similarly, DcTDM [234] extends image-based diffusion into the temporal domain but introduces dual conditioning with dense depth maps and Canny edge maps to preserve geometric and structural consistency across frames. DriveGenVLM [235] enhances long-term video generation through conditioning and sampling strategies, such as frame-by-frame generation and keyframe interpolation, offering trade-offs between quality and speed.

In contrast, DualDiff+ [236] generates videos through a dual-branch architecture that decouples foreground and background modeling. The model first projects a 3D occupancy grid into 2D space and then fuses these features with semantic inputs, including 3D bounding boxes (foreground) and maps (background).

Another line of research advances video generation by combining and adapting existing models. GenDDS [237] fine-tunes Stable Diffusion XL [243] using LoRA [66] to produce driving images, which are then extended into videos through a temporal transformer in Hotshot-XL [244]. AVD2 [241] fine-tunes the Open-Sora 1.2 model [245] on the MM-AU [240] dataset to generate videos annotated with accident causes and avoidance strategies.

Despite recent advances, diffusion-based generators for driving videos still face significant challenges. They often struggle to maintain consistent temporal and multi-view coherence, particularly over extended clips. Additionally, their understanding of the physical world's dynamics remains limited: for example, vehicles may behave in ways that defy inertia or violate occlusion logic.

C. Limitations and Future Directions

Although recent DMs support conditioning through layout masks, language tokens, or attention-based inputs, these mechanisms often remain rigid and narrowly specialized. They typically depend on manual tuning, predefined conditioning schemas, or task-specific re-training, which limits their flexibility and scalability. To address this, future research should aim to develop more generalizable conditioning frameworks that can seamlessly integrate diverse or novel inputs without requiring substantial architectural modifications or re-training.

In parallel, while DMs often achieve strong performance on statistical realism metrics, the generated trajectories and scenes frequently lack fine-grained physical plausibility. Artifacts such as implausible inertial dynamics, unnatural agent reactions, and inadequate modeling of occlusions or causal dependencies are common. One promising direction for future research is the integration of physics-informed models, which could improve the adherence to real-world physical laws and enhance the overall realism of the generated outputs.

Moreover, although LLMs are increasingly used to convert natural language inputs into guidance signals for DMs, their potential remains underutilized. Rather than serving solely as input translators, LLMs could act as embedded knowledge sources that encode rich priors about physical dynamics, semantic scene structure, and normative driving behavior. Leveraging these capabilities may substantially improve the controllability, realism, and interpretability of

diffusion-generated scenarios, particularly in complex or ambiguous environments.

VII. World Models (WMs)

WMs are generative neural network models that learn compressed spatial and temporal representations of an environment [34]. They enable agents to develop an internal model of the world to make predictions about future states of the surrounding world environment, concerning both dynamic agents and static objects. In this section, we focus on their ability to generate driving scenarios, and we categorize recent works into visual, 3D occupancy, and multi-modal generation. Moreover, we discuss related architectural innovations and benchmarks.

A. Development of World Models

Relations with Cognitive Science: The development of WMs focuses on learning compact, predictive representations of the physical world's dynamics. This concept draws inspiration from the human brain's ability to model and predict the physics of the real world [246]. Cognitive science has proposed predictive brain models to anticipate the evolution of real-world scenarios, such as the procedural and declarative models of Downing (2009) [247] . Svensson et al. (2013) [248] apply brain-like "dreaming" to simulate perception-action sequences offline, for simple robotic systems. They define mental imagery as the brain's ability to generate and manipulate internal representations of the world, in a dreaming-like process without direct interaction with the environment. Similarly, as explained in the next section, WMs can dream by generating imagined scenarios (Figure 9), without interacting with the physical world and using their learned representations to predict the future evolution of the environment.

In a related context, Plebe et al. [249], [250] propose vision autoencoders that emulate neural convergence-divergence patterns of the brain [251], to output long-term predictions of driving scenarios in the form of videos. They suggest minimizing the free energy as a training loss function, which is inspired by the Friston's theories [252] about the brain's operation. Similar minimum-free-energy principles are also proposed for the training of WMs by LeCun [246]. For a broader review of bio-inspired cognitive agents in AD, the reader can refer to [253].

We argue that such cognitive theories will inspire the next generation of WMs, which will need to learn generalizable models of the real-world dynamics from limited data. The timeline of WMs' development is shown in Figure 5.

Architecture and Evolution of World Models: The architecture of WMs is typically based on an *encoder-decoder* paradigm [34], [245], [246], [288]. As illustrated in Figure 9, the *encoder* (also called *vision* model [34]) is used to encode multimodal inputs (images, point clouds, 3D occupancy voxels, etc.) into a latent vector z_t. Then, the future predictor (*decoder* or *memory* model [34]) predicts the future latent

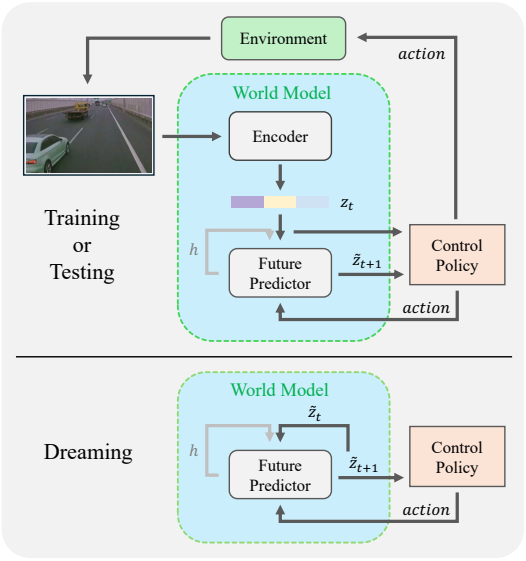

FIGURE 9. Overview of world model's training, testing and dreaming phases. In the training/testing phase (top), z_t is a latent representation of the input (e.g., image), $\tilde{z}_{t+1}$ is the prediction of the latent representation at the next time step, and h is a hidden state encoding past information. In the dreaming phase (bottom), the model generates future latent variables $\tilde{z}_t$ in an auto-regressive way: $\tilde{z}_t$ is initialized with a $\tilde{z}_0$, and then recursively fed back as input of the future predictor, which computes the next value.

representation $\tilde{z}_{t+1}$ based on z_t and an *action* provided by a given control policy. When the WM's pre-training is finished, the future predictor can be used for both motion prediction and for "dreaming" (i.e., generation) of new scenarios, never seen during training. In this regard, WMs can generate data outside the training data distribution: this is particularly valuable for AD, where rare but critical scenarios may be underrepresented in existing datasets, yet are crucial in the validation phase.

In addition to being designed individually, WMs are now increasingly integrating inspirations from LLMs, VLMs, and MLLMs, which have demonstrated a promising understanding of semantic context. More specifically, WMs are placing emphasis on using this semantic understanding for content generation [289]. Moreover, DMs play an important role as generative backbones of most modern WMs, providing stable and high-fidelity generation in both images and videos. GAIA-2 [256], DriveDreamer [288], and MagicDrive3D [264] are examples of this trend, which employ either latent or video diffusion to increase the temporal coherence and realism of the generated scenarios. Together, these examples show that WMs are developing into hybrid architectures that combine VLMs' multimodal reasoning capabilities with DMs' generative accuracy to create coherent, controllable, and semantically grounded driving simulations.

As shown in Figure 9, in the early examples [34], the vision model (encoder) was implemented as a Variational Autoencoder (VAE), and compresses high-dimensional observations into a compact latent representation. This dimensionality reduction creates a manageable state space

TABLE 9. Comparison of key research about World Models for Scenario Generation in Autonomous Driving.

Category	Input						Controllability [2]	Multi-view Generation	World Model Architecture	Model Types	Dataset	Code [3]	Paper
	Image	Text	Action	Trajectory	Geometry [1]	Map							
Visual Generation	✓	✓	✓				◑		Autoregression	Transformer [22]	In-house		GAIA-1 [254]
	✓	✓	✓		✓	✓	●	✓	Diffusion	LDM [32]	nuScenes [126]	✓	DriveDreamer [35]
	✓	✓					○		Diffusion	LDM	nuScenes, In-house		ADriver-I [255]
	✓	✓	✓	✓	✓		●	✓	Diffusion	LDM	In-house		GAIA-2 [256]
	✓	✓	✓	✓	✓	✓	●	✓	Diffusion	SVD [257]	nuScenes	✓	DriveDreamer-2 [36]
	✓	✓			✓	✓	◑		Diffusion	LDM	Waymo Open dataset [78]	✓	DriveDreamer4D [258]
	✓	✓	✓	✓			●		Diffusion	SVD	nuScenes, etc [78], [154], [259].	✓	Vista [260]
	✓			✓			○		Autoregression	Transformer	nuPlan [49], In-house	✓	DrivingWorld [261]
	✓	✓	✓		✓	✓	●	✓	Diffusion	VideoLDM [262]	nuScenes		Drive-WM [263]
	✓	✓			✓	✓	◑	✓	Diffusion	LDM	nuScenes	✓	MagicDrive [229]
		✓		✓	✓	✓	◑	✓	Diffusion	LDM	nuScenes		MagicDrive3D [264]
	✓	✓		✓	✓	✓	●	✓	Diffusion	DiT [205]	nuScenes		MagicDrive-V2 [265]
		✓			✓	✓	◑	✓	Diffusion	LDM	nuScenes, Occ3d [266]	✓	WoVoGen [267]
	✓			✓	✓	✓	◑		Diffusion	SVD	Waymo Open dataset	✓	ReconDreamer [268]
	✓	✓			✓		◑		Diffusion	DiT [205]	Cosmos [269]		Cosmos-Transfer1 [270]
	✓			✓	✓		◑		Diffusion	VideoLDM [262]	nuScenes		GeoDrive [271]
3D Occupancy Generation				✓			○		Diffusion	DiT [205]	nuScenes	✓	OccSora [272]
	✓	✓	✓	✓			●		Autoregression	Transformer	nuScenes, Lyft-Level5 [273]	✓	Drive-OccWorld [274]
				✓	✓		○		Diffusion	Latent DiT [275]	nuScenes	✓	DOME [276]
					✓		○		Autoregression	Transformer	nuScenes		RenderWorld [277]
		✓		✓	✓		◑		Autoregression	Transformer	nuScenes, etc [134], [266].		OccLLama [278]
	✓	✓	✓				◑		Autoregression	Transformer	nuScenes, Openscene [279]		DriveWorld [280]
Multi-modal Generation	✓	✓			✓	✓	◑	✓	Autoregression	Transformer	nuScenes		HoloDrive [281]
	✓		✓		✓		◑	✓	Diffusion	DDIM [31]	nuScenes, Carla		BEVWorld [282]
	✓			✓	✓		◑		Diffusion	SVD [257]	BDD [283], etc [259], [284].	✓	GEM [285]
Benchmark	✓	✓	✓	✓			●		Autoregression	Transformer	nuScenes, In-house	✓	ACT-Bench [286]
	✓	✓			✓	✓	◑	✓	Diffusion	SVD	nuScenes	✓	DriveArena [287]

[1] Geometry means 3D geometric representation and includes: 3D voxel occupancy, 3D bounding box, 3D depth, 3D segmentation and 3D point cloud.
[2] Controllability: ● Full control (models offer fine-grained scene customization with flexible control over scene elements); ◑ Partial control (models support limited or parameterized control (e.g., adjusting map); ○ No control.
[3] Code availability: "✓" means code is released open-source.

for prediction and generation. The future predictor (memory model) was implemented as a recurrent network (e.g., Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU)). The memory model captures temporal dependencies and dynamics across sequential observations, enabling the prediction of future states. Modern WMs for AD have improved this basic architecture to incorporate advanced techniques into the future predictor. For example, GAIA-1 [254] uses a transformer, and the newer GAIA-2 [256] employs a Latent Diffusion Model (LDM) [32] for future prediction and generation. Very recently, Diffusion Transformers (DiTs) [205], Stable Video Diffusion (SVD) [257] models and videoLDM [262] have gained popularity as core architectures for WMs.

As illustrated in Figure 10, WMs can generally be used for two purposes in AD: future motion prediction [290] and scenario generation [254], [256]. In this section we focus primarily on the application of WMs for scenario generation.

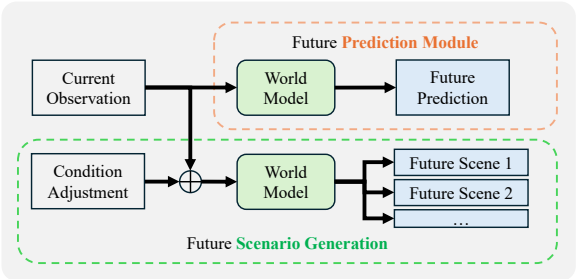

FIGURE 10. An illustration of how research on WMs for AD can be broadly categorized into two main functions: future prediction of agents' motion, and future scenario generation.

The reviewed papers, their corresponding datasets and their code availability are summarized in Table 9.

B. Scenario Generation with World Model Dreaming

WM dreaming [34] is the use of a trained WM to generate new scenarios by sampling from its learned latent space without additional real-world inputs. Once a WM has captured the underlying dynamics of an environment, it can "dream" new scenarios that follow similar physical and logical patterns as those in the training data, but with new combinations of elements and conditions that may not have been seen during training. As shown in Table 9, recent research on WMs for AD can be categorized into the following four groups.

Visual Generation: This approach focuses on creating realistic driving scenarios through the generation of images and videos. They represent the most mature category of WM applications in AD. GAIA-1 [254] pioneered the use of generative WMs for AD, by demonstrating the ability to generate diverse traffic scenarios with multiple interacting agents. GAIA-1 considers world modeling as an unsupervised sequence modeling problem, mapping multimodal inputs (video, text, and action) to discrete tokens and predicting subsequent tokens. This approach enables fine-grained control over ego-vehicle behavior and scene features, showing emerging properties such as contextual awareness and 3D geometry understanding. GAIA-2 [256] significantly advances the GAIA-1 paradigm through a latent diffusion WM that supports controllable video generation conditioned on structured inputs (e.g., ego-vehicle dynamics and agent configurations). GAIA-2 generates high-resolution, spatio-temporally consistent multi-camera videos across diverse driving environments and countries (UK, US, Germany), making it a useful tool for complex scenario simulation with good multi-view consistency.

To address the limitations of prior WMs, DriveDreamer [35] introduces a model entirely derived from real-world driving scenarios. Using its AD Diffusion Model (Auto-DMs) and a two-stage training pipeline, DriveDreamer first learns traffic structural constraints and then anticipates future states through video prediction. This approach excels in generating controllable driving videos and predicting driving policies, thereby enhancing perception tasks such as 3D detection. DriveDreamer-2 [36] extends the DriveDreamer framework [35] by incorporating an LLM to generate user-defined driving videos. DriveDreamer-2 converts user queries into agent trajectories and employs a unified multi-view model to ensure temporal and spatial coherence. It can also produce uncommon scenarios, such as abrupt vehicle cut-ins. DriveDreamer4D [258] extends the DriveDreamer framework to 4D (spatio-temporal) scene representation. By incorporating map, layout, and text conditioning, it enhances the realism of the generated data.

Unlike traditional modular designs, ADriver-I [255] introduces a unified WM using interleaved vision-action pairs to standardize visual features and control signals. Using MLLMs and diffusion, it autoregressively predicts control signals and forecasts future frames, creating a continuous simulation loop. Also following the autoregressive style, DrivingWorld [261] introduces a GPT-style WM for AD, featuring spatial-temporal fusion mechanisms. It employs next-state and next-token prediction strategies to model temporal coherence and spatial information, implementing masking and reweighting strategies to mitigate long-term drifting and improve 3D detection and motion forecasting.

As a framework for street view generation with diverse 3D geometry controls, MagicDrive [229] includes camera poses, road maps, and 3D bounding boxes, along with textual descriptions. It addresses the challenge of 3D control in traditional DMs, offering high-fidelity video generation with nuanced 3D geometry and multi-camera consistency. MagicDrive3D [264] presents a pipeline for controllable 3D street scene generation that supports multi-condition control, including BEV maps, 3D objects, and text descriptions. Unlike methods that reconstruct scenes before training, it first trains a video generation model and then reconstructs 3D scenes from generated data, enabling high-quality scene reconstruction for any-view rendering.

A world volume-aware DM is introduced by WoVoGen [267] for generating controllable multi-camera driving scenes. It operates by predicting explicit 3D world volumes to guide video generation, ensuring that multi-camera perspectives align accurately with the underlying scene geometry, and maintaining high spatial and inter-sensor consistency. While ReconDreamer [268] focuses on not only crafting WMs for driving scene reconstruction but also online restoration. It emphasizes online learning for real-time applications, allowing continuous updates to the WM as new data is acquired, which is critical for adaptability to changing conditions in AD.

With a dual-branch DM for high-fidelity video generation, GeoDrive [271] integrates 3D geometry conditions into driving WMs. It enhances spatial understanding and action controllability through 3D video rendering with dynamic editing and control for spatio-temporal consistency, improving video quality with minimal training data.

3D Occupancy Generation: 3D occupancy generation predicts and generates volumetric representations of driving environments, capturing both the spatial structure and the temporal dynamics of scenes. By treating 4D occupancy scene evolution as a video prediction task, OccSora [272] presents a novel a 4D scene tokenizer to obtain compact spatio-temporal representations. Then, it trains a diffusion transformer to generate 4D occupancy conditioned on trajectory prompts, enabling trajectory-aware simulation of various driving scenarios. To consider both static and dynamic elements in complex urban environments, Drive-OccWorld [274] combines a planner with a dynamic WM to predict 3D occupancy and flow from multi-view images. More specifically, it uses motion-aware BEV sequences as an intermediate representation, integrating multi-view video data

with motion cues to achieve robust predictions. Also aiming at improving prediction accuracy for static and dynamic objects, RenderWorld [277] further tries balancing granularity and computational efficiency. It focuses on fine-grained occupancy prediction through a novel tokenization strategy which captures spatial relationships.

Using a continuous variational autoencoder-like tokenizer, DOME [276] performs 3D occupancy prediction to preserve intricate spatial information. Unlike discrete tokenization methods, DOME's continuous approach captures subtle geometric details while maintaining computational efficiency, using probabilistic modeling to enhance robustness concerning sensor noise and occlusions. OccLLama [278] tries to integrate a multi-modal LLM as a core component for occupancy prediction. Unlike traditional models that rely solely on geometric or visual data, OccLLama uses the reasoning capabilities of LLMs to process multi-modal inputs, understanding complex scene semantics and object interactions for enhanced prediction accuracy. While DriveWorld [280] focuses on 4D scene understanding from multi-view videos. This approach is separating static spatial context from dynamic temporal changes to enable precise occupancy prediction. The model relies on self-supervised learning to reduce dependence on annotated data, thereby enhancing scalability.

Multi-modal Generation: Multi-modal generation approaches integrate multiple sensor modalities and data types as input, and output multi-modal data that can include camera images, LiDAR point clouds and depth estimation.

Aiming to address limitations of single-modality approaches, HoloDrive [281] introduces a unified framework for joint 2D-3D scene generation. It employs BEV-to-Camera and Camera-to-BEV transformation modules to bridge heterogeneous generative models. Therefore, it ensures consistency between 2D and 3D representations while using both camera images and LiDAR point clouds for the generation of consistent street scenes. Further, GEM [285] proposes a framework for generating realistic environments by integrating multi-modal sensor data, including camera images, and depth estimation. It employs a generative model based on a spatial-temporal transformer capable of predicting dynamic scene evolution regarding visual generation and depth estimation. BEVWorld [282] performed world modeling through a unified BEV latent space that also integrates multi-modal sensor inputs. The framework includes a multi-modal tokenizer and a latent BEV sequence DM that encodes multi-modal data into a unified BEV latent space. This method aims at aligning visual semantics with geometric information in a self-supervised manner.

Benchmarks: Current benchmarking frameworks provide standardized methods to assess the quality, controllability, and utility of generated scenarios, ensuring that the WMs meet the requirements for AD applications. Current evaluation frameworks mainly focus on visual realism and on the performance of downstream tasks (perception, planning, etc.). ACT-Bench [286] introduces a standardized framework to quantify action controllability, measuring how well the generated scenarios adhere to specified driving instructions. This benchmarking framework assesses the fidelity of action execution in WM-generated scenarios. DriveArena [287] is a closed-loop generative simulation platform that enables the evaluation of AD systems in dynamic and realistic environments. By simulating continuous interactions between the ego-vehicle and the environment, it bridges the gap between synthetic training and real-world deployment, supporting the iterative refinement of driving policies.

C. Limitations and Future Directions

Recent research on 3D occupancy generation with WMs has shown promising capabilities in predicting the evolution of driving environments in volumetric form. However, most models remain computationally intensive: future work should aim to develop lightweight architectures and explore finer-grained occupancy voxel representations. Recent commercial systems, such as Tesla's *Foundational Model for FSD*[9], highlight both the potential and the remaining challenges of large-scale WMs. Meanwhile, general-purpose generative WMs, such as Google DeepMind's *Genie 3* [291], can output interactive 3D environments from prompts, showing the potentital for diverse synthetic scenario generation.

Moreover, current implementations struggle with physical realism when modeling complex multi-agent interactions and real-world physics, including vehicle dynamics and kinematics laws, tire–road friction, collision forces, and weather effects. This limitation also applies to general-purpose WMs such as Genie 3, which are not tailored to AD and cannot guarantee physics-consistent modeling of vehicle dynamics and traffic rules. The generated scenarios sometimes contain physically implausible elements, such as objects that appear or disappear abruptly. Hence, the surveyed WMs can generate diverse driving scenarios but cannot accurately satisfy the laws of physics, which can lead to misleading testing results and infeasible scenarios.

VIII. Metrics Datasets, Simulators and Benchmark Challenges

In this section, we review the main evaluation metrics, datasets, simulation platforms, and benchmark challenges that serve as the foundation for scenario generation and analysis with FMs. We intentionally limit our scope to the most recent and impactful resources that are relevant for FM applications, and omit entries covered in previous work.

A. Metrics

Table 10 summarizes the main evaluation metrics from the cited papers for scenario generation and analysis with FMs.

[9]https://www.tesla.com/AI

These metrics are categorized into three main types: *(1) Framework Performance Metrics*, which assess the overall performance of frameworks; *(2) Content Quality Metrics*, which evalutate the quality and semantic accuracy of the generated or analyzed content; and *(3) Application-Specific Metrics*, which address domain-relevant aspects.

(1) Framework Performance Metrics: They evaluate the computational efficiency and operational reliability of FM-based frameworks for scenario generation and analysis.

(I) *Efficiency:* Measures the computational cost and time required for scenario generation or analysis. **Response time** refers to time from input submission to output generation, while **token usage** quantifies the total number of input and output tokens consumed during Application Programming Interface (API) key calls. These metrics are often compared against baseline approaches such as manual scripting to assess the practical benefits of FM-based frameworks [97], [100].

(II) *Effectiveness:* Refers to the operational robustness and reliability of the framework in producing valid outputs. This is commonly evaluated through **compile error rate**, the proportion of generated code or scenarios that fail to compile or parse correctly, and **execution success rate**, the percentage of scenarios that can be successfully instantiated and executed in a target environment [97], [100].

(2) Content Quality Metrics: These metrics assess the quality and semantic accuracy of the generated or analyzed content, including trajectories, semantic understanding, and language generation outputs.

(I) *Trajectory Accuracy:* Crucial for trajectory-centric generation and prediction tasks. Common metrics include **mean Average Displacement Error (mADE)**, the average Euclidean distance between predicted and ground truth trajectories across all time steps; **mean Final Displacement Error (mFDE)**, the distance at the final prediction time step from the predicted trajectory to the ground truth trajectory; and **Maximum Mean Discrepancy (MMD)**, which measures the distributional similarity between generated and real trajectory sets. Additionally, **Predictive Driver Model Score** evaluates the likelihood of predicted trajectories under human driving patterns learned from real-world data, and **Arena Driving Score** assesses overall driving competence in multi-agent scenarios by evaluating collision avoidance, goal achievement, traffic rule compliance, and interaction quality with other agents [77], [84], [88]–[90], [287], [292].

(II) *Semantic Correctness:* Assesses how well the generated scenarios or analysis outputs reflect the intended semantics of inputs like crash reports or textual prompts. Common metrics include **accuracy** or **F1 score** for evaluating scenario categorization, semantic classification, and question-answering correctness. Additionally, **completeness** and **coherence** are evaluated through human assessment, where annotators assign scores based on how thoroughly the response covers all relevant aspects and how logically consistent and well-structured the output is [79], [93], [99], [100], [102], [117], [127], [129], [131], [145], [147], [149]–[151], [172]–[174], [177], [181], [183], [184], [189], [192]

(III) *Language Quality:* Evaluates how similar generated text is to human-written reference sentences, measuring fluency, relevance, and coherence based on word overlap, structure, and meaning. Traditional metrics include **Bilingual Evaluation Understudy (BLEU)**, which measures word and phrase (n-gram) overlap focusing on precision; **Consensus-based Image Description Evaluation (CIDEr)**, which uses weighted n-grams giving more importance to informative words; **Metric for Evaluation of Translation with Explicit ORdering (METEOR)**, which considers exact matches, stem matches, and synonyms for both precision and recall; and **Recall-Oriented Understudy for Gisting Evaluation - Longest common subsequence (ROUGE-L)**, which measures content similarity using the longest common subsequence focusing on recall. However, these word-level metrics may not capture semantic nuances. To address this, **GPT Score** leverages ChatGPT's reasoning capabilities to evaluate prediction quality and semantic meaning, assigning scores. Additionally, **human evaluation scores** provide direct assessment of output quality by human annotators who rate the generated content based on observed details [117], [119], [120], [132], [152], [155], [179], [189], [192], [193].

(3) Application-Specific Metrics: They address domain-specific aspects of AD scenarios, focusing on safety-critical properties and user-specified constraints.

(I) *Safety-Criticality:* Evaluate the risk levels and safety-critical properties of generated scenarios. Key metrics include **collision rate**, the frequency of collisions occurring in the scenario; **Time-to-Collision (TTC)**, the time remaining before a potential collision; **Risk score**, a comprehensive assessment of scenario danger level; **Accuracy** for evaluating safety criticality of scenarios; and **violation discovery**, the ability to identify and detect safety-critical events or rule violations in the generated scenarios [74], [76], [121], [159], [160], [163], [164], [186], [207]–[213], [215], [217], [221].

(II) *Controllability:* Measures the framework's ability to follow user-specified constraints and control signals. Key metrics include **CLIP Alignment Score**, which measures alignment between visual content and textual prompts via cosine similarity in CLIP's shared embedding space; **Accuracy**, which evaluate the correctness of generated content against specified control signals, such as verifying the presence, location, and class of elements via object detection; and **traffic flow compliance**, which assesses adherence to constraints such as speed, waypoints, lane assignments, vehicle counts, and scene type specifications [125], [207], [209], [210], [212], [214], [222], [226]–[234], [236].

(III) *Realism:* Measures the realism of generated scenarios across multiple modalities. For traffic flow, metrics include **Wasserstein Distance (WD)** and **Kullback–Leibler divergence (KLD)** for statistical realism of motion dynamics (e.g., acceleration, jerk), **Frechet Distance** and **Symmetric Segment-Path Distance (SSPD)** for spatial differences

TABLE 10. Overview of evaluation metrics for foundation model-based scenario generation and analysis.

Category	Sub-Category	Metric	Task		Model					Output	Citations
			Gen	Ana	LLM	VLM	MLLM	DM	WM		
Framework Performance	Efficiency	Response Time	✓		✓	✓				S	[97], [100]
		Token Usage	✓		✓	✓				S	[97], [100]
	Effectiveness	Compile Error Rate	✓		✓		✓			S	[97], [100], [184]
		Execution Success Rate	✓		✓					S	[97], [100]
Content Quality	Trajectory Accuracy	mADE, mFDE	✓		✓				✓	Tr	[77], [84], [88]–[90], [286]
		MMD	✓		✓					Tr	[77], [84], [88]–[90]
		Predictive Driver Model Score	✓						✓	Tr	[287]
		Arena Driving Score	✓						✓	Tr	[287]
	Semantic Correctness	Accuracy / F1 Score	✓	✓	✓	✓	✓			T/S	e.g. [93], [99], [100], [127], [147]
		Completeness, Coherence		✓	✓					T	[102]
	Language Quality	BLEU, CIDEr, METEOR, ROUGE-L		✓		✓	✓			T	e.g. [117], [119], [138], [141], [276]
		GPT Score		✓		✓				T	[132], [152]
		Human Evaluation		✓			✓			T	[179], [192], [193]
Application-Specific	Safety-Criticality	Collision Rate	✓	✓	✓	✓		✓		S/Tr	e.g. [74], [76], [121], [210], [211]
		TTC	✓	✓		✓	✓			T/S/Tr	[76], [108], [121], [159]
		Risk Score	✓	✓	✓					T/Tr	[108]
		Accuracy		✓		✓		✓		T	e.g. [162]–[164], [180]
		Violation Discovery	✓		✓		✓			S	[186]
	Controllability	CLIP Alignment Score	✓					✓		I/V	[226], [228]
		Accuracy	✓		✓				✓	S/I/V	e.g. [125], [227]–[234], [236]
		Traffic Flow Compliance	✓					✓		Tr	e.g. [207], [209], [210], [212], [214]
	Realism	WD, KLD	✓					✓		Tr	e.g. [124], [127], [276], [293], [294]
		SSPD, Frechet Distance	✓					✓		Tr	[209], [221]
		Off-Road Rate	✓					✓		Tr	e.g. [207], [212], [213], [217], [222]
		Lane Heading Distance	✓					✓		Tr	[217]
		FID, RMSE	✓					✓		V	[125], [226]–[230], [295]
		FVD, KVD	✓					✓	✓	I/V	[231]–[236], [239], [241], [296]
		Video Panoptic Quality	✓						✓	V	[274], [294]
		mIoU	✓						✓	O	[276]
		Chamfer Distance	✓					✓		O	[212], [217]
		Human Evaluation	✓			✓				S	[124], [127]
		Scenario Consistency	✓			✓	✓			S	[124], [127], [184]
	Diversity	Statistical Distribution	✓	✓	✓	✓	✓			T/S/I	[69], [71], [125], [156], [184]
	Grounding	IoU / mIoU		✓		✓			✓	T	[138], [141], [178], [183], [276]
		3D mAP		✓		✓				T	[136], [153]
		L1/L2 Localization Error		✓			✓			T	[178], [183]
	Classification	Accuracy, Precision, Recall, Confusion Matrix		✓	✓		✓			T	[138], [141], [178], [183], [276]

Task: Gen = Generation, Ana = Analysis. Checkmarks indicate applicable tasks.
Output: T = Text (question answering), S = Script (executable code), Tr = Trajectory (single or multi-agent paths), I = Image (2D scenes), V = Video (temporal sequences), O = Others (point cloud, 3D occupancy, depth map).

between simulated and ground truth trajectories, **Off-Road Rate** for unrealistic trajectory generation, and **Lane Heading Distance** for alignment between vehicle orientation and lane direction. For image generation, **Frechet Inception Distance (FID)** measures distributional discrepancies, and **Root Mean Squared Error (RMSE)** evaluates pixel-level accuracy. For video generation, **Frechet Video Distance (FVD)**, **Kernel Video Distance (KVD)**, and **Video Panoptic Quality** assess temporal coherence and statistical similarity. For 3D scenarios, **mean Intersection-over-Union (mIoU)** evaluates occupancy prediction, and **chamfer distance** measures point cloud similarity. Additionally, **scenario consistency**, and **human evaluation** assess overall scenario quality and realism [105], [107], [124], [125], [127], [184], [207]–[209], [212], [213], [215], [217], [221], [222], [226]–[236], [239], [241], [276], [293]–[296].

(IV) *Diversity:* Captures the variability of generated scenarios by analyzing **statistical distributions** of features such as lane counts, edge counts, route lengths, and vehicle densities [69], [71], [156].

(VI) *Grounding:* Evaluates how accurately models can ground textual descriptions to visual elements and understand spatial relationships in the driving scene. Key metrics include **Intersection-over-Union (IoU)** and **mIoU** for 2D and 3D object localization accuracy, **3D mean Average Precision (mAP)** for detecting and localizing objects in 3D space, and

TABLE 11. Overview of impactful and recent datasets for foundation model-based scenario generation and analysis.

	Dataset	Year	Real	View	Sensor Data				Annotation			Traffic Condition			
					Image	LiDAR	RADAR	Traj.	3D	2D	Lane	Weather	Time	Region	Jam
Impactful	HighD [72]	2018	✓	BEV	RGB			✓		✓			D	H	✓
	nuScenes [126]	2020	✓	FPV	RGB	✓	✓	✓	✓	✓		✓	D/N	U	
	Waymo Open [78]	2020	✓	FPV	RGB	✓		✓	✓	✓	✓	✓	D/N	U/S	
	DRAMA [187]	2022	✓	FPV	RGB			✓		✓			-	U	✓
Most Recent	Comma2k19 [297]	2019	✓	FPV	RGB			✓	✓				D/N	U/S/R/H	✓
	Toronto3D [298]	2020	✓	BEV	RGB	✓		✓	✓		✓		D/N	U	✓
	A2D2 [299]	2020	✓	FPV	RGB	✓	✓	✓	✓		✓	✓	D	U/S/R/H	✓
	WADS [300]	2020	✓	FPV	RGB	✓	✓	✓	✓			✓	D/N	U/S/R	✓
	SeethroughFog [301]	2020	✓	FPV	RGB	✓	✓	✓	✓		✓	✓	D/N	U/S/R/H	✓
	Leddar PixSet [302]	2021	✓	FPV	RGB	✓		✓	✓	✓		✓	D/N	U/S/R	✓
	ZOD [303]	2022	✓	FPV	RGB	✓	✓	✓	✓	✓	✓	✓	D/N	U/S/R/H	✓
	IDD-3D [304]	2022	✓	FPV	RGB	✓			✓	✓			-	R	✓
	CODA [154]	2022	✓	FPV	RGB	✓	✓	✓	✓	✓	✓	✓	D/N	U/S/R	
	SHIFT [305]	2022	✓	FPV	RGB	✓	✓	✓	✓	✓	✓	✓	D/N	U/S/R/H	✓
	DeepAccident [195]	2023		FPV/BEV	RGB/S	✓			✓	✓	✓	✓	D/N	U/S/R/H	✓
	Dual_Radar [306]	2023	✓	FPV	RGB	✓	✓	✓	✓		✓	✓	D/N	U	
	V2V4Real [307]	2023	✓	FPV	RGB	✓		✓	✓		✓		-	U/S/H	
	SCaRL [308]	2024		FPV/BEV	RGB/S	✓	✓	✓	✓	✓	✓	✓	D/N	U/S/R/H	✓
	MARS [309]	2024	✓	FPV	RGB	✓	✓	✓	✓	✓	✓	✓	D/N	U/S/H	
	Scenes101 [310]	2024	✓	FPV	RGB			✓			✓	✓	D/N	U/S/R/H	
	TruckScenes [311]	2025	✓	FPV	RGB	✓	✓	✓	✓		✓	✓	D/N	H/U	

Impactful: We define a dataset's impact by the number of times it was used—not simply cited—by the papers included in our survey. Using this criterion, the four most impactful papers are associated with the following datasets: nuScenes (52 uses), Waymo Open (19), DRAMA (4), and HighD (3).
View indicates: FPV = First-person View, BEV = Bird's-eye View; **Image** indicates: RGB = Red, Green, Blue; S = Stereo; **Traffic Condition** includes: D/N = Day/Night; U/S/R/H = Urban/Suburban/Rural/Highway;Jam = presence of traffic congestion.

L1/L2 localization error for measuring spatial deviation between predicted and ground truth object positions [136], [138], [141], [153], [178], [183], [276].

(VII) *Classification:* Assesses the accuracy of categorizing scenarios, behaviors, or driving conditions. Common metrics include **accuracy** for scenario type identification, **confusion matrix** for understanding misclassification patterns, and **precision/recall** for specific safety-critical event detection [108], [109], [197].

B. Datasets

A typical use of FMs for scenario-based testing is to reproduce real-world scenarios in a simulation environment and reconstruct the corresponding events. LLMs typically use agents' trajectory data from given datasets, while VLMs or MLLMs can leverage additional input modalities such as LiDAR point clouds, RGB images or video streams, and rich annotations. Specifically, DMs use inputs such as RGB images, trajectories, and potentially LiDAR data to generate realistic future scenes or motion patterns through iterative refinement. In contrast, WMs aim to learn the underlying dynamics of driving environments by encoding multimodal sensor data (e.g., images, LiDAR, trajectories) and predicting future states or scene evolutions. Meanwhile, for scenario analysis, a common approach is to leverage VLMs or MLLMs to analyze driving scenes, using image or video data, with or without LiDAR or HD maps, across different tasks such as perception, prediction, and reasoning.

To assess the relevance and applicability of datasets, we adopt the categorization scheme introduced by Ding et al. [54]. This scheme enables a structured comparison across datasets, considering their sensor coverage, annotation depth, scene diversity, and potential for controllable generative tasks. In the context of FMs, which require large, diverse, and annotated data, the choice of dataset properties is fundamental to enhance the model's generalization potential. We apply this categorization to a selection of impactful and most recent datasets in Table 11, using [54] to categorize the dataset's properties given below.

(1) Sensor Data: High-quality datasets like Waymo [78] and nuScenes [126] offer diverse sensor modalities including RGB cameras, LiDAR, and RADAR. Such multimodal input is especially important for pre-training and aligning LLMs, VLMs, DMs, and WMs across visual and spatial reasoning tasks.

(2) Annotation: These datasets also include detailed 2D and 3D object annotations, lane information, and agent trajectories. This level of semantic and geometric detail supports tasks such as perception, prediction, map-conditioned scenario generation, and safety analysis.

(3) Traffic Condition: Traffic condition describes when and where the data was collected, including time of day (day/night), environment type (urban, suburban, rural,

TABLE 12. Overview of impactful and recent simulators for foundation model-based scenario generation and analysis.

	Simulator	Year	Backend	Open Source	Realistic Perception	Custom Scenario	Map Source		API Supports			DSL Support
							Real World	Human Design	Python	C++	ROS 2	
Impactful	CARLA [51]	2017	UE4	✓	✓	✓		✓	✓	✓	✓	✓
	SUMO [52]	2018	None	✓		✓	✓	✓	✓	✓		✓
	LGSVL [98]	2020	Unity	✓	✓	✓	✓	✓	✓	✓	✓	✓
	MetaDrive [86]	2021	Panda3D	✓	✓	✓	✓	✓	✓			✓
Most Recent	MATLAB AD Toolbox [312]	2018	MATLAB		✓	✓	✓	✓	✓	✓	✓	✓
	Nvidia Drive Sim [313]	2019	Nvidia Omniverse		✓	✓	✓	✓	✓	✓		
	Vista [314]	2020	None	✓	✓	✓	✓		✓			
	Nuplan [49]	2021	None	✓	✓	✓	✓	✓	✓			
	AWSIM [315]	2021	Unity	✓	✓	✓	✓	✓			✓	✓
	InterSim [316]	2022	None	✓	✓	✓	✓		✓			
	Nocturne [317]	2022	None	✓	✓	✓	✓	✓	✓	✓		
	BeamNG.tech [206]	2022	Soft-body physics		✓	✓		✓	✓		✓	✓
	Waymax [318]	2023	JAX	✓	✓	✓		✓	✓			
	TBSim [319]	2023	None	✓	✓	✓	✓	✓	✓			

Impactful: We identify the impact of a simulator following the same criterion of Table 11, based on the number of times the simulator was used—not simply cited—by the papers included in our survey. The most impactful simulators are CARLA (8 uses), MetaDrive (4), LGSVL (3), and SUMO (3).

highway), and presence of traffic congestion. These factors affect visibility, traffic flow, road layout, and driving behavior, providing diverse scenarios for evaluating autonomous driving performance.

Datasets such as Waymo Open [78] and nuScenes [126] are particularly widespread in the literature. This is largely due to their real-world fidelity, rich multisensor coverage, and comprehensive annotations, which make them ideal for training and evaluation of FMs. Additionally, it is worth noting that emerging (Visual) QA datasets relevant to scenario analysis with language FMs are discussed in Section IV-C and Section V-C.

C. Simulators

Simulation platforms are essential in the development and evaluation pipeline of AD systems. They enable safe and reproducible testing, large-scale scenario generation, and structured benchmarking. For FM-based scenario generation, simulators are particularly valuable for generating training data, enabling self-supervised pre-training, and facilitating the sim-to-real validation. FM-based scenario generation can be performed by LLMs/VLMs/MLLMs through either API functions or DSLs, allowing automatic script generation and scenario execution. Table 12 summarizes the impactful and recent simulation platforms that are relevant to scenario generation and analysis. For the classification and evaluation of the existing simulators, we extend the categorization scheme introduced by Ding et al. [54], focusing on features especially relevant to the development and application of FMs.

(1) Backend: The simulation backend defines the physical and rendering engine used to generate sensor data and simulate interactions. Platforms such as Unreal Engine 4 (UE4) or Unity enable high-fidelity rendering and realistic vehicle dynamics, which are valuable for training perception-driven foundation models. Lightweight or symbolic backends, like SUMO or Nocturne, are useful in large-scale planning and decision-making datasets where rendering realism is less critical.

(2) Realistic Perception: Simulators with realistic perception capabilities provide physics-based sensor outputs, including camera, LiDAR, or radar emulation. Such platforms are crucial for training vision-language FMs, sensor-fusion backbones, or multimodal WMs.

(3) Custom Scenario: The ability to define and customize traffic scenarios is a central requirement for both evaluation and data generation workflows. Particularly for FMs, automated and diverse scenario creation supports the pre-training of models on rare, safety-critical, or systematically varied interactions. Customization typically includes the placement and behavior of traffic participants, route definitions, or modifications of environmental conditions such as weather and lighting. Simulators like CARLA [51] offer rich APIs for manual customization, enabling users to script complex multi-agent interactions and adjust parameters such as vehicle behavior, density, and even scene appearance. More recently, platforms like BeamNG.tech [206] go a step further by supporting automated scenario generation at scale. This enables the procedural creation and batch testing of varied situations, making it well-suited for training and validating FMs in closed-loop settings.

(4) Map Source: We differentiate between scenarios based on real-world maps (e.g., OpenStreetMap) and those built from human design. Real-world maps ensure geographic realism and coverage, while human-designed maps enable controlled environments.

TABLE 13. Overview of foundation model Benchmark Challenges from 2022–2025, categorized by core capabilities.

	Name	Host	Tasks				
			Perception & Interpretation	Prediction & Planning	Reasoning & Decision	Language Understanding	Creative Generation
Autonomous Driving	CARLA AD Challenge [320]	CARLA					✓
	DRL4Real [321]	ICCV					✓
	Waymo Open Dataset Challenge [322]	Waymo / CVPR WAD	✓	✓			✓
	Argoverse 2: Scenario Mining Competition [323]	ArgoAI			✓	✓	
	Roboflow-20VL [324]	Roboflow-VL / CVPR	✓			✓	
	AVA Challenge [325]	AVA Challenge Team	✓	✓	✓	✓	
Other Fields Related to Generation and Analysis	IGLU Challenge [326]	NeurIPS / IGLU Team		✓	✓	✓	
	LLM Efficiency Challenge [327]	NeurIPS				✓	
	MMWorld [328]	CVPR			✓		
	3D Scene Understanding [329]	CVPR	✓			✓	
	Trojan Detection [330]	NeurIPS / CAIS				✓	
	SMART-101 [331]	CVPR	✓		✓	✓	
	NICE Challenge [332]	CVPR / LG Research	✓		✓	✓	
	SyntaGen [333]	CVPR	✓				✓
	Habitat Challenge [334]	CVPR / FAIR	✓	✓	✓		
	BIG-bench [335]	Google Research			✓	✓	
	BIG-bench Hard (BBH) [336]	Google Research			✓	✓	
	HELM [337]	Stanford CRFM			✓	✓	
	MMBench [338]	OpenCompass	✓		✓	✓	
	MMMU [339]	CVPR / U-Waterloo / OSU	✓		✓	✓	
	Open LLM Leaderboard [340]	VILA-Lab				✓	
	Text-to-Image Leaderboard [341]	Artificial Analysis				✓	✓
	Ego4D [342]	FAIR	✓	✓	✓	✓	
	VizWiz Grand Challenge [343]	CVPR VizWiz Workshop	✓			✓	
	MedFM [344]	NeurIPS / Shanghai AI Laboratory	✓				

(5) API-Supports: API support determines how flexibly simulators can be integrated into training pipelines. Python interfaces are especially useful for data generation and model interaction. Robot Operating System (ROS 2) compatibility allows for testing learned policies in robotics stacks, while C++ APIs provide performance for real-time validation and closed-loop deployment.

(6) Domain-Specific Language (DSL) Support: Some simulators provide DSL that enable structured, human-readable scenario specification through high-level functions or syntax. These interfaces are especially useful for integrating LLMs/VLMs/MLLMs in automated scenario generation pipelines.

Based on these criteria, two simulators stand out in Table 12 as particularly impactful in FM research: CARLA [51] and SUMO [52]. Their complementary capabilities make them well-suited to different aspects of scenario generation and evaluation. SUMO, a microscopic traffic simulator, is designed for large-scale traffic modeling and interaction-heavy scenario simulation at the population level. It supports integration with real-world maps via OpenStreetMap, allowing for geographically accurate traffic flow simulations. These features make it a practical backend for LLMs tasked with generating or editing traffic configurations using natural language prompts or structured templates. CARLA, in contrast, is a macroscopic simulator with high-fidelity physics, sensor simulation, and photorealistic rendering. It is widely used for ego-agent policy testing in closed-loop environments. Its integration with platforms like Scenic [75] enables programmatic scenario definition through interpretable formal languages, while its Python API offers fine-grained control over agent behavior, environmental settings, and sensor configurations. These characteristics make CARLA particularly suitable for LLMs, VLMs, and MLLMs in vision-language understanding, closed-loop control, and multimodal reasoning.

D. Challenges and Benchmarks

In addition to static datasets and simulation environments, open challenges and benchmarks have become useful tools to evaluate the performance of FMs. While datasets provide the raw material for training and offline testing, challenges enable comparative analysis across models in a controlled and competitive setting. To our knowledge, this is the first survey to systematically categorize and compare challenges and benchmarks relevant to scenario generation and analysis. Although many of these challenges originate in other application domains, such as medical imaging, robotics, or general-purpose language understanding, their underlying task structures often align with those found in AD. For example,

interpreting sensor input, forecasting agent behavior, making multi-step decisions, or generating new representations (e.g., scenes, trajectories, or instructions) are all core operations in scenario understanding. Table 13 presents a selection of challenges and benchmarks published between 2022 and 2025 while our work features a selective overview. The challenges highlight both direct contributions from autonomous driving, such as the Waymo Open Dataset Challenge [322], the Argoverse 2 Scenario Mining Competition [323], and the Accessibility Vision and Autonomy (AVA) Challenge [325], as well as structurally similar benchmarks from other fields. For example, while the Argoverse 2 challenge already touches on scenario analysis, it has not involved scenario generation yet. In contrast, tasks such as VQA, egocentric video understanding, or synthetic image generation often require models to interpret complex scenes and produce new, coherent outputs, an ability that is equally fundamental for scenario generation. Challenges like SyntaGen [333] and the Text-to-Image Leaderboard [341] illustrate this parallel particularly well: models are asked to generate synthetic examples that exhibit structural realism and diversity. Each challenge is categorized along five core capabilities:

(1) Perception & Interpretation: This category refers to the model's ability to process sensor inputs and extract meaningful semantic representations. Benchmarks such as MMBench [338] and MMMU [339] require fine-grained visual understanding across diagrams, images, and structured visual data. The MedFM [344] challenge focuses on extracting clinically relevant patterns from medical images such as X-rays and histology slides. Ego4D [342] evaluates perception in the context of egocentric video, where models must interpret long, unstructured streams of first-person footage.

(2) Prediction & Planning: Challenges in this category require models to forecast future events or plan a sequence of actions based on partial observations. The Waymo Open Dataset Challenge [322] is a prominent example, assessing motion forecasting from multi-agent sensor streams in real-world traffic scenarios. In the Habitat challenge [334], embodied agents must navigate photo-realistic indoor environments toward semantic or visual goals.

(3) Reasoning & Decision Making: This capability includes commonsense reasoning, causal inference, and multi-hop planning. The BIG-bench [335] and BIG-bench Hard (BBH) [336] benchmarks target difficult problems in logic, mathematics, and abstract reasoning, many of which remain unsolved even by large models. SMART-101 [331] evaluates reasoning in dialogue, specifically whether models can generate helpful, honest, and harmless responses.

(4) Language Understanding & Generation: This encompasses tasks such as instruction following, QA, summarization, and dialogue generation. The LLM Efficiency Challenge [327] evaluates how well FMs can be fine-tuned under strict computational budgets. HELM [337] offers a multi-dimensional evaluation across more than a dozen application domains, measuring not only task performance but also fairness, bias, and calibration. The Open LLM Leaderboard [340] provides a public ranking of open-source language models based on standardized evaluations across tasks such as QA or summarization.

(5) Creative Generation: Finally, this category captures the ability of a model to generate complex artifacts such as images, captions, or synthetic data samples. The Text-to-Image Leaderboard [341] evaluates diffusion-based generative models using human preference judgments over image outputs. SyntaGen [333] tests whether DMs can generate synthetic images that preserve sufficient structure and diversity to train robust perception models.

Overall, these benchmarks provide a structured landscape for measuring and comparing the capabilities of FMs beyond narrow task-specific metrics. They reflect the growing demand for models that are not only accurate but also general, adaptable, and robust across domains. For instance, the Ego4D [342] benchmark requires models to understand egocentric video data across diverse daily contexts such as households, workplaces, and outdoor activities. In contrast, MedFM [344] evaluates the ability to analyze complex medical images, requiring high precision and domain-specific knowledge. Despite their differing domains, both tasks rely on similar underlying capabilities, illustrating the versatility required from FMs.

IX. Open Research Questions and Challenges

In this paper, we illustrate how the state of the art in the emerging field of scenario generation and analysis with FMs is quite extensive. Nevertheless, there are still some open research questions and challenges. Here, we present a list of open challenges based on additional discussions with leading researchers and experts in the field. These challenges open new research questions to use FMs for scenario generation and analysis in AD.

Challenge 1 – Balancing Plausibility and Edge Case Generation: Effective scenario generation requires balancing realism with the ability to capture rare edge cases. Realistic scenarios demand that FMs abstractly understand the real-world dynamics [345]. On the other hand, edge cases essential for safety assurance [346] often approach the boundary of perceived plausibility, making them challenging for FMs to generate without producing unrealistic outcomes. When the plausibility of the generated scenarios is compromised, the resulting scenarios cannot support safety assurance arguments [53]. Thus, the key challenge is ensuring the realism of the generated scenarios, while enabling FMs to generalize and capture critical edge-case situations.

Challenge 2 – Large-Scale Multimodal Data Availability: Many FMs are trained on existing datasets that struggle to capture the full diversity of real-world driving scenarios. Moreover, the integration of multimodal data such as LiDAR, camera, RADAR, and text remains limited compared to single-modality FMs. This is due to the lack of open-source LiDAR and RADAR data at publicly accessible, internet-scale

volumes (comparable to those used to pretrain FMs on web data such as news, books, and large-scale image and video collections), and to the limited size of domain-specific multimodal datasets [41]. In addition, open datasets that include rare, diverse, and safety-critical events are still scarce. Thus, a major challenge is the limited availability of diverse, unbiased multimodal data needed to enable scenario generation with high realism and fidelity.

Challenge 3 – Standardized Evaluation Metrics and Benchmarks for Scenario Generation: Currently, there is no established standard for the automated evaluation and generation of driving scenarios. Widely accepted metrics to assess realism, plausibility, dynamic feasibility, controllability, and safety-criticality are still missing, hindering fair and meaningful comparisons among different methods. To fill this gap, open-source evaluation frameworks and community challenges or leaderboards are needed, requiring participants to generate and assess autonomous driving scenarios. Such initiatives would enable consistent benchmarking, foster the development of multi-dimensional evaluation metrics, and promote reproducible research practices. This will ultimately accelerate the integration of scenario-generation methods into safety assessment pipelines.

Challenge 4 – Safety, Robustness & Verification: Most existing methods lack formal guarantees for safety, correctness, or scenario coverage. The stochastic nature of FMs increases the risk of hallucinated outputs, limiting their reliability for AD safety assurance. A key challenge is ensuring that the generated scenarios are logically grounded and validated through formal verification, constraint satisfaction, or logic-based safety rules rather than merely correlated with the intended context.

Challenge 5 – Computational Cost and Scalability: Current FM-based generation methods demand substantial computational resources, with training requiring massive datasets, long runtimes, and high-performance hardware. Even inference and model fine-tuning are costly without advanced infrastructure. This raises unsolved challenges in scalability, accessibility, and cost-effectiveness, particularly for smaller organizations or resource-constrained applications.

Challenge 6 – Industrial Transferability and Validation: While academia offers many methods for virtual testing and evaluation, the industry must ultimately adapt them for real-world AD applications. Bridging this gap requires method validation, standardization [347], and seamless integration into existing workflows. Thus, a key research question lies in developing approaches that are not only theoretically sound but also practical, efficient, and accessible to diverse stakeholders, backed by robust industrial validation demonstrating clear benefits and adaptability.

X. Future Directions

Addressing the above-mentioned challenges in scenario generation and analysis using FMs yields several directions for future improvement and new research agendas.

Research Direction 1 – Improve Realism: Improving the realism and plausibility of the generated scenarios will require integrating domain-specific knowledge into FMs, enhancing their understanding of real-world dynamics and interactions. Hybrid approaches that combine physics-based models with data-driven FMs offer promise in generating physically coherent scenarios. Also, the exploration of dreaming with WMs [34], [291] can address gaps in sensor simulation: the data-driven nature of dreaming can capture fine-grained sensor characteristics with high fidelity.

Research Direction 2 – Create Rare Events: Capturing rare, high-risk events requires dedicated methods to systematically identify and generate such scenarios. We recommend creating targeted datasets that focus on infrequent but critical situations to improve the accuracy of models in such cases. Additionally, incorporating reasoning techniques such as causal or counterfactual reasoning [348], which may help FMs deduce plausible yet uncommon scenarios.

Research Direction 3 – Create Multimodal Datasets: Multimodal data integration remains a major challenge, requiring large-scale datasets specifically designed for scenario generation. These should combine vehicle sensor data, such as LiDAR, RADAR, and cameras, with map data, traffic rules, control actions, human feedback, and textual annotations. We also recommend developing new model architectures and training methods specifically tailored to multimodal fusion, in order to address the current limitations in scalability and integration.

Research Direction 4 – Develop Metrics and KPIs for Comparison: We heavily recommend the development of standardized evaluation methods for an objective comparison of scenarios and scenario generation approaches. This requires new benchmarks and metrics for realism, controllability, diversity, and safety-criticality, along with broad adoption by the community. Promoting these new benchmarks in competitions at the major conferences will drive progress, standardization, and community-driven innovation.

Research Direction 5 – Reduce Computational Demands: Computational efficiency and scalability present major practical constraints. Addressing them requires further investigation of techniques such as model distillation, pruning, and quantization, specifically tailored to scenario generation and analysis tasks, to minimize computational demands without sacrificing performance.

Research Direction 6 – FMs as Safe Data Flywheels: A key research direction concerns the integration of FMs into AV safety validation workflows. This includes using FMs as safe data flywheels, where the generated scenarios continuously support testing, AV models retraining, safety assessment, and performance monitoring. Future work should ensure scenario representativeness, balance real and synthetic data, and develop robust metrics to quantify the safety impact of the generated edge cases across the AV lifecycle.

Research Direction 7 – Regulatory Compliance: Ethical considerations and regulatory compliance must be integral to

future developments. Transparent methodologies are needed to identify, mitigate, and validate biases in the generation and analysis of AD scenarios. Equally important are robust approaches to data privacy management, to ensure compliance with legal and ethical standards while safeguarding sensitive training data. Advancing these aspects will also support the use of generated scenarios in safety validation and certification, contributing to structured safety arguments.

XI. Conclusion

This survey examines the state-of-the-art in FMs for autonomous driving applications, emphasizing their significant contributions to both scenario generation and scenario analysis. FMs, including LLMs, VLMs, MLLMs, DMs, and WMs, have emerged as promising tools to enhance the realism, diversity, and scalability of scenario-based testing in AD.

The versatility of FMs lies in their ability to learn from large-scale, heterogeneous datasets through self-supervised training. Their capability to generalize knowledge across various tasks has advanced the scenario-based testing paradigm, overcoming many limitations of traditional rule-based and data-driven methods. Particularly, the dual capability of scenario generation and scenario analysis presented by FMs positions them as crucial enablers for robust and efficient validation frameworks in AD systems.

Despite these advances, notable challenges persist. Achieving fine-grained controllability in safety-critical scenarios and ensuring robust realism in generated scenes are ongoing research hurdles. Computational efficiency remains a significant challenge, as many foundation models demand high memory bandwidth, high inference times, and costly GPU resources, limiting their practicality for large-scale scenario generation and real-time testing. Additionally, while the surveyed models demonstrate promising results, further research is needed to enhance the interpretability of their outputs, improve alignment with real-world traffic conditions, and systematically address out-of-distribution scenarios. Future work should also investigate if and how improvements in FM model designs and size may result in better generalization for scenario generation and analysis.

Ultimately, as autonomous vehicles approach broader operational domains and higher levels of automation, the role of advanced scenario generation and analysis methods will be paramount. FMs present a powerful framework for this evolution, promising to revolutionize both the safety and efficiency of AD development. The future trajectory of this research is expected to bring further transformative advancements, fostering safer, more reliable, and broadly accessible autonomous mobility.

ACKNOWLEDGMENT

Leadership & Structure: Y. Gao (lead), M. Piccinini, J. Betz. **Technical Sections:** Y. Gao (LLMs/VLMs/MLLMs), Y. Zhang (DMs), D. Wang and M. Piccinini (WMs), K. Moller (Datasets, Simulators, and Benchmarks). **Research Curation:** R. Brusnicki (collection of papers, GitHub repository). **Oversight & Strategy:** J. Betz, M. Piccinini, A. Gambi, K. Storms, J. F. Totz, B. Zarrouki, S. Peters, A. Stocco, B. Alrifaee, and M. Pavone contributed to critical revisions and formulating research directions.

REFERENCES

[1] J. Betz, M. Lutwitzi, and S. Peters, "A new taxonomy for automated driving: Structuring applications based on their operational design domain, level of automation and automation readiness," in *2024 IEEE Intelligent Vehicles Symposium (IV)*, 2024.

[2] Waymo, "Waymo one: The next step on our self-driving journey," 2018. [Online]. Available: https://waymo.com/blog/2018/12/waymo-one-next-step-on-our-self-driving

[3] S. International, "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles," *SAE J3016*, 2021. [Online]. Available: https://www.sae.org/standards/content/j3016_202104/

[4] L. Kolodny and J. Elias, "Waymo reports 250,000 paid robotaxi rides per week in u.s." 2025. [Online]. Available: https://www.cnbc.com/2025/04/24/waymo-reports-250000-paid-robotaxi-rides-per-week-in-us.html

[5] S. D. Pendleton, H. Andersen, X. Du, X. Shen, M. Meghjani, Y. H. Eng *et al.*, "Perception, planning, control, and coordination for autonomous vehicles," *Machines*, vol. 5, no. 1, 2017.

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

[7] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li, "End-to-end autonomous driving: Challenges and frontiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[8] W. Zheng, R. Song, X. Guo, C. Zhang, and L. Chen, "Genad: Generative end-to-end autonomous driving," *preprint arXiv:2402.11502*, 2024.

[9] T. Menzel, G. Bagschik, and M. Maurer, "Scenarios for development, test and validation of automated vehicles," in *2018 IEEE intelligent vehicles symposium (IV)*. IEEE, 2018.

[10] S. Riedmaier, T. Ponn, D. Ludwig, B. Schick, and F. Diermeyer, "Survey on scenario-based safety assessment of automated vehicles," *IEEE access*, vol. 8, 2020.

[11] A. Gambi, V. Nguyen, J. Ahmed, and G. Fraser, "Generating critical driving scenarios from accident sketches," in *2022 IEEE International Conference On Artificial Intelligence Testing (AITest)*. IEEE, 2022.

[12] A. Gambi, T. Huynh, and G. Fraser, "Generating effective test cases for self-driving cars from police reports," in *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2019.

[13] A. Cherubini, G. P. R. Papini, A. Plebe, M. Piazza, and M. D. Lio, "Bootstrapped neural models for predicting self-driving vehicle collisions with quantified confidence: Offline and online applications," *IEEE Transactions on Intelligent Vehicles*, 2024.

[14] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx *et al.*, "On the opportunities and risks of foundation models," *preprint arXiv:2108.07258*, 2021.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *preprint arXiv:2010.11929*, 2020.

[17] J. Jiang, F. Wang, J. Shen, S. Kim, and S. Kim, "A survey on large language models for code generation," *preprint arXiv:2406.00515*, 2024.

[18] Y. Huang, Y. Chen, and Z. Li, "Applications of large scale foundation models for autonomous driving," *preprint arXiv:2311.12144*, 2023.

[19] H. Gao, Z. Wang, Y. Li, K. Long, M. Yang, and Y. Shen, "A survey for foundation models in autonomous driving," *preprint arXiv:2402.01105*, 2024.

[20] Y. Wang, S. Xing, C. Can, R. Li, H. Hua, K. Tian *et al.*, "Generative ai for autonomous driving: Frontiers and opportunities," *preprint arXiv:2505.08854*, 2025.

[21] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, 2020.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[23] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[24] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, 2020.

[25] A. T. Liu, S.-w. Yang, P.-H. Chi, P.-c. Hsu, and H.-y. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

[26] P. Yin, G. Neubig, W.-t. Yih, and S. Riedel, "Tabert: Pretraining for joint understanding of textual and tabular data," *preprint arXiv:2005.08314*, 2020.

[27] H. Luo, L. Ji, B. Shi, H. Huang, N. Duan, T. Li *et al.*, "Univl: A unified video and language pre-training model for multimodal understanding and generation," *preprint arXiv:2002.06353*, 2020.

[28] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021.

[29] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2023.

[30] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, 2020.

[31] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *preprint arXiv:2010.02502*, 2020.

[32] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

[33] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," *Advances in Neural Information Processing Systems*, vol. 35, 2022.

[34] D. Ha and J. Schmidhuber, "Recurrent world models facilitate policy evolution," in *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 2018.

[35] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu, "Drivedreamer: Towards real-world-drive world models for autonomous driving," in *European Conference on Computer Vision*. Springer, 2024.

[36] G. Zhao, X. Wang, Z. Zhu, X. Chen, G. Huang, X. Bao *et al.*, "Drivedreamer-2: Llm-enhanced world models for diverse driving video generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 10, 2025.

[37] Y. Zhu, S. Wang, W. Zhong, N. Shen, Y. Li, S. Wang *et al.*, "Will large language models be a panacea to autonomous driving?" *preprint arXiv:2409.14165*, 2024.

[38] Y. Wu, D. Li, Y. Chen, R. Jiang, H. P. Zou, L. Fang *et al.*, "Multi-agent autonomous driving systems with large language models: A survey of recent advances," *preprint arXiv:2502.16804*, 2025.

[39] Y. Li, K. Katsumata, E. Javanmardi, and M. Tsukada, "Large language models for human-like autonomous driving: A survey," in *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2024.

[40] X. Zhou, M. Liu, E. Yurtsever, B. L. Zagar, W. Zimmer, H. Cao *et al.*, "Vision language models in autonomous driving: A survey and outlook," *IEEE Transactions on Intelligent Vehicles*, 2024.

[41] C. Cui, Y. Ma, X. Cao, W. Ye, Y. Zhou, K. Liang *et al.*, "A survey on multimodal large language models for autonomous driving," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

[42] S. Fourati, W. Jaafar, N. Baccar, and S. Alfattani, "Xlm for autonomous driving systems: A comprehensive review," *preprint arXiv:2409.10484*, 2024.

[43] J. Li, J. Li, G. Yang, L. Yang, H. Chi, and L. Yang, "Applications of large language models and multimodal large models in autonomous driving: a comprehensive review," *Drones*, 2025.

[44] Y. Guan, H. Liao, Z. Li, J. Hu, R. Yuan, Y. Li *et al.*, "World models for autonomous driving: An initial survey," *IEEE Transactions on Intelligent Vehicles*, 2024.

[45] S. Tu, X. Zhou, D. Liang, X. Jiang, Y. Zhang, X. Li *et al.*, "The role of world models in shaping autonomous driving: A comprehensive survey," *preprint arXiv:2502.10498*, 2025.

[46] Y. Wang, S. Xing, C. Can, R. Li, H. Hua, K. Tian *et al.*, "Generative ai for autonomous driving: Frontiers and opportunities," *preprint arXiv:2505.08854*, 2025.

[47] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan *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.

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

[49] N. Karnchanachari, D. Geromichalos, K. S. Tan, N. Li, C. Eriksen, S. Yaghoubi *et al.*, "Towards learning-based planning: The nuplan benchmark for real-world autonomous driving," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[50] M. Althoff, M. Koschi, and S. Manzinger, "Commonroad: Composable benchmarks for motion planning on roads," in *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2017.

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

[52] P. A. Lopez, E. Wiessner, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flotterod *et al.*, "Microscopic traffic simulation using sumo," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018.

[53] D. Nalic, T. Mihalj, M. Bäumler, M. Lehmann, A. Eichberger, and S. Bernsteiner, "Scenario based testing of automated driving systems: A literature survey," in *FISITA web Congress*, vol. 10, 2020.

[54] W. Ding, C. Xu, M. Arief, H. Lin, B. Li, and D. Zhao, "A survey on safety-critical driving scenario generation—a methodological perspective," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 7, 2023.

[55] B. Schütt, J. Ransiek, T. Braun, and E. Sax, "1001 ways of scenario generation for testing of self-driving cars: A survey," in *2023 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2023.

[56] Z. Yang, X. Jia, H. Li, and J. Yan, "Llm4drive: A survey of large language models for autonomous driving," *preprint arXiv:2311.01043*, 2023.

[57] H. Tian, K. Reddy, Y. Feng, M. Quddus, Y. Demiris, and P. Angeloudis, "Large (vision) language models for autonomous vehicles: Current trends and future directions," *Authorea Preprints*, 2024.

[58] A. Fu, Y. Zhou, T. Zhou, Y. Yang, B. Gao, Q. Li *et al.*, "Exploring the interplay between video generation and world models in autonomous driving: A survey," *preprint arXiv:2411.02914*, 2024.

[59] T. Feng, W. Wang, and Y. Yang, "A survey of world models for autonomous driving," *preprint arXiv:2501.11260*, 2025.

[60] S. S. Mahmud, L. Ferreira, M. S. Hoque, and A. Tavassoli, "Application of proximal surrogate indicators for safety evaluation: A review of recent developments and research needs," *IATSS research*, vol. 41, no. 4, 2017.

[61] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. pmlr, 2015.

[62] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *preprint arXiv:1301.3781*, 2013.

[63] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child *et al.*, "Scaling laws for neural language models," *preprint arXiv:2001.08361*, 2020.

[64] X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng *et al.*, "Deepseek llm: Scaling open-source language models with longtermism," *preprint arXiv:2401.02954*, 2024.

[65] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, "A systematic survey of prompt engineering in large language models: Techniques and applications," *preprint arXiv:2402.07927*, 2024.

[66] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, 2022.

[67] Y. Deng, J. Yao, Z. Tu, X. Zheng, M. Zhang, and T. Zhang, "Target: Automated scenario generation from traffic rules for testing autonomous vehicles," *preprint arXiv:2305.06018*, 2023.

[68] S. Tang, Z. Zhang, J. Zhou, L. Lei, Y. Zhou, and Y. Xue, "Legend: A top-down approach to scenario generation of autonomous driving systems assisted by large language models," in *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, 2024.

[69] E. Aasi, P. Nguyen, S. Sreeram, G. Rosman, S. Karaman, and D. Rus, "Generating out-of-distribution scenarios using language models," *preprint arXiv:2411.16554*, 2024.

[70] LangChain, "Langchain: The llm application framework," https://github.com/langchain-ai/langchain, 2023, accessed: 2025-05-26.

[71] C. Chang, S. Wang, J. Zhang, J. Ge, and L. Li, "Llmscenario: Large language model driven scenario generation," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2024.

[72] 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.

[73] C. Chang, D. Cao, L. Chen, K. Su, K. Su, Y. Su *et al.*, "Metascenario: A framework for driving scenario data description, storage and indexing," *IEEE Transactions on Intelligent Vehicles*, 2022.

[74] 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.

[75] D. J. Fremont, T. Dreossi, S. Ghosh, X. Yue, A. L. Sangiovanni-Vincentelli, and S. A. Seshia, "Scenic: a language for scenario specification and scene generation," in *Proceedings of the 40th ACM SIGPLAN conference on programming language design and implementation*, 2019.

[76] Y. Mei, T. Nie, J. Sun, and Y. Tian, "Seeking to collide: Online safety-critical scenario generation for autonomous driving with retrieval augmented large language models," *preprint arXiv:2505.00972*, 2025.

[77] S. Tan, B. Ivanovic, X. Weng, M. Pavone, and P. Kraehenbuehl, "Language conditioned traffic generation," *preprint arXiv:2307.07947*, 2023.

[78] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui *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.

[79] Y. Zhao, W. Xiao, T. Mihalj, J. Hu, and A. Eichberger, "Chat2scenario: Scenario extraction from dataset through utilization of large language model," in *2024 IEEE Intelligent Vehicles Symposium (IV)*, 2024.

[80] ASAM e.V., "Asam openscenario® dsl v2.1.0," 2024. [Online]. Available: https://www.asam.net/standards/detail/openscenario-dsl/

[81] S. Li, T. Azfar, and R. Ke, "Chatsumo: Large language model for automating traffic scenario generation in simulation of urban mobility," *IEEE Transactions on Intelligent Vehicles*, 2024.

[82] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix *et al.*, "Llama: Open and efficient foundation language models," *preprint arXiv:2302.13971*, 2023.

[83] OpenStreetMap contributors, "Planet dump retrieved from https://planet.osm.org ," https://www.openstreetmap.org, 2017.

[84] X. Li, E. Liu, T. Shen, J. Huang, and F.-Y. Wang, "Chatgpt-based scenario engineer: A new framework on scenario generation for trajectory prediction," *IEEE Transactions on Intelligent Vehicles*, 2024.

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

[86] 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, 2022.

[87] B.-K. Ruan, H.-T. Tsui, Y.-H. Li, and H.-H. Shuai, "Traffic scene generation from natural language description for autonomous vehicles with large language model," *preprint arXiv:2409.09575*, 2024.

[88] A. Aiersilan, "Generating traffic scenarios via in-context learning to learn better motion planner," *preprint arXiv:2412.18086*, 2024.

[89] S. Tan, B. Ivanovic, Y. Chen, B. Li, X. Weng, Y. Cao *et al.*, "Promptable closed-loop traffic simulation," *preprint arXiv:2409.05863*, 2024.

[90] Y. Mei, T. Nie, J. Sun, and Y. Tian, "Llm-attacker: Enhancing closed-loop adversarial scenario generation for autonomous driving with large language models," *preprint arXiv:2501.15850*, 2025.

[91] H. Tian, K. Reddy, Y. Feng, M. Quddus, Y. Demiris, and P. Angeloudis, "Enhancing autonomous vehicle training with language model integration and critical scenario generation," *preprint arXiv:2404.08570*, 2024.

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

[93] Y. Wei, Z. Wang, Y. Lu, C. Xu, C. Liu, H. Zhao *et al.*, "Editable scene simulation for autonomous driving via collaborative llm-agents," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[94] Baidu Apollo team, "Apollo: Open Source Autonomous Driving," https://github.com/ApolloAuto/apollo, 2017, accessed: 2019-02-11.

[95] National Highway Traffic Safety Administration, "National Motor Vehicle Crash Causation Survey (NMVCCS)," https://catalog.data.gov/dataset/national-motor-vehicle-crash-causation-survey-nmvccs, 2024.

[96] Ç. Güzay, E. Özdemir, and Y. Kara, "A generative ai-driven application: Use of large language models for traffic scenario generation," in *2023 14th International Conference on Electrical and Electronics Engineering (ELECO)*. IEEE, 2023.

[97] N. Petrovic, K. Lebioda, V. Zolfaghari, A. Schamschurko, S. Kirchner, N. Purschke *et al.*, "Llm-driven testing for autonomous driving scenarios," in *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*. IEEE, 2024.

[98] G. Rong, B. H. Shin, H. Tabatabaee, Q. Lu, S. Lemke, M. Možeiko *et al.*, "Lgsvl simulator: A high fidelity simulator for autonomous driving," in *2020 IEEE 23rd International conference on intelligent transportation systems (ITSC)*. IEEE, 2020.

[99] A. Guo, Y. Zhou, H. Tian, C. Fang, Y. Sun, W. Sun *et al.*, "Sovar: Build generalizable scenarios from accident reports for autonomous driving testing," in *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, 2024.

[100] X. Cai, X. Bai, Z. Cui, D. Xie, D. Fu, H. Yu *et al.*, "Text2scenario: Text-driven scenario generation for autonomous driving test," *preprint arXiv:2503.02911*, 2025.

[101] X. Zhou, Y. Huang, J. Zhang, J. Shao, D. Pan, and P. Li, "Automatic generation method for autonomous driving simulation scenarios based on large language model," in *International Conference on Artificial Intelligence and Autonomous Transportation*. Springer, 2024.

[102] L. Chen, O. Sinavski, J. Hünermann, A. Karnsund, A. J. Willmott, D. Birch *et al.*, "Driving with llms: Fusing object-level vector modality for explainable autonomous driving," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024.

[103] C. Schicktanz, L. Klitzke, K. Gimm, G. Rizzo, K. Liesner, H. H. Mosebach *et al.*, "The dlr urban traffic dataset (dlr-ut): A comprehensive traffic dataset from the aim research intersection," *TechRxiv*, 2025.

[104] X. Luo, C. Liu, F. Ding, F. Yang, Y. Zhou, J. Loo *et al.*, "Senserag: Constructing environmental knowledge bases with proactive querying for llm-based autonomous driving," in *Proceedings of the Winter Conference on Applications of Computer Vision*, 2025.

[105] A. Elhafsi, R. Sinha, C. Agia, E. Schmerling, I. A. Nesnas, and M. Pavone, "Semantic anomaly detection with large language models," *Autonomous Robots*, vol. 47, no. 8, 2023.

[106] C. Lu, T. Yue, and S. Ali, "Deepscenario: An open driving scenario dataset for autonomous driving system testing," in *2023 IEEE/ACM 20th International Conference on Mining Software Repositories (MSR)*. IEEE, 2023.

[107] J. Wu, C. Lu, A. Arrieta, T. Yue, and S. Ali, "Reality bites: Assessing the realism of driving scenarios with large language models," in *Proceedings of the 2024 IEEE/ACM First International Conference on AI Foundation Models and Software Engineering*, 2024.

[108] Y. Gao, M. Piccinini, K. Moller, A. Alanwar, and J. Betz, "From words to collisions: Llm-guided evaluation and adversarial generation of safety-critical driving scenarios," in *2025 IEEE 28th International Conference on Intelligent Transportation Systems (ITSC)*, 2025, accepted.

[109] S. You, X. Luo, X. Liang, J. Yu, C. Zheng, and J. Gong, "A comprehensive llm-powered framework for driving intelligence evaluation," *preprint arXiv:2503.05164*, 2025.

[110] O. Khattab, A. Singhvi, P. Maheshwari, Z. Zhang, K. Santhanam, S. Vardhamanan *et al.*, "Dspy: Compiling declarative language model calls into self-improving pipelines," 2024.

[111] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham *et al.*, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*. PMLR, 2021.

[112] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022.

[113] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, 2022.

[114] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *preprint arXiv:2304.10592*, 2023.

[115] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford *et al.*, "Zero-shot text-to-image generation," in *International conference on machine learning*. Pmlr, 2021.

[116] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023.

[117] Y. Inoue, Y. Yada, K. Tanahashi, and Y. Yamaguchi, "Nuscenes-mqa: Integrated evaluation of captions and qa for autonomous driving datasets using markup annotations," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

[118] S. Wang, Z. Yu, X. Jiang, S. Lan, M. Shi, N. Chang *et al.*, "Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning," *preprint arXiv:2405.01533*, 2024.

[119] M. Nie, R. Peng, C. Wang, X. Cai, J. Han, H. Xu *et al.*, "Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving," in *European Conference on Computer Vision*. Springer, 2024.

[120] A. Gopalkrishnan, R. Greer, and M. Trivedi, "Multi-frame, lightweight & efficient vision-language models for question answering in autonomous driving," *preprint arXiv:2403.19838*, 2024.

[121] Z. Sheng, Z. Huang, Y. Qu, Y. Leng, S. Bhavanam, and S. Chen, "Curricuvlm: Towards safe autonomous driving via personalized safety-critical curriculum learning with vision-language models," *preprint arXiv:2502.15119*, 2025.

[122] Q. Lu, X. Wang, Y. Jiang, G. Zhao, M. Ma, and S. Feng, "Multimodal large language model driven scenario testing for autonomous vehicles," *preprint arXiv:2409.06450*, 2024.

[123] W. Bao, Q. Yu, and Y. Kong, "Uncertainty-based traffic accident anticipation with spatio-temporal relational learning," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

[124] Y. Miao, G. Fainekos, B. Hoxha, H. Okamoto, D. Prokhorov, and S. Mitra, "From dashcam videos to driving simulations: Stress testing automated vehicles against rare events," *preprint arXiv:2411.16027*, 2024.

[125] A. Marathe, D. Ramanan, R. Walambe, and K. Kotecha, "Wedge: A multi-weather autonomous driving dataset built from generative vision-language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023.

[126] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu *et al.*, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[127] S. Luo, Y. Zhang, Y. Deng, and X. Zheng, "From accidents to insights: Leveraging multimodal data for scenario-driven ads testing," *preprint arXiv:2502.02025*, 2025.

[128] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman *et al.*, "Gpt-4 technical report," *preprint arXiv:2303.08774*, 2023.

[129] T. Choudhary, V. Dewangan, S. Chandhok, S. Priyadarshan, A. Jain, A. K. Singh *et al.*, "Talk2bev: Language-enhanced bird's-eye view maps for autonomous driving," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024.

[130] J. Mao, M. Niu, C. Jiang, H. Liang, J. Chen, X. Liang *et al.*, "One million scenes for autonomous driving: Once dataset," *preprint arXiv:2106.11037*, 2021.

[131] A. Ishaq, J. Lahoud, K. More, O. Thawakar, R. Thawkar, D. Dissanayake *et al.*, "Drivelmm-o1: A step-by-step reasoning dataset and large multimodal model for driving scenario understanding," *preprint arXiv:2503.10621*, 2025.

[132] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie *et al.*, "Drivelm: Driving with graph visual question answering," in *European Conference on Computer Vision*. Springer, 2024.

[133] A.-M. Marcu, L. Chen, J. Hünermann, A. Karnsund, B. Hanotte, P. Chidananda *et al.*, "Lingoqa: Visual question answering for autonomous driving," in *European Conference on Computer Vision*. Springer, 2024.

[134] T. Qian, J. Chen, L. Zhuo, Y. Jiao, and Y.-G. Jiang, "Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024.

[135] B. Khalili and A. W. Smyth, "Autodrive-qa-automated generation of multiple-choice questions for autonomous driving datasets using large vision-language models," *preprint arXiv:2503.15778*, 2025.

[136] M. Najibi, J. Ji, Y. Zhou, C. R. Qi, X. Yan, S. Ettinger *et al.*, "Unsupervised 3d perception with 2d vision-language distillation for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[137] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss *et al.*, "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.

[138] Y. Zhou, L. Cai, X. Cheng, Z. Gan, X. Xue, and W. Ding, "Openannotate3d: Open-vocabulary auto-labeling system for multi-modal 3d data," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024.

[139] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[140] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *Computer vision–ECCV 2008: 10th European conference on computer vision, marseille, France, October 12-18, 2008, proceedings, part i 10*. Springer, 2008.

[141] W.-B. Kou, Q. Lin, M. Tang, S. Wang, R. Ye, G. Zhu *et al.*, "Enhancing large vision model in street scene semantic understanding through leveraging posterior optimization trajectory," *preprint arXiv:2501.01710*, 2025.

[142] I. de Zarzà, J. de Curtò, G. Roig, and C. T. Calafate, "Llm multimodal traffic accident forecasting," *Sensors*, vol. 23, no. 22, 2023.

[143] H. Maeda, Y. Sekimoto, T. Seto, T. Kashiyama, and H. Omata, "Road damage detection using deep neural networks with images captured through a smartphone," *arXiv preprint arXiv:1801.09454*, 2018.

[144] K. Tong and S. Solmaz, "Connectgpt: Connect large language models with connected and automated vehicles," in *2024 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2024, pp. 581–588.

[145] X. Zheng, L. Wu, Z. Yan, Y. Tang, H. Zhao, C. Zhong *et al.*, "Large language models powered context-aware motion prediction in autonomous driving," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024.

[146] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu *et al.*, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[147] E. Rivera, J. Lübberstedt, N. Uhlemann, and M. Lienkamp, "Scenario understanding of traffic scenes through large visual language models," *preprint arXiv:2501.17131*, 2025.

[148] 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.

[149] L. Wen, X. Yang, D. Fu, X. Wang, P. Cai, X. Li *et al.*, "On the road with gpt-4v (ision): Early explorations of visual-language model on autonomous driving," *preprint arXiv:2311.05332*, 2023.

[150] X. Cao, T. Zhou, Y. Ma, W. Ye, C. Cui, K. Tang *et al.*, "Maplm: A real-world large-scale vision-language benchmark for map and traffic

scene understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[151] A. Keskar, S. Perisetla, and R. Greer, "Evaluating multimodal vision-language model prompting strategies for visual question answering in road scene understanding," in *Proceedings of the Winter Conference on Applications of Computer Vision (WACV) Workshops*, 2025.

[152] S. Xie, L. Kong, Y. Dong, C. Sima, W. Zhang, Q. A. Chen *et al.*, "Are vlms ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives," *preprint arXiv:2501.04003*, 2025.

[153] F. Li, H. Jin, B. Gao, L. Fan, L. Jiang, and L. Zeng, "Nugrounding: A multi-view 3d visual grounding framework in autonomous driving," *preprint arXiv:2503.22436*, 2025.

[154] K. Li, K. Chen, H. Wang, L. Hong, C. Ye, J. Han *et al.*, "Coda: A real-world road corner case dataset for object detection in autonomous driving," in *European Conference on Computer Vision*. Springer, 2022.

[155] K. Chen, Y. Li, W. Zhang, Y. Liu, P. Li, R. Gao *et al.*, "Automated evaluation of large vision-language models on self-driving corner cases," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025.

[156] Y. Wang, A. Alhuraish, S. Yuan, and H. Zhou, "Openlka: An open dataset of lane keeping assist from recent car models under real-world driving conditions," *preprint arXiv:2505.09092*, 2025.

[157] H. Hwang, S. Kwon, Y. Kim, and D. Kim, "Is it safe to cross? interpretable risk assessment with gpt-4v for safety-aware street crossing," in *2024 21st International Conference on Ubiquitous Robots (UR)*. IEEE, 2024.

[158] F.-H. Chan, Y.-T. Chen, Y. Xiang, and M. Sun, "Anticipating accidents in dashcam videos," in *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part IV 13*. Springer, 2017.

[159] J. Zhang, Y. Guan, C. Wang, H. Liao, G. Zhang, and Z. Li, "Latte: Lightweight attention-based traffic accident anticipation engine," *preprint arXiv:2504.04103*, 2025.

[160] M. P. Ronecker, M. Foutter, A. Elhafsi, D. Gammelli, I. Barakaiev, M. Pavone *et al.*, "Vision foundation model embedding-based semantic anomaly detection," *preprint arXiv:2505.07998*, 2025.

[161] Q. Zhang, M. Zhu, and H. F. Yang, "Think-driver: From driving-scene understanding to decision-making with vision language models," in *European Conference on Computer Vision Workshop*, 2024.

[162] J. Lee, J. Cho, H. Suk, and S. Kim, "SFF rendering-based uncertainty prediction using visionLLM," in *AAAI 2025 Workshop LM4Plan*, 2025. [Online]. Available: https://openreview.net/forum?id=q8ptjh1pDl

[163] D. Chen, Z. Zhang, Y. Liu, and X. T. Yang, "Insight: Enhancing autonomous driving safety through vision-language models on context-aware hazard detection and edge case evaluation," *preprint arXiv:2502.00262*, 2025.

[164] Y. Wang and H. Zhou, "Bridging human oversight and black-box driver assistance: Vision-language models for predictive alerting in lane keeping assist systems," *preprint arXiv:2505.11535*, 2025.

[165] D. Wu, W. Han, Y. Liu, T. Wang, C.-z. Xu, X. Zhang *et al.*, "Language prompt for autonomous driving," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 8, 2025.

[166] D. Wu, W. Han, T. Wang, X. Dong, X. Zhang, and J. Shen, "Referring multi-object tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023.

[167] W. Hong, W. Wang, Q. Lv, J. Xu, W. Yu, J. Ji *et al.*, "Cogagent: A visual language model for gui agents," *preprint arXiv:2312.08914*, 2024.

[168] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[169] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," *preprint arXiv:2306.02858*, 2023.

[170] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut *et al.*, "Gemini: a family of highly capable multimodal models," *preprint arXiv:2312.11805*, 2023.

[171] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai *et al.*, "Qwen2. 5-omni technical report," *arXiv preprint arXiv:2503.20215*, 2025.

[172] S.-Y. Park, C. Cui, Y. Ma, A. Moradipari, R. Gupta, K. Han *et al.*, "Nuplanqa: A large-scale dataset and benchmark for multi-view driving scene understanding in multi-modal large language models," *preprint arXiv:2503.12772*, 2025.

[173] X. Ding, J. Han, H. Xu, X. Liang, W. Zhang, and X. Li, "Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[174] A. Ishaq, J. Lahoud, F. S. Khan, S. Khan, H. Cholakkal, and R. M. Anwer, "Tracking meets large multimodal models for driving scenario understanding," *preprint arXiv:2503.14498*, 2025.

[175] S. Yang, J. Liu, R. Zhang, M. Pan, Z. Guo, X. Li *et al.*, "Lidar-llm: Exploring the potential of large language models for 3d lidar understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 9, 2025.

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

[177] S. Park, M. Lee, J. Kang, H. Choi, Y. Park, J. Cho *et al.*, "Vlaad: Vision and language assistant for autonomous driving," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

[178] X. Ding, J. Han, H. Xu, W. Zhang, and X. Li, "Hilm-d: Towards high-resolution understanding in multimodal large language models for autonomous driving," *preprint arXiv:2309.05186*, 2023.

[179] Y. Ma, Y. Cao, J. Sun, M. Pavone, and C. Xiao, "Dolphins: Multimodal language model for driving," in *European Conference on Computer Vision*. Springer, 2024.

[180] J. Fan, J. Wu, J. Gao, J. Yu, Y. Wang, H. Chu *et al.*, "Mllm-sul: Multimodal large language model for semantic scene understanding and localization in traffic scenarios," *preprint arXiv:2412.19406*, 2024.

[181] Y. Zhang and Y. Nie, "Interndrive: A multimodal large language model for autonomous driving scenario understanding," in *Proceedings of the 2024 4th International Conference on Artificial Intelligence, Automation and High Performance Computing*, 2024.

[182] P. Gao, J. Han, R. Zhang, Z. Lin, S. Geng, A. Zhou *et al.*, "Llama-adapter v2: Parameter-efficient visual instruction model," *preprint arXiv:2304.15010*, 2023.

[183] X. Zhou, K. Larintzakis, H. Guo, W. Zimmer, M. Liu, H. Cao *et al.*, "Tumtraffic-videoqa: A benchmark for unified spatio-temporal video understanding in traffic scenes," *preprint arXiv:2502.02449*, 2025.

[184] Q. Lu, M. Ma, X. Dai, X. Wang, and S. Feng, "Realistic corner case generation for autonomous vehicles with multimodal large language model," *preprint arXiv:2412.00243*, 2024.

[185] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, "Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[186] H. Tian, X. Han, G. Wu, Y. Zhou, S. Li, J. Wei *et al.*, "An llm-enhanced multi-objective evolutionary search for autonomous driving test scenario generation," *preprint arXiv:2406.10857*, 2024.

[187] S. Malla, C. Choi, I. Dwivedi, J. H. Choi, and J. Li, "Drama: Joint risk localization and captioning in driving," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023.

[188] J. M. Hankey, M. A. Perez, and J. A. McClafferty, "Description of the shrp 2 naturalistic database and the crash, near-crash, and baseline data sets," Virginia Tech Transportation Institute, Tech. Rep., 2016.

[189] T. Zeng, L. Wu, L. Shi, D. Zhou, and F. Guo, "Are vision llms road-ready? a comprehensive benchmark for safety-critical driving video understanding," *preprint arXiv:2504.14526*, 2025.

[190] H.-k. Chiu, R. Hachiuma, C.-Y. Wang, S. F. Smith, Y.-C. F. Wang, and M.-H. Chen, "V2v-llm: Vehicle-to-vehicle cooperative autonomous driving with multi-modal large language models," *arXiv preprint arXiv:2502.09980*, 2025.

[191] 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.

[192] S. Jain, S. Thapa, K.-T. Chen, A. L. Abbott, and A. Sarkar, "Semantic understanding of traffic scenes with large vision language models," in *2024 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2024.

[193] Q. Kong, Y. Kawana, R. Saini, A. Kumar, J. Pan, T. Gu *et al.*, "Wts: A pedestrian-centric traffic video dataset for fine-grained spatial-temporal understanding," in *European Conference on Computer Vision*. Springer, 2024.

[194] J. Lübberstedt, E. Rivera, N. Uhlemann, and M. Lienkamp, "V3lma: Visual 3d-enhanced language model for autonomous driving," *preprint arXiv:2505.00156*, 2025.

[195] T. Wang, S. Kim, J. Wenxuan, E. Xie, C. Ge, J. Chen *et al.*, "Deepaccident: A motion and accident prediction benchmark for v2x autonomous driving," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024.

[196] L. Wang, Y. Ren, H. Jiang, P. Cai, D. Fu, T. Wang *et al.*, "Accidentgpt: Accident analysis and prevention from v2x environmental perception with multi-modal large model," *preprint arXiv:2312.13156*, 2023.

[197] L. Shi, B. Jiang, and F. Guo, "Scvlm: a vision-language model for driving safety critical event understanding," *preprint arXiv:2410.00982*, 2024.

[198] M. Abu Tami, H. I. Ashqar, M. Elhenawy, S. Glaser, and A. Rakotonirainy, "Using multimodal large language models (mllms) for automated detection of traffic safety-critical events," *Vehicles*, vol. 6, no. 3, 2024.

[199] G. Elghazaly, R. Frank, S. Harvey, and S. Safko, "High-definition maps: Comprehensive survey, challenges, and future perspectives," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 4, pp. 527–550, 2023.

[200] H. Xiang, Z. Zheng, X. Xia, R. Xu, L. Gao, Z. Zhou *et al.*, "V2x-real: a largs-scale dataset for vehicle-to-everything cooperative perception," in *European Conference on Computer Vision*. Springer, 2024, pp. 455–470.

[201] R. Xu, X. Xia, J. Li, H. Li, S. Zhang, Z. Tu *et al.*, "V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 13 712–13 722.

[202] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, 2021.

[203] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew *et al.*, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *preprint arXiv:2112.10741*, 2021.

[204] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023.

[205] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023.

[206] P. Maul, M. Mueller, F. Enkler, E. Pigova, T. Fischer, and L. Stamatogiannakis, "Beamng.tech technical paper," BeamNG GmbH, Technical Report, 2023.

[207] Z. Zhong, D. Rempe, D. Xu, Y. Chen, S. Veer, T. Che *et al.*, "Guided conditional diffusion for controllable traffic simulation," in *2023 IEEE international conference on robotics and automation (ICRA)*, 2023.

[208] H. Lin, X. Huang, T. Phan-Minh, D. S. Hayden, H. Zhang, D. Zhao *et al.*, "Causal composition diffusion model for closed-loop traffic generation," *preprint arXiv:2412.17920*, 2024.

[209] C. Xu, D. Zhao, A. Sangiovanni-Vincentelli, and B. Li, "Diffscene: Diffusion-based safety-critical scenario generation for autonomous vehicles," in *The Second Workshop on New Frontiers in Adversarial Machine Learning*, 2023.

[210] J. Lu, S. Azam, G. Alcan, and V. Kyrki, "Data-driven diffusion models for enhancing safety in autonomous vehicle traffic simulations," *preprint arXiv:2410.04809*, 2024.

[211] Y. Xie, X. Guo, C. Wang, K. Liu, and L. Chen, "Advdiffuser: Generating adversarial safety-critical driving scenarios via guided diffusion," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024.

[212] W.-J. Chang, F. Pittaluga, M. Tomizuka, W. Zhan, and M. Chandraker, "Safe-sim: Safety-critical closed-loop traffic simulation with diffusion-controllable adversaries," in *European Conference on Computer Vision*. Springer, 2024.

[213] Z. Huang, Z. Zhang, A. Vaidya, Y. Chen, C. Lv, and J. F. Fisac, "Versatile behavior diffusion for generalized traffic agent simulation," *preprint arXiv:2404.02524*, 2024.

[214] Z. Zhong, D. Rempe, Y. Chen, B. Ivanovic, Y. Cao, D. Xu *et al.*, "Language-guided traffic simulation via scene-level diffusion," in *Conference on Robot Learning*. PMLR, 2023.

[215] M. Peng, Y. Xie, X. Guo, R. Yao, H. Yang, and J. Ma, "Ld-scene: Llm-guided diffusion for controllable generation of adversarial safety-critical driving scenarios," *preprint arXiv:2505.11247*, 2025.

[216] S. Zhang, J. Tian, Z. Zhu, S. Huang, J. Yang, and W. Zhang, "Drivegen: Towards infinite diverse traffic scenarios with large models," *preprint arXiv:2503.05808*, 2025.

[217] E. Pronovost, M. R. Ganesina, N. Hendy, Z. Wang, A. Morales, K. Wang *et al.*, "Scenario diffusion: Controllable driving scenario generation with diffusion," *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[218] M. Jiang, Y. Bai, A. Cornman, C. Davis, X. Huang, H. Jeon *et al.*, "Scenediffuser: Efficient and controllable driving simulation initialization and rollout," *Advances in Neural Information Processing Systems*, vol. 37, 2024.

[219] S. Sun, Z. Gu, T. Sun, J. Sun, C. Yuan, Y. Han *et al.*, "Drivescenegen: Generating diverse and realistic driving scenarios from scratch," *IEEE Robotics and Automation Letters*, 2024.

[220] K. Chitta, D. Dauner, and A. Geiger, "Sledge: Synthesizing driving environments with generative models and rule-based traffic," in *European Conference on Computer Vision*. Springer, 2024.

[221] L. Rowe, R. Girgis, A. Gosselin, L. Paull, C. Pal, and F. Heide, "Scenario dreamer: Vectorized latent diffusion for generating driving simulation environments," *preprint arXiv:2503.22496*, 2025.

[222] S. Yu, K. Kim, D. Kim, H. Han, and J. Lee, "Direct preference optimization-enhanced multi-guided diffusion model for traffic scenario generation," *preprint arXiv:2502.12178*, 2025.

[223] J. Zhou, L. Wang, Q. Meng, and X. Wang, "Diffroad: Realistic and diverse road scenario generation for autonomous vehicle testing," *preprint arXiv:2411.09451*, 2024.

[224] E. Pronovost, K. Wang, and N. Roy, "Generating driving scenes with diffusion," *preprint arXiv:2305.18452*, 2023.

[225] J. Lu, K. Wong, C. Zhang, S. Suo, and R. Urtasun, "Scenecontrol: Diffusion for controllable traffic scene generation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[226] S. Gu, J. Su, Y. Duan, X. Chen, J. Luo, and H. Zhao, "Text2street: Controllable text-to-image generation for street views," in *International Conference on Pattern Recognition*. Springer, 2025.

[227] K. Chen, E. Xie, Z. Chen, Y. Wang, L. Hong, Z. Li *et al.*, "Geodiffusion: Text-prompted geometric control for object detection data generation," *preprint arXiv:2306.04607*, 2023.

[228] K. Yang, E. Ma, J. Peng, Q. Guo, D. Lin, and K. Yu, "Bevcontrol: Accurately controlling street-view elements with multi-perspective consistency via bev sketch layout," *preprint arXiv:2308.01661*, 2023.

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

[230] H. Li, Z. Yang, Z. Qian, G. Zhao, Y. Huang, J. Yu *et al.*, "Dualdiff: Dual-branch diffusion model for autonomous driving with semantic fusion," *preprint arXiv:2505.01857*, 2025.

[231] Y. Wen, Y. Zhao, Y. Liu, F. Jia, Y. Wang, C. Luo *et al.*, "Panacea: Panoramic and controllable video generation for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[232] X. Li, Y. Zhang, and X. Y. Drivingdiffusion, "Layout-guided multi-view driving scene video generation with latent diffusion model," *preprint arXiv:2310.07771*, vol. 2, no. 3, 2023.

[233] J. Jiang, G. Hong, M. Zhang, H. Hu, K. Zhan, R. Shao *et al.*, "Dive: Efficient multi-view driving scenes generation based on video diffusion transformer," *preprint arXiv:2504.19614*, 2025.

[234] X. Bai, Y. Luo, L. Jiang, and S. Ostadabbas, "Dual-conditioned temporal diffusion modeling for long driving video generation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[235] Y. Fu, A. Jain, X. Chen, Z. Mo, and X. Di, "Drivegenvlm: Real-world video generation for vision language model based autonomous driving," in *2024 IEEE International Automated Vehicle Validation Conference (IAVVC)*. IEEE, 2024.

[236] Z. Yang, Z. Qian, X. Li, W. Xu, G. Zhao, R. Yu *et al.*, "Dualdiff+: Dual-branch diffusion for high-fidelity video generation with reward guidance," *preprint arXiv:2503.03689*, 2025.

[237] Y. Fu, Y. Li, and X. Di, "Gendds: Generating diverse driving video scenarios with prompt-to-video generative model," *preprint arXiv:2408.15868*, 2024.

[238] Y. Yao, X. Wang, M. Xu, Z. Pu, Y. Wang, E. Atkins *et al.*, "Dota: Unsupervised detection of traffic anomaly in driving videos," *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[239] Z. Guo, Y. Zhou, and C. Gou, "Drivinggen: Efficient safety-critical driving video generation with latent diffusion models," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 2024.

[240] J. Fang, L.-l. Li, J. Zhou, J. Xiao, H. Yu, C. Lv *et al.*, "Abductive ego-view accident video understanding for safe driving perception," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[241] C. Li, K. Zhou, T. Liu, Y. Wang, M. Zhuang, H.-a. Gao *et al.*, "Avd2: Accident video diffusion for accident video description," *preprint arXiv:2502.14801*, 2025.

[242] L. Baresi, D. Y. X. Hu, A. Stocco, and P. Tonella, "Efficient domain augmentation for autonomous driving testing using diffusion models," in *Proceedings of 47th International Conference on Software Engineering*, ser. ICSE '25. IEEE, 2025.

[243] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller *et al.*, "SDXL: Improving latent diffusion models for high-resolution image synthesis," in *The Twelfth International Conference on Learning Representations*, 2024.

[244] J. Mullan, D. Crawbuck, and A. Sastry, "Hotshot-xl," https://github.com/hotshotco/hotshot-xl, 2023, online.

[245] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing *et al.*, "Video generation models as world simulators," *OpenAI Blog*, vol. 1, 2024.

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

[247] K. L. Downing, "Predictive models in the brain," *Connection Science*, vol. 21, no. 1, 2009.

[248] H. Svensson, S. Thill, and T. Ziemke, "Dreaming of electric sheep? exploring the functions of dream-like mechanisms in the development of mental imagery simulations," *Adaptive Behavior*, 2013.

[249] A. Plebe, R. Donà, G. P. Papini Rosati, and M. Da Lio, "Mental imagery for intelligent vehicles," in *Proceedings of the 5th International Conference on Vehicle Technology and Intelligent Transport Systems - VEHITS*, INSTICC. SciTePress, 2019.

[250] A. Plebe and M. D. Lio, "On the road with 16 neurons: Towards interpretable and manipulable latent representations for visual predictions in driving scenarios," *IEEE Access*, vol. 8, 2020.

[251] K. Meyer and A. Damasio, "Convergence and divergence in a neural architecture for recognition and memory," *Trends in Neurosciences*, vol. 32, no. 7, 2009.

[252] K. Friston, "The free-energy principle: a unified brain theory?" *Nature Reviews Neuroscience*, vol. 11, no. 2, 2010.

[253] A. Plebe, H. Svensson, S. Mahmoud, and M. Da Lio, "Human-inspired autonomous driving: A survey," *Cognitive Systems Research*, 2024.

[254] A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall *et al.*, "Gaia-1: A generative world model for autonomous driving," *preprint arXiv:2309.17080*, 2023.

[255] F. Jia, W. Mao, Y. Liu, Y. Zhao, Y. Wen, C. Zhang *et al.*, "Adriver-i: A general world model for autonomous driving," *preprint arXiv:2311.13549*, 2023.

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

[257] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *preprint arXiv:2311.15127*, 2023.

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

[259] J. Yang, S. Gao, Y. Qiu, L. Chen, T. Li, B. Dai *et al.*, "Generalized predictive model for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[260] S. Gao, J. Yang, L. Chen, K. Chitta, Y. Qiu, A. Geiger *et al.*, "Vista: A generalizable driving world model with high fidelity and versatile controllability," *preprint arXiv:2405.17398*, 2024.

[261] X. Hu, W. Yin, M. Jia, J. Deng, X. Guo, Q. Zhang *et al.*, "Drivingworld: Constructingworld model for autonomous driving via video gpt," *preprint arXiv:2412.19505*, 2024.

[262] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler *et al.*, "Align your latents: High-resolution video synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023.

[263] 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.

[264] R. Gao, K. Chen, Z. Li, L. Hong, Z. Li, and Q. Xu, "Magicdrive3d: Controllable 3d generation for any-view rendering in street scenes," *preprint arXiv:2405.14475*, 2024.

[265] R. Gao, K. Chen, B. Xiao, L. Hong, Z. Li, and Q. Xu, "Magicdrivedit: High-resolution long video generation for autonomous driving with adaptive control," *preprint arXiv:2411.13807*, 2024.

[266] X. Tian, T. Jiang, L. Yun, Y. Mao, H. Yang, Y. Wang *et al.*, "Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving," *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[267] 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, 2024.

[268] C. Ni, G. Zhao, X. Wang, Z. Zhu, W. Qin, G. Huang *et al.*, "Recondreamer: Crafting world models for driving scene reconstruction via online restoration," *preprint arXiv:2411.19548*, 2024.

[269] N. Agarwal, A. Ali, M. Bala, Y. Balaji, E. Barker, T. Cai *et al.*, "Cosmos world foundation model platform for physical ai," *preprint arXiv:2501.03575*, 2025.

[270] H. A. Alhaija, J. Alvarez, M. Bala, T. Cai, T. Cao, L. Cha *et al.*, "Cosmos-transfer1: Conditional world generation with adaptive multimodal control," *preprint arXiv:2503.14492*, 2025.

[271] A. Chen, W. Zheng, Y. Wang, X. Zhang, K. Zhan, P. Jia *et al.*, "Geodrive: 3d geometry-informed driving world model with precise action control," *preprint arXiv:2505.22421*, 2025.

[272] L. Wang, W. Zheng, Y. Ren, H. Jiang, Z. Cui, H. Yu *et al.*, "Occsora: 4d occupancy generation models as world simulators for autonomous driving," *preprint arXiv:2405.20337*, 2024.

[273] J. Houston, G. Zuidhof, L. Bergamini, Y. Ye, L. Chen, A. Jain *et al.*, "One thousand and one hours: Self-driving motion prediction dataset," in *Conference on Robot Learning*. PMLR, 2021.

[274] Y. Yang, J. Mei, Y. Ma, S. Du, W. Chen, Y. Qian *et al.*, "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.

[275] X. Ma, Y. Wang, G. Jia, X. Chen, Z. Liu, Y.-F. Li *et al.*, "Latte: Latent diffusion transformer for video generation," *preprint arXiv:2401.03048*, 2024.

[276] S. Gu, W. Yin, B. Jin, X. Guo, J. Wang, H. Li *et al.*, "Dome: Taming diffusion model into high-fidelity controllable occupancy world model," *preprint arXiv:2410.10429*, 2024.

[277] Z. Yan, W. Dong, Y. Shao, Y. Lu, L. Haiyang, J. Liu *et al.*, "Renderworld: World model with self-supervised 3d label," *preprint arXiv:2409.11356*, 2024.

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

[279] O. Contributors, "Openscene: The largest up-to-date 3d occupancy prediction benchmark in autonomous driving," 2023.

[280] C. Min, D. Zhao, L. Xiao, J. Zhao, X. Xu, Z. Zhu *et al.*, "Driveworld: 4d pre-trained scene understanding via world models for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[281] Z. Wu, J. Ni, X. Wang, Y. Guo, R. Chen, L. Lu *et al.*, "Holodrive: Holistic 2d-3d multi-modal street scene generation for autonomous driving," *preprint arXiv:2412.01407*, 2024.

[282] Y. Zhang, S. Gong, K. Xiong, X. Ye, X. Tan, F. Wang *et al.*, "Bevworld: A multimodal world model for autonomous driving via unified bev latent space," *preprint arXiv:2407.05679*, 2024.

[283] 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.

[284] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras *et al.*, "Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[285] M. Hassan, S. Stapf, A. Rahimi, P. Rezende, Y. Haghighi, D. Brüggemann *et al.*, "Gem: A generalizable ego-vision multimodal

world model for fine-grained ego-motion, object dynamics, and scene composition control," *preprint arXiv:2412.11198*, 2024.

[286] H. Arai, K. Ishihara, T. Takahashi, and Y. Yamaguchi, "Act-bench: Towards action controllable world models for autonomous driving," *preprint arXiv:2412.05337*, 2024.

[287] X. Yang, L. Wen, Y. Ma, J. Mei, X. Li, T. Wei *et al.*, "Drivearena: A closed-loop generative simulation platform for autonomous driving," *preprint arXiv:2408.00415*, 2024.

[288] 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.

[289] T. Wiedemer, Y. Li, P. Vicol, S. S. Gu, N. Matarese, K. Swersky *et al.*, "Video models are zero-shot learners and reasoners," *arXiv preprint arXiv:2509.20328*, 2025.

[290] A. Hu, G. Corrado, N. Griffiths, Z. Murez, C. Gurau, H. Yeo *et al.*, "Model-based imitation learning for urban driving," *Advances in Neural Information Processing Systems*, vol. 35, 2022.

[291] P. J. Ball, J. Bauer, F. Belletti, B. Brownfield, A. Ephrat, S. Fruchter *et al.*, "Genie 3: A new frontier for world models," 2025.

[292] D. Dauner, M. Hallgarten, T. Li, X. Weng, Z. Huang, Z. Yang *et al.*, "Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking," *Advances in Neural Information Processing Systems*, vol. 37, 2024.

[293] H. G. Barrow, J. M. Tenenbaum, R. C. Bolles, and H. C. Wolf, "Parametric correspondence and chamfer matching: Two new techniques for image matching," in *Proceedings: Image Understanding Workshop*. Science Applications, Inc, 1977.

[294] D. Kim, S. Woo, J.-Y. Lee, and I. S. Kweon, "Video panoptic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[295] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[296] T. Unterthiner, S. Van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," *preprint arXiv:1812.01717*, 2018.

[297] H. Schafer, E. Santana, A. Haden, and R. Biasini, "A commute in data: The comma2k19 dataset," *preprint arXiv:1812.05752*, 2018.

[298] W. Tan, N. Qin, L. Ma, Y. Li, J. Du, G. Cai *et al.*, "Toronto-3d: A large-scale mobile lidar dataset for semantic segmentation of urban roadways," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2020.

[299] J. Geyer, Y. Kassahun, M. Mahmudi, X. Ricou, R. Durgesh, A. S. Chung *et al.*, "A2d2: Audi autonomous driving dataset," *preprint arXiv:2004.06320*, 2020.

[300] A. Kurup and J. P. Bos, "Winter adverse driving dataset (wads): year three," in *Autonomous Systems: Sensors, Processing and Security for Ground, Air, Sea and Space Vehicles and Infrastructure 2022*, M. C. Dudzik, T. J. Axenson, and S. M. Jameson, Eds. SPIE, 2022.

[301] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer *et al.*, "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020.

[302] J.-L. Deziel, P. Merriaux, F. Tremblay, D. Lessard, D. Plourde, J. Stanguennec *et al.*, "Pixset: An opportunity for 3d computer vision to go beyond point clouds with a full-waveform lidar dataset," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021.

[303] M. Alibeigi, W. Ljungbergh, A. Tonderski, G. Hess, A. Lilja, C. Lindström *et al.*, "Zenseact open dataset: A large-scale and diverse multimodal dataset for autonomous driving," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023.

[304] S. Dokania, A. H. A. Hafez, A. Subramanian, M. Chandraker, and C. Jawahar, "Idd-3d: Indian driving dataset for 3d unstructured road scenes," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2023.

[305] T. Sun, M. Segu, J. Postels, Y. Wang, L. Van Gool, B. Schiele *et al.*, "Shift: A synthetic driving dataset for continuous multi-task domain adaptation," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022.

[306] X. Zhang, L. Wang, J. Chen, C. Fang, G. Yang, Y. Wang *et al.*, "Dual radar: A multi-modal dataset with dual 4d radar for autononous driving," *Scientific Data*, vol. 12, no. 1, 2025.

[307] R. Xu, X. Xia, J. Li, H. Li, S. Zhang, Z. Tu *et al.*, "V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023.

[308] A. N. Ramesh, A. Correas-Serrano, and M. Gonzalez-Huici, "Scarl-a synthetic multi-modal dataset for autonomous driving," in *ICMIM 2024; 7th IEEE MTT Conference*, 2024.

[309] Y. Li, Z. Li, N. Chen, M. Gong, Z. Lyu, Z. Wang *et al.*, "Multiagent multitraversal multimodal self-driving: Open mars dataset," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024.

[310] J. Zürn, P. Gladkov, S. Dudas, F. Cotter, S. Toteva, J. Shotton *et al.*, "Wayvescenes101: A dataset and benchmark for novel view synthesis in autonomous driving," *preprint arXiv:2407.08280*, 2024.

[311] F. Fent, F. Kuttenreich, F. Ruch, F. Rizwin, S. Juergens, L. Lechermann *et al.*, "Man truckscenes: A multimodal dataset for autonomous trucking in diverse conditions," *Advances in Neural Information Processing Systems*, vol. 37, 2024.

[312] (2025) Automated Driving Toolbox. MathWorks. [Online]. Available: https://www.mathworks.com/products/automated-driving.html

[313] NVIDIA, "Nvidia drive sim," https://developer.nvidia.com/drive/simulation, 2019.

[314] A. Amini, T.-H. Wang, I. Gilitschenski, W. Schwarting, Z. Liu, S. Han *et al.*, "Vista 2.0: An open, data-driven simulator for multimodal sensing and policy learning for autonomous vehicles," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022.

[315] (2021) AWSIM: Autonomous Driving Simulator for Autoware. Tier IV Inc. Accessed: 2025-05-13. [Online]. Available: https://tier4.github.io/AWSIM/

[316] Q. Sun, X. Huang, B. C. Williams, and H. Zhao, "Intersim: Interactive traffic simulation via explicit relation modeling," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022.

[317] E. Vinitsky, N. Lichtlé, X. Yang, B. Amos, and J. Foerster, "Nocturne: a scalable driving benchmark for bringing multi-agent learning one step closer to the real world," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2022.

[318] C. Gulino, J. Fu, W. Luo, G. Tucker, E. Bronstein, Y. Lu *et al.*, "Waymax: an accelerated, data-driven simulator for large-scale autonomous driving research," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2023.

[319] D. Xu, Y. Chen, B. Ivanovic, and M. Pavone, "Bits: Bi-level imitation for traffic simulation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023.

[320] "CARLA Autonomous Driving Leaderboard," https://leaderboard.carla.org/, 2019.

[321] "DRL for Real ICCV 2025 Workshop," https://drl-for-real.github.io/DRL-for-Real/, 2025.

[322] (2025) Waymo open dataset challenges. Waymo LLC. [Online]. Available: https://waymo.com/open/challenges

[323] C. Davidson, D. Ramanan, and N. Peri, "Refav: Towards planning-centric scenario mining," *preprint arXiv:2505.20981*, 2025.

[324] P. Robicheaux, M. Popov, A. Madan, I. Robinson, J. Nelson, D. Ramanan *et al.*, "Roboflow100-vl: A multi-domain object detection benchmark for vision-language models," *preprint arXiv:2505.20612*, 2025.

[325] AVA Challenge Team. (2024) Accessibility vision and autonomy (ava) challenge. [Online]. Available: https://accessibility-cv.github.io/

[326] J. Kiseleva, A. Skrynnik, A. Zholus, S. Mohanty, N. Arabzadeh, M.-A. Côté *et al.*, "Interactive grounded language understanding in a collaborative environment: Retrospective on iglu 2022 competition," in *Proceedings of the NeurIPS 2022 Competitions Track*, ser. Proceedings of Machine Learning Research, M. Ciccone, G. Stolovitzky, and J. Albrecht, Eds., vol. 220. PMLR, 2022.

[327] M. Saroufim, Y. Perlitz, L. Choshen, L. Antiga, G. Bowyer, C. Puhrsch *et al.*, "Neurips 2023 llm efficiency fine-tuning competition," *preprint arXiv:2503.13507*, 2025.

[328] X. He, W. Feng, K. Zheng, Y. Lu, W. Zhu, J. Li *et al.*, "Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos," *preprint arXiv:2406.08407*, 2024.
[329] "3D Scene Understanding at CVPR 2025," https://scene-understanding.com/index.html, 2025.
[330] N. Maloyan, E. Verma, B. Nutfullin, and B. Ashinov, "Trojan detection in large language models: Insights from the trojan detection challenge," *preprint arXiv:2404.13660*, 2024.
[331] A. Cherian, K.-C. Peng, S. Lohit, K. A. Smith, and J. B. Tenenbaum, "Are deep neural networks smarter than second graders?" in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023.
[332] T. Kim, P. Ahn, S. Kim, S. Lee, M. Marsden, A. Sala *et al.*, "Nice: Cvpr 2023 challenge on zero-shot image captioning," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2024.
[333] K. Singh, T. Navaratnam, J. Holmer, S. Schaub-Meyer, and S. Roth, "Is synthetic data all we need? benchmarking the robustness of models trained with synthetic images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024.
[334] K. Yadav, J. Krantz, R. Ramrakhya, S. K. Ramakrishnan, J. Yang, A. Wang *et al.*, "Habitat Challenge 2023," https://aihabitat.org/challenge/2023/, 2023.
[335] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch *et al.*, "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models," *TRANSACTIONS ON MACHINE LEARNING RESEARCH*, 2022.
[336] M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung *et al.*, "Challenging big-bench tasks and whether chain-of-thought can solve them," *preprint arXiv:2210.09261*, 2022.
[337] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga *et al.*, "Holistic evaluation of language models," *preprint arXiv:2211.09110*, 2022.
[338] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao *et al.*, *MMBench: Is Your Multi-modal Model an All-Around Player?* Springer Nature Switzerland, 2024.
[339] X. Yue, Y. Ni, T. Zheng, K. Zhang, R. Liu, G. Zhang *et al.*, "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024.
[340] A. Myrzakhan, S. M. Bsharat, and Z. Shen, "Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena," *preprint arXiv:2406.07545*, 2024.
[341] Artificial Analysis. (2024) Text-to-image leaderboard. [Online]. Available: https://artificialanalysis.ai/text-to-image
[342] K. Grauman, A. Westbury, E. Byrne, V. Cartillier, Z. Chavis, A. Furnari *et al.*, "Ego4D: Around the World in 3,000 Hours of Egocentric Video," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2022.
[343] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman *et al.*, "Vizwiz grand challenge: Answering visual questions from blind people," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018.
[344] Shanghai AI Laboratory. (2023) Medfm: Foundation model prompting for medical image classification. [Online]. Available: https://www.shlab.org.cn/medfm
[345] M. E. Mostadi, H. Waeselynck, and J.-M. Gabriel, "Seven technical issues that may ruin your virtual tests for adas," in *2021 IEEE Intelligent Vehicles Symposium (IV)*, 2021.
[346] D. Karunakaran, J. S. Berrio Perez, and S. Worrall, "Generating edge cases for testing autonomous vehicles using real-world data," *Sensors*, vol. 24, no. 1, 2024. [Online]. Available: https://www.mdpi.com/1424-8220/24/1/108
[347] Q. Song, E. Engström, and P. Runeson, "Industry practices for challenging autonomous driving systems with critical scenarios," *ACM Trans. Softw. Eng. Methodol.*, vol. 33, no. 4, 2024.
[348] P. Mondorf and B. Plank, "Beyond accuracy: Evaluating the reasoning behavior of large language models–a survey," *preprint arXiv:2404.01869*, 2024.

Yuan Gao is a Ph.D. student at the Autonomous Vehicle Systems (AVS) lab at the Technical University of Munich (TUM). He received a B.Sc. in Mechanical Engineering from Hefei University of Technology in 2017 and two M.Sc. degrees from TUM: Mechatronics and Robotics and Development, Production, and Management in Mechanical Engineering in 2023. His research focuses on scenario generation and analysis for autonomous vehicles using FMs.

Mattia Piccinini is a TUM Global Post-doctoral Fellow at the Autonomous Vehicle Systems (AVS) lab at the Technical University of Munich (TUM). He received an M.Sc. in mechatronics engineering (cum laude) and a Ph.D. (cum laude) in autonomous systems from the University of Trento, Italy, in 2019 and 2024 respectively. He obtained the ITSS Best Dissertation Award and the Humboldt Post-doctoral Fellowship (2025). In 2022, he was a visiting Ph.D. student at the Universität der Bundeswehr, Munich, Germany. His research focuses on physics-guided motion generation and control of mobile ground robots.

Yuchen Zhang is a Ph.D. student at the Autonomous Vehicle Systems (AVS) lab at the Technical University of Munich (TUM). She holds a bachelor's degree in Mechanical Engineering from Shanghai University (2021) and a master's degree in Robotic Systems Engineering from RWTH Aachen University (2023). Since April 2024, she has been part of the AVS Lab, where her research focuses on perception systems for off-road vehicles.

Dingrui Wang is a PhD student at the Autonomous Vehicle Systems (AVS) Lab at the Technical University of Munich (TUM). He received his B.Sc. from Tianjin University in 2021 and M.Sc. from KU Leuven in 2022. His research focuses on investigating the applications of world models and end-to-end learning to develop reliable, data-driven systems capable of handling the complex decision-making processes required for autonomous navigation.

Korbinian Moller received a B.Sc. degree and an M.Sc. degree in mechanical engineering from the Technical University of Munich (TUM) in 2021 and 2023, respectively. He is currently pursuing a Ph.D. degree at the Autonomous Vehicle Systems (AVS) lab at TUM. His research interests include edge-case scenario simulation, the optimization of vehicle behavior, and motion planning in autonomous driving.

Roberto Brusnicki is a PhD student at the Autonomous Vehicle Systems (AVS) lab at the Technical University of Munich. He received his B.Sc. in 2017 and M.Sc. in 2022 from the Aeronautics Institute of Technology. His research focuses on enhancing autonomous vehicle performance using large language models for perceptual accuracy, scene understanding, and decision-making in ambiguous scenarios.

Baha Zarrouki is a PhD Researcher at the Autonomous Vehicle Systems (AVS) Lab at the Technical University of Munich (TUM). He received a B.Sc. and an M.Sc. in Electrical Engineering from TU Berlin in 2018 and 2020, respectively. His PhD research focused on fusing Deep Reinforcement Learning and Model Predictive Control for Nonlinear Motion Control of AV Systems. His current work explores World Model Predictive Control for autonomous driving.

Alessio Gambi is a Scientist at the Austrian Institute of Technology (AIT), Vienna, Austria. His research interests include automated software testing and analysis of complex and autonomous systems, simulation- and scenario-based testing, and test regression optimization. He is a recipient of a Facebook Testing and Verification Award (2019) and the Best Paper Award of the International Conference on Web Engineering (ICWE 2010). He serves on the program committees of flagship software engineering conferences (e.g., ASE, FSE, ISSTA, ICST) and reviews for top-tier journals (e.g., TSE, TOSEM, TAAS).

Jan Frederik Totz completed his PhD (Dr. rer. nat.) in Theoretical Physics at the Technical University of Berlin in 2017 on the topic of neuromorphic spiking neural networks. He received the Springer Thesis Award, the Chorafas prize, and the Carl-Ramsauer Award for his thesis. He continued with a postdoctoral position in the Departments of Mathematics and Mechanical Engineering at the Massachusetts Institute of Technology, funded by a Feodor Lynen Research Fellowship of the Humboldt Foundation. Currently, he serves as an R&D Engineer at Audi, specializing in Autonomous Driving.

Kai Storms studied Mechanical and Process Engineering at the Technical University of Darmstadt, where he received the M.Sc. in 2020 and completed his PhD (Dr.-Ing.) degree in February 2024. He is currently the chief engineer and vice head of the Institute of Automotive Engineering at the Technical University of Darmstadt, Germany. His topic was context-aware data reduction for highly automated driving. His research interests include the verification and validation of automated vehicles.

Steven Peters was born in 1987 and received his PhD (Dr.-Ing.) in 2013, at Karlsruhe Institute of Technology, Karlsruhe, Baden-Württemberg, Germany. From 2016 to 2022 he worked as Manager of Artificial Intelligence Research at Mercedes-Benz AG in Germany. He is a Full Professor at the Technical University of Darmstadt, Darmstadt, Germany and heads the Institute of Automotive Engineering in the Department of Mechanical Engineering since 2022.

Andrea Stocco is an Assistant Professor at the Technical University of Munich at the Chair of Software Engineering for Data-intensive Applications. He is also the head of the Automated Software Testing unit at fortiss. His research focuses on the interface between software engineering and deep learning with the goals of improving the robustness, reliability, and dependability of data-intensive software systems. He has received multiple best and distinguished paper awards at leading international conferences, including ICST (2025), QUATIC (2023), and ICWE (2016). He serves on the program committees of top-tier software engineering conferences (e.g., ICSE, FSE, ISSTA, ICST) and reviews for software engineering journals (e.g., TSE, TOSEM, EMSE, JSS, IST).

Bassam Alrifaee is a professor at the University of the Bundeswehr Munich, directs the Professorship for Adaptive Behavior of Autonomous Vehicles. Formerly at RWTH Aachen University, he founded the Cyber-Physical Mobility (CPM) group and the CPM Lab (2017–2024). He held a Visiting Scholar role at the University of Delaware in 2023. His research focuses on distributed control, service-oriented architectures, and connected and automated vehicles. Prof. Alrifaee secured grants and received awards for his advisory and editorial work. He holds Senior Member status at IEEE.

Marco Pavone is an Associate Professor of Aeronautics and Astronautics at Stanford University, where he directs the Autonomous Systems Laboratory and the Center for Automotive Research at Stanford. He is also a Distinguished Research Scientist at NVIDIA where he leads autonomous vehicle research. Before joining Stanford, he was a Research Technologist within the Robotics Section at the NASA Jet Propulsion Laboratory. He received a Ph.D. degree in Aeronautics and Astronautics from the Massachusetts Institute of Technology in 2010. His main research interests are in the development of methodologies for the analysis, design, and control of autonomous systems, with an emphasis on self-driving cars, autonomous aerospace vehicles, and future mobility systems. He is a recipient of a number of awards, including a Presidential Early Career Award for Scientists and Engineers from President Barack Obama.

Johannes Betz is an assistant professor in the Department of Mobility Systems Engineering at the Technical University of Munich (TUM), where he is leading the Autonomous Vehicle Systems (AVS) lab. He is one of the founders of the TUM Autonomous Motorsport team. His research focuses on developing adaptive dynamic path planning and control algorithms, decision-making algorithms that work under high uncertainty in multi-agent environments, and validating the algorithms on real-world robotic systems. Johannes earned a B.Eng. (2012) from the University of Applied Science Coburg, an M.Sc. (2012) from the University of Bayreuth, an M.A. (2021) in philosophy from TUM, and a Ph.D. (2019) from TUM.