# Literature Review

## Generative AI and Foundation Models in Automated Driving

**Kh Safkat Amin, M.Sc.**
RWTH Aachen University

Date: April 15, 2025

## Contents

## 1 Introduction

Automated driving systems (ADS) has rapidly evolved from rule-based systems into sophisticated AI-driven architectures capable of understanding and navigating complex environments. This transformation has been enabled by advances in sensing technologies, such as cameras, LiDAR, and radar, and the rise of deep learning models that interpret multimodal data in real time. At the core of modern driving stacks lies a multi-stage pipeline encompassing perception, prediction, planning, and control. Perception modules perform tasks like object detection, tracking, and scene understanding, which feed into downstream decision-making processes. The quality of these outputs directly impacts the safety and intelligence of autonomous behavior. This progression has defined what NVIDIA terms the *Perception AI* era, where machine learning systems interpret visual and textual inputs to support driving decisions [HUA25].

However, the field is now undergoing a paradigm shift, from *Perception AI* to *Generative AI*, and further into *Agentic AI* [HUA25]. Generative AI models not only interpret data but also synthesize high-fidelity images, text, and video, opening new opportunities in data augmentation, simulation, and contextual understanding. These capabilities are being accelerated by the rise of *Foundation Models*, large-scale, general-purpose models trained on massive, multimodal datasets. Unlike task-specific models that require tailored datasets and architectures, foundation models can generalize across a wide range of tasks with minimal fine-tuning. This transition from narrow, objective-specific training to broad, transferable intelligence is redefining the way ADS are designed, trained, and deployed.

The implications of generative AI and foundation models are far-reaching. In the offline development phase, these models are being used for automatic annotation, scenario simulation, and world modeling, dramatically reducing the effort required to create diverse and scalable training environments. In the on-vehicle context, foundation models are beginning to be deployed in real-time modules, offering new capabilities in perception, intent recognition, and decision-making. This marks the rise of Agentic AI, systems that exhibit reasoning, contextual understanding, planning, and goal-directed behavior akin to cognitive agents.

Agentic AI holds the potential to significantly enhance the robustness and generalization of ADS, particularly in handling rare or long-tail scenarios. Much like human drivers draw on a lifetime of diverse experiences to reason through unfamiliar situations, these AI agents aim to bring similar adaptability to machines. This is crucial for safe deployment in dynamic and unpredictable environments.

However, despite rapid progress, significant challenges remain. Generative and foundation models pose computational burdens, require vast data resources, and raise questions around safety, interpretability, and real-time deployment. This review explores how generative and agentic AI paradigms are reshaping the landscape of automated driving, focusing on both offline processes and on-vehicle deployment, highlighting recent advancements, emerging trends, and critical open questions shaping the future of intelligent driving systems.

## 2 Background and Advances

While modern ADS have become increasingly sophisticated, safety remains the primary bottleneck for large-scale deployment. One of the major limitations of current systems is their vulnerability to unseen or rare scenarios, situations that lie outside the distribution of training data. These edge cases can expose the brittle nature of specialized models, highlighting the need for systems that can reason under uncertainty, adapt in real time, and make context-aware decisions. Foundation models, with their inherent capacity for transfer learning, reasoning, and multimodal understanding, present a promising solution to bridge this gap.

In a recent webinar hosted by ADaS Lab, Marco Pavone, Director of the Autonomous Systems Lab at Stanford University and Research Scientist at NVIDIA, emphasized how AI is reshaping the development of ADS by dividing the influence of foundation models into two major paradigms [ADA25]: Offline processes, where AI enhances simulation, data generation, labeling, and pre-training. On-vehicle autonomy, where foundation models contribute directly to perception, planning, and control in real-time driving stacks.

The following sections explore recent progress within these two paradigms, with a focus on works highlighted in the webinar and other notable advancements in the field.

### 2.1 Offline Processing: Data-Driven Learning and Realistic Simulation

One of the most promising applications of generative AI and foundation models in offline ADS development lies in enhancing perception pipelines. Perception forms the backbone of any ADS, and improving it directly impacts downstream prediction and decision-making tasks. Traditionally, perception has been limited by expensive data acquisition, extensive manual labeling, and constrained augmentation techniques. Generative AI opens a new frontier by augmenting, repairing, and enriching perception data across multiple fronts.

#### 2.1.1 Enhancing Perception and Multimodal Sensor Understanding

The development of robust ADS hinges on two interconnected challenges: high-quality data annotation and multimodal sensor robustness. Generative AI and foundation models are revolutionizing both fronts by automating semantic understanding, enhancing sensor data quality, and enabling intelligent cross-modal fusion.

Manual labeling of multimodal sensor data remains a bottleneck, but vision-language models (VLMs) like *Contrastive Language-Image Pre-Training (CLIP)* [RAD21], *Large Language and Vision Assistant (LLaVA)* [LIU23], and *Generative Pre-trained Transformer-4 with vision (GPT-4V)* [OPE23] now enable scalable, context-aware auto-labeling. These models align visual and textual semantics to infer object classes, scene dynamics, and even road conditions with minimal supervision. For instance, CLIP's zero-shot capabilities allow it to assign labels to rare or novel objects in driving scenes (e.g., "construction barrier" or "debris"), while the *Segment Anything Model (SAM)* [KIR23] enables prompt-driven segmentation of regions described in natural language (e.g., "occluded pedestrian"). Extending this to 3D, Open-Vocabulary *SAM3D* [YAN23b] applies similar principles to LiDAR point clouds, segmenting and labeling objects based on textual prompts like "parked delivery van" or "icy

road section." Human-in-the-loop systems further refine these annotations, where generative AI proposes labels that experts verify or correct, reducing labeling costs by orders of magnitude while preserving accuracy.

*Generative Adversarial Networks (GANs)* have emerged as powerful tools for enhancing degraded visual data, particularly in tasks like super-resolution, deblurring, and inpainting. Unlike traditional methods that optimize for pixel-wise accuracy, GANs prioritize perceptual realism by training a generator-discriminator pair. This setup allows models to synthesize sharper, more natural-looking images, even if they exhibit lower PSNR values. Building on this, *Conditional Enhancement GAN (CE-GAN)* [JIA22] leverages complementary LiDAR input to restore corrupted camera images — such as those blurred by motion or obscured by mud-splattered lenses. By incorporating a tailored combination of loss functions (e.g., L1, structure-aware, and blurriness loss), CE-GAN mitigates over-smoothing and preserves fine texture details, outperforming conventional approaches under adverse conditions. A recent survey by Wei et al. [MO25] also summarizes enhancement techniques for in-vehicle vision systems, including motion deblurring, deraining, and dehazing. These findings highlight that generative models consistently outperform traditional techniques under various challenging conditions, reinforcing their growing role in real-world ADS. Beyond vision, generative models also enhance LiDAR and radar perception, enabling denoising and point cloud completion, transforming sparse 3D data into dense, coherent representations that improve scene understanding in low-visibility scenarios.

Furthermore, emerging multimodal generative models like *ImageBind* [GIR23] push multimodal integration further by learning a unified embedding space across 6 diverse modalities—including images, LiDAR, radar, audio and even text. This enables intelligent cross-modal reasoning: for instance, refining camera-based detections using LiDAR geometry, or contextualizing sparse radar signals with visual cues. Such fusion is particularly valuable in safety-critical scenarios—detecting a pedestrian occluded in the camera view but visible in the LiDAR scan, or recognizing that noisy radar reflections stem from heavy rain rather than solid obstacles. By aligning complementary sensor perspectives, these models enhance robustness and situational awareness in edge-case environments where single-modality systems tend to fail.

Emerging sensing modalities like 4D radar represent a breakthrough in automotive perception, overcoming many of the limitations of traditional radar systems. By capturing elevation and micro-Doppler information, 4D radar delivers richer motion cues and spatial awareness, enabling robust object detection at distances exceeding 300 meters—even in poor lighting or adverse weather where cameras and LiDAR struggle. This has sparked growing interest in the research community, especially as 4D radar begins to complement, and in some cases surpass, conventional sensors. While it can generate point clouds similar to LiDAR, these remain comparatively sparse. However, the raw radar outputs—high-dimensional tensors expressed in the frequency domain—carry nuanced information often imperceptible to human interpretation. Generative AI offers a powerful solution here, capable of translating these unstructured signals into dense, interpretable representations. By learning to interpolate missing data and enhance low-resolution radar frames, generative models unlock the hidden potential of radar, transforming it into a perception modality with both robustness and precision.

Collaborative perception via *Vehicle-to-Everything (V2X)* communication extends the reach of individual sensors, enabling vehicles to share real-time observations with each other and the infrastructure. By aggregating these distributed inputs, generative models can reconstruct occluded scenes—such as detecting a jaywalker behind a parked truck or visualizing a collapsed bridge based on roadside reports. Transformer-based models like *V2X-ViT* [XU22] fuse multi-vehicle perspectives into a unified view, while generative components help complete missing regions using learned priors. This improves perception beyond line-of-sight, especially in dense urban environments where occlusions are common.

Together, these developments form a generative perception pipeline: degraded or sparse sensor data is enhanced, semantics are added through auto-labeling, and complementary modalities are fused into coherent scene representations. This pipeline not only improves robustness in edge cases but also generates rich training data for planning and prediction tasks. A related trend is *low-carbon multimodal perception*—reducing dependence on power-hungry sensors like LiDAR or thermal cameras by using generative models to simulate or refine their outputs from lower-cost inputs [XU24]. This shift opens the door to energy-efficient, scalable perception systems that maintain performance while minimizing cost and hardware demands—critical for sustainable deployment at scale.

### 2.1.2 3D Scene Reconstruction and Holistic Simulation

Modern ADS are shifting from modular pipelines—where perception, prediction, and planning operate independently—towards end-to-end architectures that process fused multimodal data in a unified framework. This evolution demands a holistic approach to data generation, where entire driving scenes are simulated with realistic agent interactions, environmental complexity, and edge-case scenarios. Traditional real-world datasets, while valuable, fall short in capturing rare or dangerous events, making simulation a critical tool for robust validation and verification. Generative AI offers promising capabilities in this domain: GANs, diffusion models, and neural rendering techniques like *Neural Radiance Fields (NeRF)* and *Gaussian Splatting* can construct photorealistic 3D environments and dynamically simulate traffic participants—enabling diverse, controllable, and high-fidelity training and testing environments for ADS development.

One of the earlier breakthroughs in applying neural rendering to large-scale automated driving environments is *Block-NeRF* [TAN22], which extends traditional NeRFs beyond constrained, object-centric scenes. Designed for city-scale applications, Block-NeRF decomposes expansive urban environments into modular NeRF “blocks,” each trained independently and stitched together to render large, navigable spaces. This structure not only supports efficient scaling but also enables targeted updates—crucial for maintaining evolving maps without retraining from scratch. To ensure temporal and environmental consistency, Block-NeRF integrates appearance embeddings, exposure conditioning, and learned pose refinement, addressing variations in lighting, weather, and data collection across time. These capabilities make it especially useful for automated driving, offering a scalable method to build and maintain high-fidelity digital twins that support simulation, planning, and continuous validation.

While Block-NeRF effectively models static environments at large scales, it does not account for dynamic scene elements such as moving vehicles and pedestrians—key components in traffic simulation. To address this limitation, *EmerNeRF* [YAN23a] introduces a framework for reconstructing dynamic scenes from sparse visual data. EmerNeRF tackles the challenge of building 4D (space-time) representations by disentangling static and dynamic components within the scene. It self-supervises the reconstruction of dynamic elements without requiring dense ground-truth labels, leveraging 3D scene flow estimation to model how objects move over time. This makes EmerNeRF particularly suited for generating time-evolving scenarios that mirror real-world traffic dynamics. For ADS developers, EmerNeRF offers an advanced tool for simulating complex interactions and edge-case behaviors—such as sudden pedestrian crossings or rare vehicle maneuvers—thereby extending the reach and realism of simulation-based testing.

While powerful, EmerNeRF still faces challenges with completeness and real-time inference due to its per-scene optimization and limited motion modeling. *STORM* [YAN24] addresses these limitations by proposing a fully feed-forward, Transformer-based approach for real-time dynamic scene reconstruction. It introduces a novel mechanism for spatio-temporal aggregation: 3D Gaussians from multiple frames are aligned using self-supervised scene flow and fused in a single forward pass. This enables STORM to not only outperform existing methods in reconstructing complex, dynamic regions but also operate at speeds suitable for interactive applications—reconstructing large-scale outdoor scenes in  200ms. Moreover, STORM learns richer motion representations, allowing high-fidelity mask generation and enhanced dynamic simulation from sparse observations.

NeRFs and Gaussian Splatting have also been used to reposition road users within 3D reconstructed scenes, enhancing automated driving testing by enabling fine-grained scenario manipulation. Instead of regenerating entire environments, engineers can interactively modify specific actors—such as changing a pedestrian's path or relocating a vehicle to simulate near-miss situations—while preserving photorealistic context and scene integrity. This allows for targeted generation of edge cases and counterfactual scenarios, which are crucial for stress-testing perception, prediction, and planning modules. Moreover, the ability to alter traffic participants within real-world reconstructions bridges the gap between simulation and reality, offering a more scalable and controllable framework for evaluating automated systems in diverse, safety-critical situations.

A prominent example of this approach is *NeuroNCAP* [LJU25], a NeRF-based simulator that enables closed-loop safety testing by dynamically editing real-world driving scenes—such as relocating vehicles or altering trajectories—to synthesize collisions or near-miss scenarios. The system leverages a trained NeRF to render photorealistic camera inputs from modified scenes, which are then processed by the automated vehicle's planner to generate trajectories. These trajectories are converted into control commands (e.g., steering, acceleration) via a *linear quadratic regulator (LQR)* controller and executed using a kinematic bicycle model with real-world constraints (e.g., EuroNCAP-compliant braking limits). By simulating the vehicle's full reactive pipeline in manipulated but realistic environments, NeuroNCAP efficiently evaluates planner robustness for edge cases like stationary obstructions or lateral collisions, bridging the gap between simulation and real-world data while avoiding costly physical testing.

These manipulated scenes are invaluable for *out-of-distribution (OOD)* testing. For instance, after training a trajectory prediction model on existing datasets, novel scenes generated using generative AI can evaluate the model's generalization capabilities, testing its behavior in unseen situations without requiring new real-world data collection. However, a current limitation is that generated agents do not interact dynamically—other agents are merely replayed and do not respond to changes in the manipulated agent, creating unrealistic scenarios where vehicles ignore newly inserted obstacles or the ego vehicle's maneuvers. *Ghost Gym* [WAY23] addresses this gap by enabling closed-loop evaluation with reactive agents in counterfactual simulations. By simulating interactive behaviors grounded in real-world dynamics, Ghost Gym enhances the credibility of manipulated scenes, facilitating more reliable testing of automated driving stacks.

Additionally, modifying scenes often introduces visual artifacts like violations of physical laws, unrealistic shadows, floating objects, or texture distortions. These limitations can be addressed through hybrid approaches combining neural rendering with game engines, where physics-based simulators handle reactive agent behaviors and collision avoidance while the neural component maintains visual fidelity, with the physics engine also serving to detect and correct rendering inconsistencies. One such initiative is *PRISM-1* [WAY24a], which focuses on creating high-fidelity 3D reconstructions from real-world scenes. These reconstructions are designed to maintain physical plausibility and can be integrated into simulations like Ghost Gym to test ADS under realistic conditions

To move beyond handcrafted rules and avoid physically implausible artifacts in simulation, recent advancements in generative AI have explored learning the dynamics of the real world directly from data. These approaches aim to build world models—neural models that capture the structure, causality, and physics of an environment in a way that enables autonomous agents to reason, plan, and simulate interactions realistically. Rather than relying solely on static reconstructions or replayed trajectories, world models learn how things move and respond, modeling both agents and environments in a cohesive framework. A prime example of this direction is *Generative AI for Autonomy (GAIA)-1* [HU23]. Trained on large-scale real-world driving data collected from British cities, GAIA-1 leverages a combination of VQ-GAN for spatial encoding, an autoregressive transformer for temporal dynamics, and a diffusion-based video decoder to generate high-fidelity, physically plausible future driving scenes. By operating on reconstructions produced by PRISM-1, GAIA-1 enables realistic manipulation of objects and actors within the scene to synthesize novel or edge-case scenarios.

What sets GAIA-1 apart is its ability to generate diverse, multimodal driving scenarios from a short sequence of input video, action, and text. For instance, it can simulate a vehicle being forced onto a sidewalk, weather transitions, or novel traffic compositions involving unseen combinations of vehicles and pedestrians—behaviors that are risky to record in real life and scarce in existing datasets. These simulations are not mere replays of prior footage; instead, they are imaginative extrapolations grounded in GAIA-1's learned understanding of traffic rules, infrastructure layouts, and physical context. When combined with PRISM-1's high-fidelity 3D reconstructions, GAIA-1 can manipulate scene elements to generate new or edge-case scenarios within physically realistic environments, enabling more nuanced simulation and testing. It also demonstrates counterfactual reasoning, such as imagining what would happen if a vehicle didn't yield or if a pedestrian crossed unexpectedly. Moreover,

its ability to preserve coherent agent behavior and spatial geometry over time allows for robust scenario generation that can be used to stress-test ADS under complex, rare, and interactive conditions. This marks a significant advancement toward intelligent, data-driven simulation frameworks for AV development.

In addition to GAIA-1, several other generative models have emerged that extend the concept of world modeling for automated driving. *DriveDreamer* [WAN24], for example, differs from GAIA-1 by incorporating richer input modalities such as HD maps and 3D bounding boxes, enabling finer control and more accurate spatial understanding during scenario generation. While GAIA-1 focuses on generative imagination based on real-world driving data, DriveDreamer emphasizes precision and structured control over predicted futures, offering enhanced utility for decision-making systems. Beyond these, [GUA24] highlighted many other models that explore the world modeling paradigm, by integrating elements like language, point clouds, 3D occupancy grids, and LIDAR data. These advances underscore a growing trend toward multi-modal, interpretable, and physics-aware world models that can simulate complex and diverse driving scenarios with increasing realism and fidelity

One of the exciting advancements in world modeling for automated driving is the capability for prompt-based scene generation. These models can now create detailed traffic scenarios based on high-level textual inputs, such as natural language prompts or crash reports. This approach allows for the generation of complex, specific scenes in response to user queries, making it possible to simulate rare or edge-case scenarios without the need for extensive real-world data collection. *LCTGen* [TAN23] is a notable example of this development, which leverages crash reports to generate traffic scenes from textual descriptions. By using pre-trained *large language models (LLMs)*, LCTGen sidesteps the need for costly paired language-traffic datasets and offers a highly flexible way to create critical and diverse traffic scenarios. This ability to generate scenes based on text not only enhances scenario testing for safety validation but also paves the way for more human-centric and scalable simulation frameworks in ADS development.

Another powerful capability of these world models is novel view synthesis, where the models can generate new perspectives or viewpoints of a given scene. This is particularly valuable for ADS, as it allows for the simulation of traffic scenarios from multiple angles without needing to capture all possible views through real-world sensors or cameras. By generating different viewpoints, these models can simulate how an automated vehicle would perceive its environment from various positions, helping to evaluate sensor performance, validate perception algorithms, and test decision-making strategies under diverse visual contexts. This ability to create synthetic, high-fidelity visual perspectives in a controlled, scalable manner significantly accelerates the development of ADS by providing rich, varied datasets for training and testing, without the logistical and safety challenges of real-world data collection.

World models play a crucial role not only in scenario generation but also in learning and refining driving behaviors, planning, and control strategies for ADS. For example, *Model-based Imitation Learning (MILE)* [HU22] enhances driving performance by predicting future environments based on offline datasets, improving driving scores and enabling vehicles to operate without HD maps. Similarly, SEM2 builds upon world models by addressing challenges like task-irrelevant information and

data imbalance, leading to better generalization and handling of unexpected situations in driving scenarios. Multi-view modeling, as seen in Drive-WM, adds another layer by ensuring consistent predictions across multiple camera views, enhancing safety and performance in end-to-end automated driving. Other models, like UniWorld, push boundaries by utilizing point cloud fusion for more accurate environmental understanding. Additionally, models like TrafficBots emphasize agent behavior prediction, enabling more scalable and efficient planning. These advancements demonstrate the growing sophistication of world models in improving the performance, robustness, and safety of automated vehicles in diverse and dynamic environments.

## 2.2 On-Vehicle Execution: Real-Time Reasoning and Decision-Making on the Road

While the previous section focused on how foundation models support the offline development of ADS, through simulation, data augmentation, and 3D reconstruction, this section explores their growing role in online development, where real-time decision-making, adaptability, and robustness are crucial. Online development workflows focus on embedding intelligence directly into the vehicle's operational pipeline, making it capable of handling dynamic driving environments. This shift raises the question: how far can foundation models go in addressing the challenges of real-time autonomy?

Foundation models, trained on internet-scale multimodal data, offer a unique advantage: exposure to an immense variety of scenarios, including edge cases that conventional systems often fail to handle. This vast prior knowledge enables reasoning about novel or ambiguous situations. As a result, researchers are exploring how such models can enhance the perception, prediction, and planning stack of ADS—paving the way for holistic, interpretable, and human-like driving intelligence.

One of the most prominent developments in this direction is *Dolphins* [MA23], a VLM tailored for automated driving. Dolphins addresses a key limitation of current AV systems: their inability to holistically understand complex scenes, adapt rapidly to unfamiliar scenarios, and recover from errors in real time. By integrating multimodal inputs—video, text, and past control signals—Dolphins performs step-by-step reasoning through a mechanism called Grounded Chain of Thought (GCoT). Instruction-tuned on driving-specific datasets like *BDD-X* [KIM18], the model handles perception, control prediction, and dialogue comprehension within a unified framework. Its in-context learning and reflective capabilities allow it to dynamically adapt to new conditions and correct its behavior. However, inference latency and computational demands still present significant hurdles for real-time deployment.

Another compelling framework is language agent develped by [MAO24], which leverages LLMs as cognitive agents within an AV system. Unlike traditional pipelines that strictly segment perception, prediction, and planning, Agent-Driver unifies these modules under a reasoning-driven architecture. It uses a Tool Library to selectively invoke perception modules, a Cognitive Memory to recall traffic rules and past experiences, and a Reasoning Engine that performs chain-of-thought inference and hierarchical planning. Remarkably, Agent-Driver achieves significant performance gains with minimal training data, demonstrating robust few-shot learning capabilities. Its design also emphasizes interpretability, with decision logs explaining not just what actions are taken, but why. The primary

trade-off lies in the inference overhead of LLMs, making real-time responsiveness a challenge without further optimization.

While these models focus on language-driven reasoning, another approach emphasizes multimodal integration. *LINGO-2* [WAY24b], developed by Wayve, marks a milestone as the first closed-loop *vision-language-action model (VLAM)* deployed on public roads. It enhances traditional end-to-end driving systems by enabling the vehicle to generate natural language explanations alongside driving commands. This dual-output system not only improves transparency but also allows for interaction via language prompts (e.g., “turn left,” “slow down”)—a feature that brings user-in-the-loop control closer to reality. Tested in simulation and real-world environments, LINGO-2 demonstrates the ability to answer real-time queries and adapt behavior accordingly. Despite this, aligning language outputs with physical actions remains an open challenge, requiring improvements in grounding and consistency to avoid unsafe behaviors.

Beyond language-driven intelligence, architectural innovation is also reshaping how online development workflows are designed. *PARA-Drive* [WEN24] proposes a fully parallelized end-to-end AV architecture that co-trains modules for perception, prediction, planning, and mapping using a shared bird’s-eye-view representation. By removing inter-module dependencies, PARA-Drive eliminates redundant computation and achieves near 3× speedups compared to sequential designs. Its modularity also allows for component deactivation at runtime—enabling more efficient inference without significant accuracy loss. While PARA-Drive excels in open-loop evaluations, its real-time capabilities in closed-loop and real-world contexts are areas for future research.

Together, these advances signal a paradigm shift in the online development of automated vehicles. Traditional modular pipelines remain important for interpretability and safety auditing, but integrating foundation models into these systems introduces flexibility, generalization, and adaptability previously unattainable. Whether through language-based reasoning, multimodal integration, or architectural redesign, foundation models are redefining what it means to “drive” in real time. Challenges such as inference speed, energy consumption, and data alignment persist, but the trajectory is clear: online development workflows are increasingly becoming AI-native, driven by models that reason, reflect, and learn—much like humans do behind the wheel.

The use of AI in real-time, safety-critical scenarios such as automated driving inevitably raises questions about reliability, robustness, and trust. Ensuring safe behavior under uncertainty and rare edge cases remains a core challenge. At the same time, AI—particularly large-scale foundation models—can also play a crucial role in improving safety by offering richer semantic understanding of complex environments. One promising direction is the use of LLMs for semantic anomaly detection, where failures arise not from unfamiliar inputs but from unusual combinations of familiar elements. In the paper [ELH23], the authors introduce a zero-shot framework that transforms visual inputs into textual descriptions and uses LLMs to reason about contextual inconsistencies through chain-of-thought prompting. Applied to automated driving scenarios in the CARLA simulator, the approach outperforms traditional out-of-distribution detection methods by capturing subtle semantic mismatches, such as traffic lights mounted on trucks. These results highlight the potential of foundation models to serve as an additional semantic layer for detecting safety-relevant anomalies, paving the way toward more context-aware and trustworthy AI systems.

Foundation models, particularly LLMs, are highly effective in anomaly detection, offering contextual information and zero-shot reasoning that can identify out-of-distribution failure modes in ADS. However, their computational expense makes real-time implementation challenging, especially for agile robots with limited computational resources. *AESOP* [SIN24] addresses this by mimicking human behavior during anomalous situations—just as we slow down and focus more deeply when encountering something strange while driving. AESOP creates a cache of LLM embeddings from nominal scenes collected during training. At test time, when the robot encounters a new scene, it retrieves the full embedding from a vector database and performs further analysis to detect anomalies. This two-stage framework—combining a fast anomaly detection phase with a slower, deeper reasoning process—enables safe and efficient operation, ensuring that the system can handle unexpected situations while maintaining safety and performance in real-time.

## 3 Conclusion

Foundation models are transforming the landscape of ADS, offering a significant leap in the generalization capabilities required for fully automated vehicle development. By leveraging vast amounts of data and cutting-edge technology, these models facilitate advancements in both offline and real-time processing of ADS. The potential for unification and specialization within these models is immense, allowing not only for the replacement of existing pipelines but also for their enhancement. This opens the door to parallel architectures that enable fast inference and low latency, revolutionizing how ADS are developed and deployed.

However, challenges remain in realizing the full potential of these models. The computational cost of high-dimensional data, such as video, and issues related to physical embodiment—such as noisy robots and uncontrolled environments—continue to pose significant obstacles. Additionally, ensuring the safety and ethical deployment of these technologies is critical for their responsible use.

Despite these hurdles, embodied AI platforms are already demonstrating their potential, and ongoing research is actively addressing these concerns. High-quality, large-scale data, ranging from diverse driving scenarios to embodied and internet-scale videos, is crucial for advancing these systems. Foundation models learn from both driving and non-driving experiences in ways similar to humans, enhancing their ability to adapt and generalize across various environments.

With continued innovation, research, and an expanding open-source ecosystem, foundation models are well-positioned to revolutionize ADS, enhance safety, and drive the development of intelligent, adaptable AI platforms capable of meeting the demands of the future.

## 4 Abbreviations

| | |
|---|---|
| ADS | automated driving systems |
| GAN | generative adversarial network |
| LiDAR | light detection and ranging |
| LLM | large language model |
| RADAR | radio detection and ranging |
| V2X | vehicle-to-everything |

## 5 References

[ADA25] ADAS LAB
Webinar on Foundation Models for Autonomous Driving
https://www.linkedin.com/events/webinaronfoundationmodelsforaut7307414676617732096/comments/, Accessed: 2025-04-01, Apr. 2025

[ELH23] ELHAFSI, A., SINHA, R., AGIA, C., SCHMERLING, E., NESNAS, I. A. D., PAVONE, M.
Semantic anomaly detection with large language models
Autonomous Robots 47.8 (2023), pp. 1035–1055, ISSN: 1573-7527, DOI: 10.1007/s10514-023-10132-6, URL: https://doi.org/10.1007/s10514-023-10132-6

[GIR23] GIRDHAR, R., EL-NOUBY, A., LIU, Z., SINGH, M., ALWALA, K. V., JOULIN, A., MISRA, I.
ImageBind: One Embedding Space To Bind Them All
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 15180–15190

[GUA24] GUAN, Y., LIAO, H., LI, Z., HU, J., YUAN, R., ZHANG, G., XU, C.
World Models for Autonomous Driving: An Initial Survey
IEEE Transactions on Intelligent Vehicles (2024), pp. 1–17, DOI: 10.1109/TIV.2024.3398357

[HU22] HU, A., CORRADO, G., GRIFFITHS, N., MUREZ, Z., GURAU, C., YEO, H., KENDALL, A., CIPOLLA, R., SHOTTON, J.
Model-Based Imitation Learning for Urban Driving
Advances in Neural Information Processing Systems, ed. by KOYEJO, S., MOHAMED, S., AGARWAL, A., BELGRAVE, D., CHO, K., OH, A., vol. 35, Curran Associates, Inc., 2022, pp. 20703–20716, URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/827cb489449ea216e4a257c47e407d18-Paper-Conference.pdf

[HU23] HU, A., RUSSELL, L., YEO, H., MUREZ, Z., FEDOSEEV, G., KENDALL, A., SHOTTON, J., CORRADO, G.
GAIA-1: A Generative World Model for Autonomous Driving
2023, arXiv: 2309.17080 [cs.CV], URL: https://arxiv.org/abs/2309.17080

[HUA25] HUANG, J.
NVIDIA Keynote at CES 2025
https://www.ces.tech/videos/2025/january/nvidia-keynote/, Accessed: 2025-04-10, Jan. 2025

[JIA22] JIANG, S., GUO, Z., ZHAO, S., WANG, H., JING, W.
CE-GAN : A Camera Image Enhancement Generative Adversarial Network for Autonomous Driving
2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA), 2022, pp. 1–6, DOI: 10.1109/DSAA54385.2022.10032427

[KIM18] KIM, J., ROHRBACH, A., DARRELL, T., CANNY, J., AKATA, Z.
Textual Explanations for Self-Driving Vehicles
Proceedings of the European Conference on Computer Vision (ECCV) (2018)

[KIR23] KIRILLOV, A., MINTUN, E., RAVI, N., MAO, H., ROLLAND, C., GUSTAFSON, L., XIAO, T., WHITEHEAD, S., BERG, A. C., LO, W.-Y., DOLLÁR, P., GIRSHICK, R.
Segment Anything
2023, arXiv: 2304.02643 [cs.CV], URL: https://arxiv.org/abs/2304.02643

[LIU23] LIU, H., LI, C., WU, Q., LEE, Y. J.
Visual Instruction Tuning
2023, arXiv: 2304.08485 [cs.CV], URL: https://arxiv.org/abs/2304.08485

[LJU25] LJUNGBERGH, W., TONDERSKI, A., JOHNANDER, J., CAESAR, H., ÅSTRÖM, K., FELSBERG, M., PETERSSON, C.
NeuroNCAP: Photorealistic Closed-Loop Safety Testing for Autonomous Driving
Computer Vision – ECCV 2024, ed. by LEONARDIS, A., RICCI, E., ROTH, S., RUSSAKOVSKY, O., SATTLER, T., VAROL, G., Cham: Springer Nature Switzerland, 2025, pp. 161–177, ISBN: 978-3-031-73404-5

[MA23] MA, Y., CAO, Y., SUN, J., PAVONE, M., XIAO, C.
Dolphins: Multimodal Language Model for Driving
2023, arXiv: 2312.00438 [cs.CV], URL: https://arxiv.org/abs/2312.00438

[MAO24] MAO, J., YE, J., QIAN, Y., PAVONE, M., WANG, Y.
A Language Agent for Autonomous Driving
2024, arXiv: 2311.10813 [cs.CV], URL: https://arxiv.org/abs/2311.10813

[MO25] MO, T., ZHENG, S., CHAN, W.-Y., YANG, R.
Review of AI Image Enhancement Techniques for In-Vehicle Vision Systems Under Adverse Weather Conditions
World Electric Vehicle Journal 16.2 (2025), ISSN: 2032-6653, DOI: 10.3390/wevj16020072, URL: https://www.mdpi.com/2032-6653/16/2/72

[OPE23] OPENAI
GPT-4 with Vision (GPT-4V)
Accessed: 2025-04, 2023, URL: https://openai.com/research/gpt-4v-system-card

[RAD21] RADFORD, A., KIM, J. W., HALLACY, C., RAMESH, A., GOH, G., AGARWAL, S., SASTRY, G., ASKELL, A., MISHKIN, P., CLARK, J., KRUEGER, G., SUTSKEVER, I.
Learning Transferable Visual Models From Natural Language Supervision
2021, arXiv: 2103.00020 [cs.CV], URL: https://arxiv.org/abs/2103.00020

[SIN24] SINHA, R., ELHAFSI, A., AGIA, C., FOUTTER, M., SCHMERLING, E., PAVONE, M.
Real-Time Anomaly Detection and Reactive Planning with Large Language Models
2024, arXiv: 2407.08735 [cs.RO], URL: https://arxiv.org/abs/2407.08735

[TAN23] TAN, S., IVANOVIC, B., WENG, X., PAVONE, M., KRAEHENBUEHL, P.
Language Conditioned Traffic Generation
2023, arXiv: 2307.07947 [cs.CV], URL: https://arxiv.org/abs/2307.07947

[TAN22] TANCIK, M., CASSER, V., YAN, X., PRADHAN, S., MILDENHALL, B., SRINIVASAN, P. P., BARRON, J. T., KRETZSCHMAR, H.
Block-NeRF: Scalable Large Scene Neural View Synthesis
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2022, pp. 8248–8258

[WAN24] WANG, X., ZHU, Z., HUANG, G., CHEN, X., ZHU, J., LU, J.
DriveDreamer: Towards Real-World-Drive World Models for Autonomous Driving
European Conference on Computer Vision, Springer, 2024, pp. 55–72

[WAY23] WAYVE
Ghost Gym: A Neural Simulator for Autonomous Driving
Accessed: 2025-04-14, Dec. 2023, URL: https://wayve.ai/thinking/ghost-gym-neural-simulator/

[WAY24a] WAYVE
Introducing PRISM-1: Photorealistic reconstruction in static and dynamic scenes
Accessed: 2025-04-14, June 2024, URL: https://wayve.ai/thinking/prism-1/

[WAY24b] WAYVE
LINGO-2: Driving with Natural Language
Accessed: 2025-04-14, Wayve, Apr. 2024, URL: https://wayve.ai/thinking/lingo-2-driving-with-language/

[WEN24] WENG, X., IVANOVIC, B., WANG, Y., WANG, Y., PAVONE, M.
PARA-Drive: Parallelized Architecture for Real-time Autonomous Driving
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2024, pp. 15449–15458

[XU24] XU, M., NIYATO, D., KANG, J., XIONG, Z., JAMALIPOUR, A., FANG, Y., KIM, D. I., XUEMIN, SHEN
Integration of Mixture of Experts and Multimodal Generative AI in Internet of Vehicles: A Survey
2024, arXiv: 2404.16356 [cs.NI], URL: https://arxiv.org/abs/2404.16356

[XU22] XU, R., XIANG, H., TU, Z., XIA, X., YANG, M.-H., MA, J.
V2X-ViT: Vehicle-to-Everything Cooperative Perception with Vision Transformer
Computer Vision – ECCV 2022, ed. by AVIDAN, S., BROSTOW, G., CISSÉ, M., FARINELLA, G. M., HASSNER, T., Cham: Springer Nature Switzerland, 2022, pp. 107–124, ISBN: 978-3-031-19842-7

[YAN24] YANG, J., HUANG, J., CHEN, Y., WANG, Y., LI, B., YOU, Y., SHARMA, A., IGL, M., KARKUS, P., XU, D., IVANOVIC, B., WANG, Y., PAVONE, M.
STORM: Spatio-Temporal Reconstruction Model for Large-Scale Outdoor Scenes
2024, arXiv: 2501.00602 [cs.CV], URL: https://arxiv.org/abs/2501.00602

[YAN23a] YANG, J., IVANOVIC, B., LITANY, O., WENG, X., KIM, S. W., LI, B., CHE, T., XU, D., FIDLER, S., PAVONE, M., WANG, Y.
EmerNeRF: Emergent Spatial-Temporal Scene Decomposition via Self-Supervision
arXiv preprint arXiv:2311.02077 (2023)

[YAN23b] YANG, Y., WU, X., HE, T., ZHAO, H., LIU, X.
SAM3D: Segment Anything in 3D Scenes
2023, arXiv: 2306.03908 [cs.CV], URL: https://arxiv.org/abs/2306.03908