Exploring the Integration of Agentic AI with Multimodal Foundation Models in Autonomous Driving

J.Dhana Santhosh Reddy
M.Eng Robotics
UID: 120405570
UMD College Park
js5162@umd.edu

Chandhan Saai Katuri
M.Eng Robotics
UID: 120387364
UMD College Park
chandhan@umd.edu

Naga Kambhampati
M.Eng Robotics
UID: 119484128
UMD College Park
saigopal@umd.edu

Abstract

This paper investigates the integration of agentic AI with multimodal foundation models to advance autonomous driving. Agentic AI systems, distinguished by their proactive decision-making and adaptive reasoning, are combined with multimodal foundation models capable of processing and fusing diverse sensor inputs including vision, LiDAR, and radar to achieve comprehensive environmental understanding. This synergy enhances core autonomous driving functions such as perception, prediction, and planning, enabling vehicles to navigate complex, real-world environments with minimal human intervention. We highlight recent progress in multimodal data integration, discuss sophisticated decision-making frameworks, and explore strategies to ensure robustness against out-of-distribution (OOD) scenarios. Additionally, we examine emerging challenges, address safety and reliability considerations, and propose directions for future research and development. By unifying agentic intelligence with multimodal foundation models, this work provides a blueprint for more reliable, adaptable, and ultimately safer autonomous driving systems.

1 Introduction

Autonomous driving (AD) systems have made significant progress through advancements in machine learning, sensor technologies, and data fusion techniques[4, 19]. Yet, the inherently dynamic and unpredictable nature of real-world roadways continues to challenge these systems, especially when encountering long-tail events, adverse weather, and complex traffic interactions. Such conditions demand more than just incremental improvements to existing modular pipelines they require a paradigm shift toward more adaptive, intelligent, and context-aware solutions.

This paper explores the integration of agentic AI with multimodal foundation models to address these challenges [1, 2, 3]. Agentic AI, characterized by proactive, goal-oriented decision-making and adaptive reasoning [1], complements multimodal foundation models that process and fuse diverse inputs such as RGB images, LiDAR point clouds, radar signals, and textual priors into a holistic understanding of the driving environment [5].

In addition to introducing this integrative framework, this survey examines state-of-the-art techniques, highlights recent achievements, identifies open problems, and outlines a research roadmap[8]. Our intent is not only to demonstrate the potential of combining agentic AI and multimodal foundation models but also to provide researchers, practitioners, and policymakers with a clearer perspective on the broader implications and future directions for advancing autonomous driving technologies.

This survey highlights recent achievements, identifies open problems, and outlines a roadmap for advancing autonomous driving technologies.

Motivation

Agentic AI, as discussed by Rick Spair [1], represents a transformative approach to artificial intelligence, characterized by autonomous decision-making and adaptability in dynamic environments. Its goal-driven behavior addresses key challenges in applications like autonomous vehicles and smart infrastructure. By combining reinforcement learning and multi-agent coordination, agentic AI enables systems to effectively navigate complex, uncertain scenarios with minimal human intervention. This concept inspired my survey, highlighting its potential to revolutionize autonomous intelligence for reliable, adaptable performance in real-world challenges. [1].

2 Related Work

The integration of agentic AI with multimodal models has been explored in several recent works, focusing on improving perception, decision-making, and planning in autonomous driving systems.

2.1 Multimodal Foundation Models in Autonomous Driving

Several multimodal models have advanced autonomous driving capabilities:

EMMA (End-to-End Multimodal Model for Autonomous Driving): EMMA utilizes a multimodal large language model (MLLM) foundation to process raw sensor inputs, such as camera data, into driving-specific outputs like planner trajectories, object detections, and road graph elements[14]. By leveraging natural language representations for inputs and outputs, EMMA demonstrates strong performance in motion planning and 3D object detection. However, EMMA's reliance on language representations introduces computational overhead and limits its applicability to scenarios with limited labeled data such as KITTI and Waymo Datasets.[18].

KOMA (Knowledge-Driven Multi-Agent Framework for Autonomous Driving): KOMA addresses dynamic environments by combining visual perception with language-driven reasoning [12]. While EMMA [14] establishes a strong foundation for multimodal reasoning in autonomous driving, its limitations in handling dynamic multi-agent scenarios become the focus of KOMA [12] The integration of multi-step planning and shared memory enables cooperative decision-making among agents, particularly in interaction-heavy scenarios like intersections. While KOMA's approach enhances robustness and coordination, challenges remain in scaling the framework to real-world deployments, especially under time-critical constraints.

RoboFusion: RoboFusion is designed to enhance 3D object detection using visual foundation models (VFMs) like the Segment Anything Model (SAM) [2]. KOMA's emphasis on coordination in dynamic environments necessitates robust perception capabilities, which RoboFusion [2] addresses. It introduces SAM-AD, which adapts SAM for autonomous driving, and innovative components like wavelet decomposition for noise reduction and adaptive fusion for feature refinement. RoboFusion excels in handling adverse weather and noisy conditions, but the complexity of integrating VFMs in real-time systems may hinder scalability.

The self-attention mechanism, central to RoboFusion's feature refinement process, is mathematically defined as:

$$F' = \text{softmax}(W_Q F \cdot (W_K F)^T) W_V F, \tag{1}$$

where F represents the fused feature, W_Q, W_K, W_V are learnable parameters, and F' is the refined feature. This formulation enables the model to emphasize informative components while suppressing irrelevant noise. RoboFusion demonstrates significant resilience in out-of-distribution (OOD) scenarios.

2.2 AI with Multimodal Perception for Self-Driving Cars

The research on multimodal sensor fusion [19] further strengthens the perceptual foundation of this integrated model. Subsequent studies exploring multimodal fusion strategies[5, 17, 13] demonstrated that early fusion of RGB and depth (e.g., LiDAR) data [6] outperforms later fusion strategies due to improved spatial alignment.Beyond improved spatial alignment, early fusion facilitates joint feature learning from the outset, allowing the network to exploit complementary cues between modalities (such as texture from RGB and geometry from depth) before they are abstracted into separate feature spaces. This early integration reduces the risk of losing fine-grained relational information that often dissipates when modalities are fused at later stages.By optimizing early fusion strategies for RGB images and depth data, this research contributes to a more comprehensive and accurate understanding of the driving environment, benefiting both KOMA's multi-agent [12] coordination and RoboFusion's object detection [2] . Using the CARLA simulator and Conditional Imitation Learning (CIL), the research optimizes control signals u_t, emphasizing early fusion strategies. The vehicle control output u_t is modeled as:

$$u_t = f_\theta(x_t, d_t, n_t), \tag{2}$$

Early fusion integrates RGB and depth data at the input level:

$$F_{\text{fused}} = \text{concat}(x_t, d_t), \tag{3}$$

where F_{fused} is processed through shared layers to predict vehicle control signals. Early fusion outperforms mid and late fusion strategies due to better spatial alignment of RGB and depth features in early layers and achieves a 76.76 % drift reduction in CARLA simulations [13] .Critically, these gains in perception and control fidelity support the overarching objectives of robust, adaptable autonomous vehicles.

2.3 Combining Deep Learning and Decision-Making AI for Autonomous Navigation

This research combines Monte Carlo Tree Search (MCTS) [20] and Deep Reinforcement Learning (DRL) [3] to improve decision-making, extending AlphaGo Zero-style models for continuous action spaces. Although MCTS+DRL frameworks improve long-horizon planning and exploration, they often incur significant computational costs. This complexity may limit real-time deployment unless combined with efficient sampling strategies or approximate search methods. Furthermore, reward shaping in dense traffic conditions remains a non-trivial challenge, potentially leading to suboptimal policies if not carefully tuned. The decision-making problem is formulated as a Partially Observable Markov Decision Process (POMDP) [25], where high-level actions (e.g., lane change) are optimized using an adaptive cruise control system. The optimal policy $\pi^*(s)$ is derived as:

$$\pi^*(s) = \arg\max_a Q(s, a), \tag{4}$$

where $Q(s, a)$ represents the value function, updated iteratively:

$$Q(s, a) \leftarrow \frac{\sum_{i=1}^{N} R_i(s, a)}{N(s, a)}. \tag{5}$$

Integrating MCTS with DRL improves exploration efficiency and decision-making under dynamic conditions.

2.4 Multi-Agent Reinforcement Learning for Autonomous Driving

Multi-agent reinforcement learning (MARL) addresses the complexity of coordinating multiple autonomous vehicles by enabling cooperation, competition, and communication among agents [24]. Approaches such as centralized training with decentralized execution (e.g., MADDPG [27]) enhance policy learning, while independent policy optimization (IPO) [9] supports scalability in large-scale traffic environments. However, MARL still contends with partial observability, non-stationarity, and credit assignment challenges.

Incorporating transformer-based communication can effectively handle diverse agent behaviors, building upon methods like KOMA [12], and thereby moving closer to real-world deployment. Although MARL can improve scalability and efficiency, training large populations of agents remains computationally intensive, raising feasibility concerns, particularly in heterogeneous networks.

Formally, for each agent i, the Bellman equation for the state-action value function $Q_i(s, a)$ is:

$$Q_i(s, a) = \mathbb{E}\big[R_i + \gamma \max_{a'} Q_i(s', a')\big]. \tag{6}$$

Recent trends integrate MARL with multimodal foundation models and agentic AI to achieve proactive, context-aware decision-making, real-time distributed execution, socially-aware coordination, and robust perception under out-of-distribution conditions [12].

2.5 Decision-Making AI for Self-Driving Cars

The research on Partially Observable Markov Decision Processes [25] provides a framework for handling uncertainty in dynamic traffic environments. Integrating POMDP-based reasoning into the model complements MARL's [24] coordination strategies, enabling robust decision-making under uncertainty and contributing to the overall safety and reliability of the system [10] .Although POMDP formulations elegantly capture uncertainty, exact inference over complex road scenarios may be computationally intractable. Real-world systems must rely on approximations or sampling-based methods, potentially sacrificing optimality for tractability. Balancing these computational constraints against the benefits of uncertainty-awareness remains a key research challenge. A POMDP is represented as a tuple:

$$M = (S, A, O, T, Z, R, \gamma), \tag{7}$$

Optimization: The policy π maximizes the expected cumulative reward:

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right]. \tag{8}$$

Learning-based approaches using POMDPs handle uncertainty more effectively than classical decision-making systems [20].

2.6 Multimodal Transformer for End-to-End Autonomous Driving

The TransFuser model [17] offers valuable insights into optimizing multimodal integration strategies. Its exploration of early, deep, and late fusion provides guidance for effectively combining the diverse sensor inputs and language representations used in the integrated model [5] .

Early fusion combines raw sensor inputs (X_1, X_2) at the initial stage, expressed as $F = f_{\text{early}}(X_1, X_2)$. This approach enables joint feature learning from the start, leveraging richer contextual information [13] . However, it is sensitive to sensor misalignment and synchronization errors.

Deep fusion integrates modality-specific features at intermediate layers of the neural network, defined as $F = h(f(X_1), g(X_2))$. This strategy balances modularity and accuracy but requires careful alignment of intermediate features, making implementation complex [19].

Late fusion aggregates outputs from independently processed modalities at the decision level, formulated as $F = f_{\text{late}}(H_1, H_2)$. While this approach provides modularity and robustness to sensor failures, it often underperforms early and deep fusion due to its inability to exploit complementary low-level information [4].

Although TransFuser suggests no one-size-fits-all solution, early fusion may be ideal for stable sensor setups. TransFuser achieves superior performance by leveraging global context and optimizing the spatial alignment of multimodal inputs. Early fusion provides **higher accuracy**, while late fusion ensures **modularity and robustness**. The choice of strategy depends on task requirements and sensor configurations.whereas deep fusion provides nuanced feature interactions at the cost of higher model complexity. Late fusion's robustness to sensor failures trades off against less efficient feature synergy.

2.7 VLP: Vision-Language Planning for Autonomous Driving

The Vision-Language Planning (VLP) [15] framework enhances autonomous driving systems (ADS) by integrating large language models (LLMs) with Bird's Eye View (BEV) features for semantic reasoning and motion planning. It comprises two key components: the Agent-centric Learning

Paradigm (ALP) and the Self-driving-car-centric Learning Paradigm (SLP), which together refine ADS reasoning and decision-making [14].

Planning Query Optimization: The planning query Q is optimized using BEV features F_{BEV} and goal information G_{goal}:

$$Q = \text{LM}(F_{\text{BEV}}, G_{\text{goal}}), \tag{9}$$

where LM aligns visual features with high-level semantic goals, ensuring contextual understanding in the ego-vehicle planning process [15].

Semantic Reasoning and Planning: By integrating LLMs, VLP improves the semantic representation of BEV features, enabling [3]:

- Comprehension of driving environments with human-like semantic interpretations.
- Effective generalization to unseen scenarios, such as new urban environments.
- Reduced collision rates through refined contextual motion planning.

ALP enhances BEV features by aligning them with expected semantic descriptions using contrastive learning. SLP optimizes ego-vehicle queries by linking them to goal-oriented linguistic features.

VLP significantly reduces average L2 error and collision rates, It demonstrates robust generalization to diverse scenarios and achieves an intuitive integration of vision and language, making ADS more reliable and adaptable.

3 Proposed System Architecture

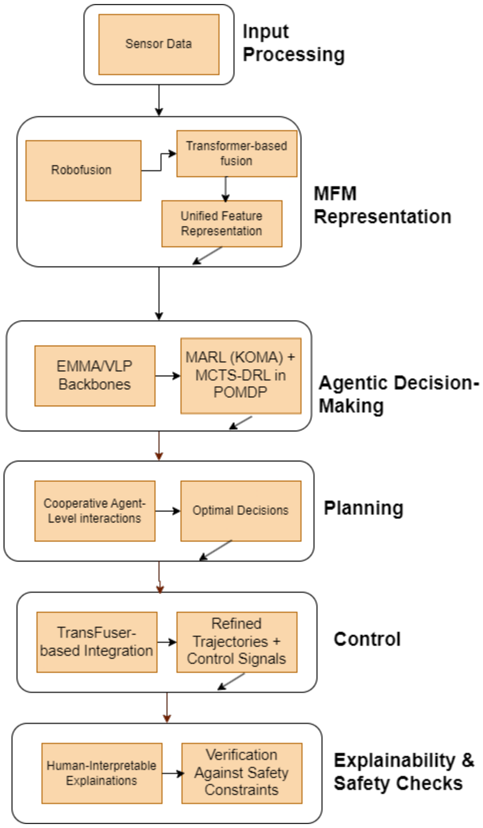
Sensor Data
Input Processing
Robofusion
Transformer-based fusion
Unified Feature Representation
MFM Representation
EMMA/VLP Backbones
MARL (KOMA) + MCTS-DRL in POMDP
Agentic Decision-Making
Cooperative Agent-Level interactions
Optimal Decisions
Planning
TransFuser-based Integration
Refined Trajectories + Control Signals
Control
Human-Interpretable Explainations
Verification Against Safety Constraints
Explainability & Safety Checks

Figure 1: System Architecture for Proposed model

The proposed architecture comprises the following key components [4, 19], organized according to their role in the overall data processing and decision-making pipeline: The data flow within the architecture can be summarized as follows:

- **Input Processing:** Sensor data → RoboFusion & Transformer-based fusion → Unified feature representation [2].
- **MFM Representation:** Unified features → EMMA/VLP backbones for semantic parsing → Environment states, language-aligned semantics, and preliminary trajectory hints [14, 15].
- **Agentic Decision-Making:** Semantic environment state + MARL (KOMA) → Cooperative agent-level interactions [12]. MCTS+DRL within POMDP → Optimal high-level decisions under uncertainty [20].
- **Planning & Control:** High-level decisions + TransFuser-based integration → Refined trajectories and control signals [17].
- **Output:** Safe, context-aware, and interpretable driving actions (control signals to actuators) [8].
- **Explainability & Safety Checks:** Final decisions → Human-interpretable explanations, verification against safety constraints [16].

3.1 Multimodal Sensor Input and Preprocessing

Sensor Suite The system utilizes a comprehensive sensor suite, including RGB cameras, LiDAR point clouds, radar signals, GPS, and V2X communication [19]. This diverse set of sensors provides redundant and complementary information about the driving environment, ensuring robustness to individual sensor limitations and varying environmental conditions. This is an improvement on using a lesser number of sensors or sensors which are all of the same type.

RoboFusion with VFMs (e.g., SAM-AD) Raw sensor data is pre-processed using a RoboFusion-like component that leverages Visual Foundation Models (VFMs) such as SAM-AD [2]. VFMs, trained on vast datasets, enhance data quality by mitigating noise, handling out-of-distribution (OOD) scenarios, and performing semantic segmentation. This approach offers significant advantages over traditional computer vision models, which often degrade under adverse conditions or sensor imperfections. The use of VFMs ensures a more reliable and adaptable perception layer, crucial for real-world deployment.

Early/Deep Multimodal Fusion The pre-processed data from each sensor modality is then integrated using either an early or deep fusion strategy [13]. Early fusion combines raw sensor data, while deep fusion integrates features extracted from each modality [17]. Unlike late fusion, which merges outputs at the decision level, early/deep fusion captures inter-modality correlations more effectively. This results in improved spatial alignment and increased robustness to noise. This approach allows the system to exploit the synergistic relationship between different sensor modalities.

3.2 Multimodal Feature Representation and Semantic Understanding

The fused sensor data is further processed using a Transformer-based fusion mechanism [17, 5]. The self-attention mechanism in Transformers excels at modeling long-range dependencies and aligning complex multi-channel inputs. Compared to traditional fusion techniques (e.g., concatenation, averaging), Transformer-based fusion selectively weights informative features across modalities, leading to a more coherent and informative representation [4]. Multimodal Foundation Model (MFM) Backbone, this unified representation is then fed into an MFM backbone, specifically employing models like EMMA and VLP [14, 15].

EMMA for Semantic Understanding : EMMA processes the fused sensor data alongside natural language instructions to generate structured, semantically rich environment representations [18]. Unlike unimodal or purely vision-based models, EMMA's language grounding provides a more human-like, interpretable understanding of the scene, facilitating explainability and ensuring that downstream modules have access to a semantically meaningful representation.

VLP for Goal-Directed Planning : VLP integrates vision-language planning capabilities, bridging the gap between raw perception and high-level semantic goals [15]. VLP enables the system to reason with linguistic instructions, improving adaptability to new tasks and complex driving rules. This is a significant advancement over traditional vision-only planners, which are limited in their ability to understand semantic goals or contextual cues [7].

3.3 Agentic Decision-Making and Reasoning

Multi-Agent Reasoning and Coordination (KOMA + MARL) This layer employs the KOMA framework combined with MARL strategies to handle multi-agent interactions and ensure social compliance [12, 24]. Unlike single-agent reinforcement learning, MARL introduces communication and cooperation protocols, enabling vehicles to navigate complex traffic scenarios safely and efficiently. Transformer-based communication further enhances the scalability and coherence of multi-agent interactions [9].

3.4 Decision-Making Under Uncertainty (MCTS + DRL within POMDP)

To address the challenges of partial observability and uncertainty, the architecture incorporates MCTS for search-based planning and DRL for adaptive policy learning within a POMDP framework [20, 25]. This hybrid approach outperforms purely learned policies by efficiently exploring decision spaces and improving performance in uncertain, partially observable scenarios [10]. Compared to heuristic-based planners, Monte Carlo Tree Search (MCTS) + Deep Reinforcement Learning (DRL) adapts to novel circumstances and continuously refines policies through ongoing experience.

3.5 Planning and Control

High-Level Planning and Trajectory Generation (VLP + TransFuser) : This layer leverages VLP's language-driven reasoning [15] and TransFuser's robust multimodal integration [17] to generate contextually aware, human-aligned trajectories. This approach surpasses classical planners that rely on hand-crafted features and rules, leading to more flexible and intuitive driving behaviors [3].

Trajectory Refinement:: The generated trajectories are iteratively refined based on feedback from the Agentic AI layer and multimodal cues, ensuring optimal adaptation to dynamic conditions [7].

Low-Level Control Execution : The refined trajectories are translated into precise control signals for the vehicle's actuators [13]. The MFM backbone's semantically rich and noise-robust features contribute to reliable control even under challenging conditions, unlike conventional systems that may degrade sharply with sensor anomalies [10].

3.6 Explainability and Safety

XAI Tools and Natural Language Explanations: The system incorporates XAI tools, leveraging EMMA and KOMA's language-driven reasoning to provide human-interpretable explanations for its decisions [8]. This is crucial for user acceptance, regulatory compliance, and debugging [16]. Unlike opaque end-to-end deep networks, the proposed architecture inherently produces structured, interpretable semantic representations that enable more transparent and meaningful explanations [11].

Safety & Ethical Constraints: Rule-based safety checks and monitors are integrated alongside learned policies to ensure compliance with traffic laws, ethical guidelines, and fail-safe protocols [16, 10].

4 Safety & Ethical Constraints

4.1 Safety & Ethical Constraints

System reliability and robustness in autonomous systems is critical to ensure consistent performance. Multimodal foundation models, such as RoboFusion and TransFuser, are essential for achieving enhanced robustness in rare and out-of-distribution (OOD) scenarios, particularly in adverse conditions like extreme weather [2, 17]. These models must reliably process noisy environments and mitigate uncertainties through aleatoric and epistemic prediction frameworks to enhance planning accuracy in dynamic settings [20, 10].

Decision-making and transparency play a vital role in gaining user trust. Explainable AI (XAI)[8] tools provide interpretable decisions, enabling human operators to understand and validate system reasoning [8]. This transparency not only ensures trust but also addresses potential safety concerns. By understanding the decision-making process, public acceptance of autonomous systems can improve significantly. Furthermore, real-time verification methods and techniques like Chain-of-Thought prompting help mitigate hallucinations in generative and decision-making AI systems, ensuring reliability in complex scenarios [16].

Ethical compliance and human-centric design must be prioritized to adhere to established standards and ensure fairness [11]. Rule-based safety monitors, alongside learned policies, ensure systems follow ethical guidelines and traffic laws. Additionally, AI models must actively minimize biases to prevent the propagation of harmful stereotypes during reasoning and interaction [16].

Safety in real-time operations is a core challenge [10]. Addressing computational constraints on edge devices ensures timely decision-making and prevents delays that could lead to safety-critical failures. Resilience to adversarial attacks and anomalies is equally important for maintaining control and robustness in unpredictable environments [7].

4.2 Remaining Challenges and Future Directions

Scalability and generalization across diverse geographical and environmental conditions remain significant challenges. Enhancements in multimodal models, such as Vision-Language Planning (VLP), are necessary to improve performance and adaptability in varied scenarios [15, 3]. Real-time adaptation and computation demand optimized transformers and generative models that can operate efficiently on edge devices while supporting high-frequency planning and decision-making [17, 24].

Data and multimodal fusion strategies also require further refinement. Improving early and deep multimodal fusion approaches can ensure better alignment of diverse sensor modalities, addressing synchronization issues that arise in dynamic environments [13, 19].

Ethical and regulatory frameworks must evolve to develop comprehensive safety scorecards for evaluating AI system risks [16]. These frameworks should provide formal guarantees for stability, ethical compliance, and operational safety . Lastly, enhancing explainability through advanced semantic understanding will bridge the gap between human-like reasoning and AI decision-making, ensuring transparency and trustworthiness in autonomous systems [8, 11].

4.3 Conclusion

The integration of agentic AI with multimodal foundation models represents a transformative approach in autonomous driving technology. The integration of agentic AI with multimodal foundation models has immense potential to revolutionize autonomous driving. By enabling vehicles to perceive, reason and act autonomously, The combination of EMMA's structured reasoning, KOMA's multi-agent coordination, and RoboFusion's robust perception provides a promising foundation for next-generation autonomous vehicles.

However, significant challenges remain. Real-world validation under extreme out-of-distribution conditions requires more rigorous testing protocols. Computational overhead from integrating multiple sophisticated models needs optimization for practical deployment. Questions about long-term system stability and drift in real-world conditions require further investigation.

The path forward requires balancing technical advancement with practical considerations. Future research must focus on establishing standardized benchmarks, optimizing model efficiency, and ensuring robust performance across diverse environmental conditions. Additionally, careful attention to ethical implications, safety standards, and regulatory compliance will be crucial for successful deployment.

While challenges persist, the integration of agentic AI with multimodal foundation models shows great promise for creating more capable, reliable, and safer autonomous vehicles. This technology has the potential to revolutionize transportation systems, improving efficiency, safety, and accessibility for all.

References

[1] R. Spair., "Agentic AI: Revolutionizing Autonomous Intelligence," LinkedIn, Oct. 10, 2024. [Online]. Available: https://www.linkedin.com/pulse/agentic-ai-revolutionizing-autonomous-intelligence-rick-spair-7q2ef.

[2] Z. Song et al., "RoboFusion: Towards Robust Multi-Modal 3D Object Detection via SAM," arXiv preprint arXiv:2401.03907, Jan. 2024.

[3] L. Wang et al., "Efficient Reinforcement Learning for Autonomous Driving with Parameterized Skills and Priors," arXiv preprint arXiv:2305.04412, May 2023.

[4] M. A. Manzoor, S. Albarri, Z. Xian, Z. Meng, P. Nakov, and S. Liang, "Multimodality Representation Learning: A Survey on Evolution, Pretraining and Its Applications," arXiv preprint arXiv:2302.00389, Feb. 2023.

[5] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal Learning with Transformers: A Survey," arXiv preprint arXiv:2206.06488, Jun. 2022.

[6] TELFOR 2018: 26th Telecommunications Forum, Belgrade, 20 and 21 November 2018, the SAVA Center. Telecommunications Society: Academic Mind, 2018.

[7] S.-H. Lee, Y. Jung, and S.-W. Seo, "Imagination-Augmented Hierarchical Reinforcement Learning for Safe and Interactive Autonomous Driving in Urban Environments," arXiv preprint arXiv:2311.10309, Nov. 2023.

[8] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, "Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions," arXiv preprint arXiv:2112.11561, Dec. 2021.

[9] Z. Zhang et al., "AD-H: Autonomous Driving with Hierarchical Agents," arXiv preprint arXiv:2406.03474, Jun. 2024.

[10] W. Shao, J. Xu, Z. Cao, H. Wang, and J. Li, "Uncertainty-Aware Prediction and Application in Planning for Autonomous Driving: Definitions, Methods, and Comparison," arXiv preprint arXiv:2403.02297, Mar. 2024.

[11] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, "Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions," arXiv preprint arXiv:2112.11561, Dec. 2021.

[12] K. Jiang et al., "KoMA: Knowledge-driven Multi-agent Framework for Autonomous Driving with Large Language Models," arXiv preprint arXiv:2407.14239, Jul. 2024.

[13] Y. Xiao, F. Codevilla, A. Gurram, O. Urfalioglu, and A. M. López, "Multimodal End-to-End Autonomous Driving," IEEE Transactions on Intelligent Transportation Systems, Jun. 2019, doi: 10.1109/TITS.2020.3013234.

[14] J.-J. Hwang et al., "EMMA: End-to-End Multimodal Model for Autonomous Driving," arXiv preprint arXiv:2410.23262, Oct. 2024.

[15] C. Pan et al., "VLP: Vision Language Planning for Autonomous Driving," arXiv preprint arXiv:2401.05577, Jan. 2024.

[16] J. Jabbour and V. J. Reddi, "Generative AI Agents in Autonomous Machines: A Safety Perspective," in Proceedings of the ACM Conference, Oct. 2024, doi: 10.1145/3676536.3698390.

[17] A. Prakash, K. Chitta, and A. Geiger, "Multi-Modal Fusion Transformer for End-to-End Autonomous Driving," arXiv preprint arXiv:2104.09224, Apr. 2021.

[18] J.-J. Hwang et al., "EMMA: End-to-End Multimodal Model for Autonomous Driving," arXiv preprint arXiv:2410.23262, Oct. 2024.

[19] K. Huang, B. Shi, X. Li, X. Li, S. Huang, and Y. Li, "Multi-modal Sensor Fusion for Auto Driving Perception: A Survey," arXiv preprint arXiv:2202.02703, Feb. 2022.

[20] C.-J. Hoel, K. Driggs-Campbell, K. Wolff, L. Laine, and M. J. Kochenderfer, "Combining Planning and Deep Reinforcement Learning in Tactical Decision Making for Autonomous Driving," IEEE Transactions on Intelligent Vehicles, May 2019, doi: 10.1109/TIV.2019.2955905.

[21] "Bridging the Gap Between Perception and Navigation," 2023.

[22] Y. Xiao, F. Codevilla, A. Gurram, O. Urfalioglu, and A. M. López, “Multimodal End-to-End Autonomous Driving,” IEEE Transactions on Intelligent Transportation Systems, Jun. 2019, doi: 10.1109/TITS.2020.3013234.

[23] Y. Xiao, F. Codevilla, A. Gurram, O. Urfalioglu, and A. M. López, “Multimodal End-to-End Autonomous Driving,” IEEE Transactions on Intelligent Transportation Systems, Jun. 2019, doi: 10.1109/TITS.2020.3013234.

[24] R. Zhang et al., “Multi-Agent Reinforcement Learning for Autonomous Driving: A Survey,” arXiv preprint arXiv:2408.09675, Aug. 2024.

[25] Q. Liu, X. Li, S. Yuan, and Z. Li, "Decision-Making Technology for Autonomous Vehicles: Learning-Based Methods, Applications and Future Outlook," Decision-making AI for Self-Driving Cars, Beijing Institute of Technology and Delft University of Technology, 2024.

[26] Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments

[27] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, “Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments,” Jun. 2017, [Online]. Available: http://arxiv.org/abs/1706.02275