A Survey of Multi-Modal Large Models in Robotics and Autonomous Systems

www.surveyx.cn

Abstract

This survey explores the transformative impact of multi-modal large models in robotics and autonomous systems, highlighting their integration of diverse data modalities and learning techniques to enhance decision-making, perception, and interaction capabilities. Central to these advancements is the use of multi-fidelity reinforcement learning (MFRL) and embodied intelligence, which optimize trajectory planning and improve robustness in autonomous navigation. The survey examines applications across various domains, including autonomous driving, drone autonomy, and industrial automation, emphasizing the role of sensor fusion and hybrid models in enhancing system accuracy and reliability. Challenges such as scalability, computational complexity, and multi-modal integration are discussed, underscoring the need for flexible frameworks and robust verification mechanisms. Future research directions focus on improving model generalization, data collection methods, and safety evaluation frameworks to advance the field. By addressing these challenges, multi-modal large models promise to revolutionize robotics, enabling the development of intelligent, adaptive, and resilient autonomous systems capable of thriving in complex environments.

1 Introduction

1.1 Concept of Multi-Modal Large Models in Robotics

Multi-modal large models in robotics and autonomous systems represent an advanced framework that integrates diverse data streams and learning paradigms to enhance robotic adaptability and functionality. Central to this framework is the use of multi-fidelity reinforcement learning (MFRL), which optimizes trajectory planning for unmanned aerial vehicles (UAVs) by leveraging various fidelity levels in simulation environments [1]. This principle of combining multiple data modalities and learning techniques underpins improved decision-making and operational efficiency.

In autonomous driving, Multi-Modal Large Language Models (MLLMs) enhance system explainability and generalization, addressing the limitations of traditional black-box models [2]. By incorporating diverse behavioral models of road users, these models significantly bolster the robustness and reliability of navigation systems [3].

The application of multi-modal large models extends to vision-based learning for drones, where visual perception and decision-making are critical for enhancing autonomy [4]. This integration of sensory data with sophisticated learning algorithms refines aerial robotics' autonomy and decision-making capabilities.

Embodied intelligence is crucial for developing these models, as it addresses the limitations of existing large language model (LLM) agents in executing complex tasks in dynamic environments, such as Minecraft [5]. This necessity for models to adapt and learn from their physical surroundings is echoed in machine learning applications in robotics [6].

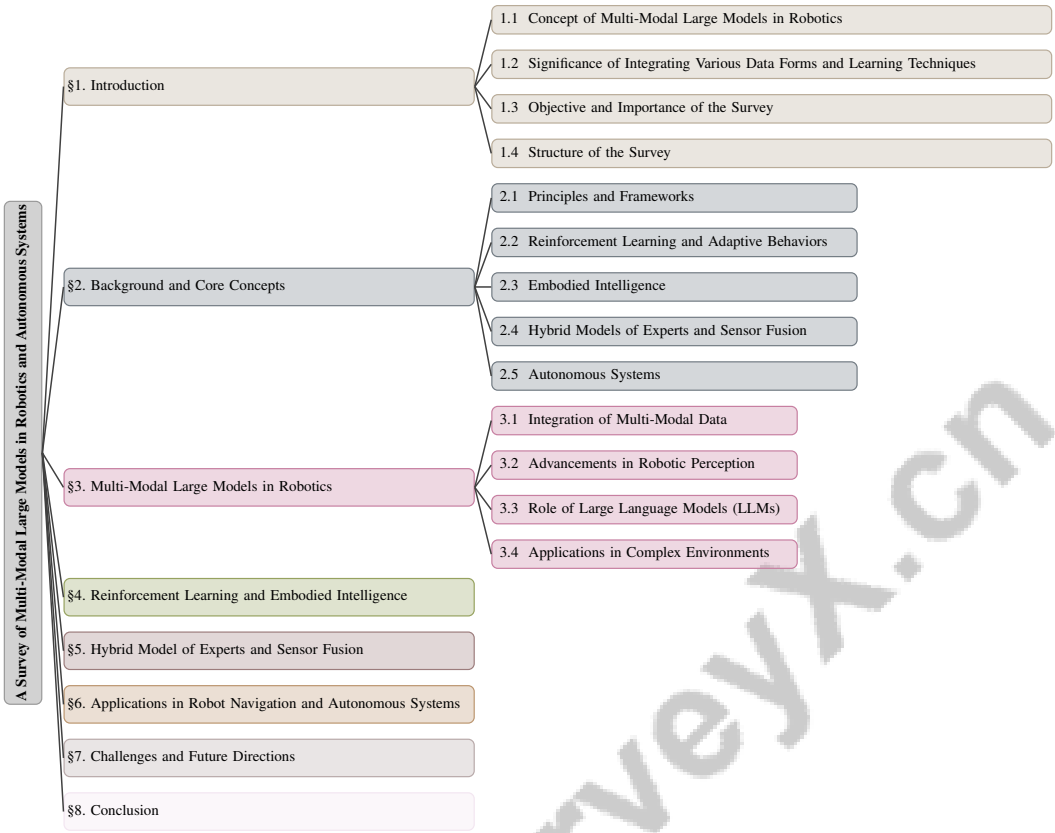

Figure 1: chapter structure

The transition from traditional Automated Guided Vehicles (AGVs) to more versatile Autonomous Mobile Robots (AMRs) in manufacturing illustrates the application of multi-modal large models for automating internal logistics and inventory tasks [7]. This shift underscores the need for flexible and adaptive models to meet the complex operational demands of modern industries.

Scalability of foundation models (FMs) poses a challenge due to the scarcity of high-quality internet data for training, particularly in multi-robot systems where navigation failures often require human intervention [8, 9].

Recent surveys in embodied visual navigation highlight advancements in navigation agents empowered by deep learning methods, showcasing the potential of these models in complex environments [10]. By integrating diverse data modalities and advanced learning techniques, multi-modal large models in robotics foster a transformative approach to enhancing robotic intelligence and autonomy. Innovations such as the REAL framework integrate LLMs into mission planning, significantly improving resilience in unmodeled situations. These advancements facilitate the autonomous collection and utilization of data, refining instruction-following capabilities and expanding operational scopes across various applications [11, 12, 13, 14, 15].

1.2 Significance of Integrating Various Data Forms and Learning Techniques

Integrating diverse data forms and learning techniques is vital for advancing robotic systems, particularly in enhancing adaptability and operational efficiency in complex environments. Current machine learning methods often struggle with the complexities of real-world environments, highlighting the need for robust integration strategies [6]. By combining different data modalities, robots can process and interpret extensive information, thereby improving decision-making capabilities.

A self-organized control paradigm allows robots to generate meaningful behaviors based solely on sensor feedback, exemplifying the importance of integrating various data forms [16]. This approach enhances adaptability to dynamic environments by leveraging sensor data for informed control

strategies. The challenges of applying deep reinforcement learning algorithms to motion planning further illustrate the necessity of integrating multiple data forms and learning techniques for robust robotic performance [17].

Incorporating semantic information into robotic systems improves navigation and intelligent human interaction, addressing limitations in traditional methods that often lack contextual understanding [18]. The integration of evolutionary dynamics with reinforcement learning enhances learning efficiency and performance in robotic systems [19].

The use of YOLO-based environmental perception frameworks underscores the significance of integrating visual data for improved decision-making in robotic systems [20]. This integration is crucial for enabling effective visual input processing, enhancing situational awareness and interaction capabilities. The multi-fidelity reinforcement learning method, which combines reinforcement learning with multi-fidelity Bayesian optimization, demonstrates the importance of real-time trajectory adaptation through data modality integration [1].

Challenges in embodied visual navigation, such as high costs of real-world data collection and model instability from partial observations, emphasize the need for integrating various data forms and learning techniques to bridge the gap between simulated and real-world environments [10]. This integration enhances robotic adaptability and efficiency, paving the way for the development of intelligent, resilient autonomous systems capable of operating in diverse environments.

1.3 Objective and Importance of the Survey

This survey aims to systematically explore the transformative impact of multi-modal large models on robotics, focusing on enhancing autonomous decision-making and task execution capabilities. By integrating diverse data modalities and advanced learning techniques, these models promise to advance robotic intelligence and adaptability [11]. The survey provides a comprehensive examination of these models, elucidating their role in overcoming existing limitations and guiding future research directions [17].

A significant aspect is highlighting the integration of large language models (LLMs) with reinforcement learning (RL), facilitating scalable, unsupervised learning for complex tasks like object rearrangement [2]. This integration is crucial for developing robust algorithms capable of real-time learning and adaptation, achieving fully autonomous operations in Unmanned Aerial Systems (UAS) and other robotic platforms [15].

The survey also addresses challenges and opportunities for humanoid robots to autonomously perform complex loco-manipulation tasks, advancing humanoid robotics [21]. It explores navigation policies that enable robots to navigate crowded environments socially compliant, utilizing model-based reinforcement learning for improvement [22].

Additionally, innovative methods like the Language Frontier Guide (LFG) are proposed to enhance robot navigation using semantic knowledge, crucial for improving robot-environment interaction [23]. By examining the integration of foundation models such as Sora into robotic systems, the survey underscores their potential to enhance capabilities and automate tasks, contributing to the development of intelligent autonomous systems [24].

The survey identifies challenges and advancements in developing large model architectures (LMAs), emphasizing the need for standardized evaluation frameworks to assess multidimensional intelligence, including crystallized, fluid, social, and embodied intelligence. It aims to analyze current technologies and identify critical gaps in robotic systems, particularly in multimodal communication for military applications. By reviewing advancements in human-robot interaction, unmanned autonomous systems, and vision-based learning, the survey seeks to facilitate the creation of intelligent, adaptive, and resilient autonomous systems that perform complex tasks in unstructured environments, enhancing operational decision-making and communication capabilities in military contexts [12, 4, 25].

1.4 Structure of the Survey

This survey is systematically organized to provide a comprehensive understanding of multi-modal large models in robotics and autonomous systems. It begins with an introduction to the concept and significance of these models, emphasizing their potential to enhance navigation and decision-making

capabilities through the integration of diverse data forms and learning techniques. The introductory section outlines the survey's objectives and importance in advancing robotic research.

Following the introduction, the survey delves into the background and core concepts essential for understanding multi-modal large models, exploring fundamental principles and frameworks, the role of reinforcement learning in adaptive behaviors, the concept of embodied intelligence, and contributions of hybrid models and sensor fusion to robotic accuracy and reliability. It concludes with an examination of autonomous systems in the context of multi-modal large models.

The subsequent section focuses on applications of multi-modal large models in robotics, discussing data integration and learning techniques that enhance perception, interpretation, and interaction. It highlights advancements in robotic perception, the role of large language models, and applications in complex environments.

The survey comprehensively explores the integration of reinforcement learning and embodied intelligence to enhance robotic capabilities, emphasizing the unique challenges of machine learning in robotics, such as adaptability in diverse environments and safety in interactions. It includes a detailed analysis of hybrid models that combine expert systems with sensor fusion techniques, highlighting their role in improving decision-making processes and environmental perception in robotic systems. Additionally, it reviews the impact of Large Language Models (LLMs) on embodied navigation, discussing their potential to facilitate sophisticated decision-making and environmental understanding in robotic applications [6, 26]. This includes discussions on sensor fusion, data quality, and applications in navigation and task execution.

The practical applications of multi-modal large models in robot navigation and autonomous systems are explored next, with case studies in robot-assisted therapy, human-robot interaction, optimization in autonomous driving and traffic systems, and real-world implementations in industrial settings. The section highlights advancements in multi-robot systems and collaborative tasks.

Finally, the survey addresses challenges and future directions in developing and implementing multi-modal large models in robotics. It discusses challenges in multi-modal integration and generalization, scalability and computational complexity, and potential future research directions. The survey concludes by summarizing key findings and reinforcing the importance of multi-modal large models in advancing robotics and autonomous systems. It introduces a framework categorizing robots into physical, human, and digital types, conceptualizing the integration of AI technologies into a 'foundation robot' system [24].The following sections are organized as shown in Figure 1.

2 Background and Core Concepts

2.1 Principles and Frameworks

The development of multi-modal large models in robotics and autonomous systems is rooted in principles that integrate diverse data modalities and learning paradigms, enhancing robotic autonomy and adaptability. Probabilistic representations, as used in RAG-Driver, improve trajectory forecasting and decision-making by utilizing a hybrid vector and textual database for autonomous driving [2]. The integration of foundation models into robotics is categorized into decision-making, perception, and multimodal integration stages, systematically enhancing robotic capabilities [27]. The MAZero algorithm exemplifies efficient policy search in cooperative environments through a centralized model and Monte Carlo Tree Search (MCTS), showcasing model-based multi-agent reinforcement learning's potential [28].

Vision-based control methods are classified into indirect, semi-direct, and end-to-end approaches, facilitating visual data integration in drones and improving autonomy [4]. A self-improvement framework emphasizes models' ability to generate their own training data, setting them apart from traditional supervised and reinforcement learning [8]. Theoretical perspectives on machine learning in robotics highlight the need for inductive biases specific to embodied learning, emphasizing action-perception interplay [6]. The Versatile Instructable Motion prior (VIM) framework enables legged robots to learn agile locomotion by imitating animal and human motions, integrating reinforcement learning with biologically inspired models [29].

Embodied visual navigation methods underscore deep learning approaches' effectiveness, integrating deep learning with embodied intelligence to enhance navigation in complex environments [10].

Collectively, these principles and frameworks advance multi-modal large models, fostering intelligent, adaptive, and resilient robotic systems capable of autonomous operation in dynamic environments.

2.2 Reinforcement Learning and Adaptive Behaviors

Reinforcement learning (RL) is crucial for developing adaptive robotic behaviors in complex environments. Integrating RL with semantic information allows robots to adjust behaviors in response to human movements, enhancing adaptability in crowded settings [18]. The Domain-Enriched Reinforcement Learning (DERL) framework exemplifies RL's role in evolving adaptive behaviors, contributing to embodied intelligence by enabling autonomous learning without manual resets or pre-engineered reward functions [19, 30].

In autonomous mobile robots, RL fosters resilience to unforeseen circumstances during missions [11]. The Multi-Fidelity Reinforcement Learning (MFRL) method enhances adaptability by integrating RL with multi-fidelity evaluations for trajectory planning [1]. Despite advancements, challenges persist, including the need for large datasets, difficulties in generalizing learned behaviors, and performance limitations of modular systems [17]. Addressing sample inefficiency in multi-agent reinforcement learning (MARL) is crucial for improving data efficiency and real-time performance in robotic applications [28, 27].

Training legged robots for agile locomotion tasks while adapting to sensory inputs highlights RL's potential in enhancing robotic agility [29]. These advancements and challenges underscore RL's pivotal role in fostering adaptive behaviors, enabling robots to learn from diverse data sources and adapt to intricate environments, thereby improving intelligence and resilience [11, 12].

2.3 Embodied Intelligence

Embodied intelligence integrates a robot's physical form with control mechanisms to enhance environmental interaction, particularly in neuromorphic robotics, where systems mimic biological neural structures for improved adaptability [31]. The relationship between morphological computation and controller complexity is crucial, as a robot's physical structure can offload computational tasks, leading to more efficient behaviors [32]. The MEIA model exemplifies embodied intelligence by translating high-level tasks in natural language into executable actions, integrating perception, cognition, and action for autonomous task performance [33].

In designing robots with embodied intelligence, coupling the controller and morphology is essential, especially in synthetic cells with limited computational capabilities [34]. Embodied intelligence encompasses tasks such as navigation and question answering, necessitating sensory processing, decision-making, and action execution [35]. Integrating embodied intelligence with Large Language Models (LLMs) presents challenges, such as adversarial attacks manipulating text prompts, emphasizing the need for robust security measures [36].

Research in social robotic navigation categorizes approaches into stages involving CNN-based path planning, socially aware motion models using deep reinforcement learning (DRL), and incremental learning approaches [37]. The theory of intelligence as an emergent property from crystallized, fluid, social, and embodied types provides a comprehensive framework for understanding intelligent robots' capabilities [38]. In complex task execution, embodied intelligence is vital for overcoming challenges faced by existing LLM-based agents, underscoring the necessity for models integrating sensory inputs with action-oriented decision-making [5].

2.4 Hybrid Models of Experts and Sensor Fusion

Hybrid models of experts and sensor fusion enhance robotic systems' accuracy and reliability in complex environments by leveraging multiple data sources and expert knowledge. Integrating trajectory forecasting with planning and control is essential, as highlighted by structured dynamical system representations for interpretable outcomes [39]. Markov Logic Networks (MLNs) offer flexible performance in data processing compared to traditional hard-rule approaches, effectively addressing sensor fusion challenges [40].

In vision-based learning for drones, challenges include integrating complex visual perception systems and ensuring real-time decision-making [4]. A dataset encompassing a wide range of cognitive

tasks aids in evaluating dimensions of intelligence integral to developing hybrid models [41]. The necessity for new theoretical frameworks for embodied intelligence highlights hybrid models' role in integrating cognitive science principles into robotics [6]. Despite advancements, limitations persist in complex environments and generalizing from simulations to real-world tasks [10], indicating a need for ongoing research in hybrid models and sensor fusion.

These advancements emphasize hybrid models' pivotal role in improving robotic systems' accuracy and reliability. By integrating diverse data sources and expert insights, these models enhance autonomous systems' functionality, enabling efficient navigation and operation in intricate environments, expanding applicability across various domains, including commercial and military uses [15, 13].

2.5 Autonomous Systems

Autonomous systems are foundational in deploying multi-modal large models, enabling robots to perform complex tasks with minimal human intervention while maintaining efficiency and adaptability. These systems operate in dynamic environments by processing multiple data streams to perceive, interpret, and interact with surroundings. Integrating multi-modal large models within autonomous systems enhances robotic intelligence, allowing them to respond to complex user inputs and execute intricate tasks based on natural language instructions and visual perceptions. Robust evaluation frameworks and security measures address performance and safety challenges, improving robotic systems' efficacy in mission-oriented tasks [42, 43, 44, 45].

In autonomous driving, multi-modal multi-task visual understanding foundation models are crucial for interpreting road scenes, enhancing navigation safety and reliability [46]. The modular architecture of GRID exemplifies rapid prototyping and adaptation potential across robotics applications, essential in dynamic environments where robotic trajectories must adapt to sparse measurements [47]. Autonomous systems face challenges in estimating appropriate actions in diverse driving scenarios [48]. Limitations of service robots, such as the lack of real-time validation mechanisms, underscore the need for robust systems capable of safe operations [49]. Addressing inefficiencies in modular designs that separate perception, prediction, planning, and control is crucial for enhancing performance [50].

Embodied intelligence and real-time computation are vital for autonomous systems, facilitating decision-making and environmental interaction [51]. The one-to-many supervision problem emphasizes effective decision-making frameworks in uncertain environments [9]. Coordinating multiple Autonomous Mobile Robots (AMRs) in dynamic manufacturing environments illustrates autonomous systems' role in automating logistics efficiently [7]. In multi-modal large models, autonomous systems must enable vehicles to interact competently with various road users, addressing the interplay between autonomous entities and human drivers [3]. Autonomous systems are integral to implementing multi-modal large models, providing the foundation for developing intelligent, adaptive, and resilient robotic systems. By leveraging advanced learning architectures and decision-making frameworks, these systems promise to revolutionize robotics, enabling robots to operate independently in complex environments.

3 Multi-Modal Large Models in Robotics

The exploration of multi-modal large models in robotics reveals their pivotal role in augmenting robotic capabilities across diverse applications. By integrating visual, auditory, and contextual data, these models offer a comprehensive approach to perception and interaction, enhancing adaptability and efficiency in real-world environments. Figure 2 illustrates the hierarchical structure of these multi-modal large models, highlighting key categories such as data integration, advancements in perception, the role of large language models, and applications in complex environments. Each of these categories is further divided into subcategories, detailing specific frameworks, methods, and applications that enhance robotic capabilities in perception, decision-making, and interaction. The following subsections explore specific strategies for multi-modal data integration and its impact on advancing robotic capabilities.

3.1 Integration of Multi-Modal Data

Integrating multi-modal data is essential for enhancing robotic perception and interaction, enabling precise and adaptable task execution. Systems like LHControl utilize learning-based hierarchical

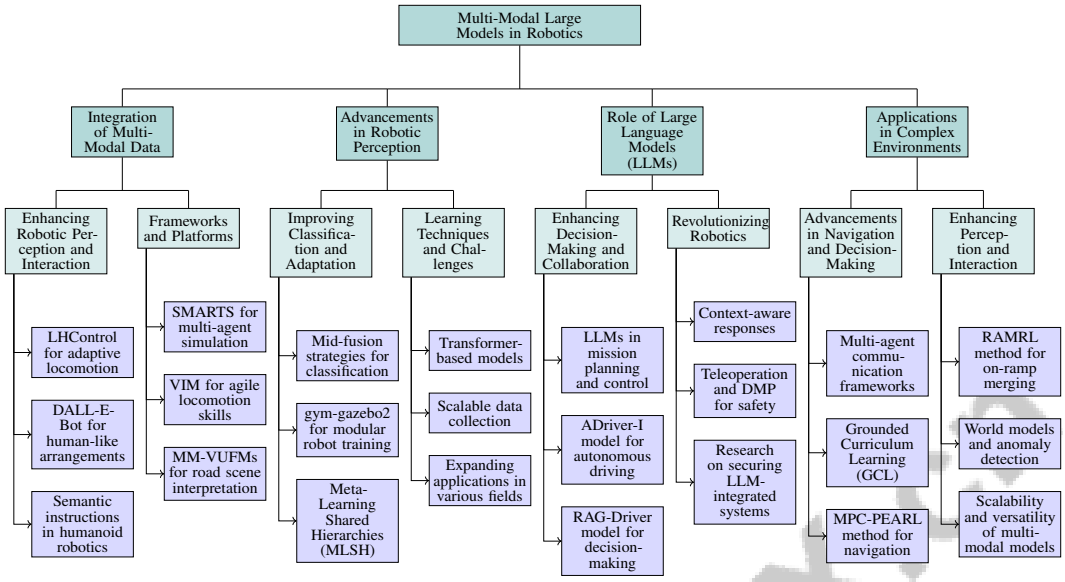

Figure 2: This figure illustrates the hierarchical structure of multi-modal large models in robotics, highlighting key categories such as data integration, advancements in perception, the role of large language models, and applications in complex environments. Each category is further divided into subcategories, detailing specific frameworks, methods, and applications that enhance robotic capabilities in perception, decision-making, and interaction.

control to adapt locomotion strategies based on environmental feedback [52]. Frameworks such as DALL-E-Bot leverage multi-modal data to autonomously infer and execute human-like arrangements, enhancing perceptual capabilities [53].

As illustrated in Figure 3, the integration of multi-modal data enhances robotic capabilities across three main areas: Robotic Perception, Task Execution, and Autonomous Navigation. Each category highlights key frameworks and systems that leverage multi-modal data for improved performance.

In humanoid robotics, integrating multi-modal data for task planning and execution through semantic instructions significantly boosts perception and interaction [21]. Multi-modal cues from human teachers, like gaze and speech, further emphasize human-robot collaboration's role in advancing robotic capabilities [54].

The SMARTS platform exemplifies improved perception and interaction through multi-agent simulation environments [3]. The VIM framework enables legged robots to learn agile locomotion skills by integrating existing knowledge and structured feedback [29].

In road scene interpretation, MM-VUFMs integrate visual tasks to enhance navigation safety and reliability [46]. The GRID framework's Foundation Mosaic reduces sample complexity for pre-training Robotics Foundation Models [47], crucial for efficient multi-modal models in complex environments [55].

The DERL framework integrates evolutionary processes and sensory data learning to enhance perception and interaction [19]. The MFRL method uses fidelity evaluations to improve perception in dynamic environments [1].

In autonomous navigation, integrating geometric control, trajectory optimization, and deep learning underscores the comprehensive approach needed for effective operation [51]. The EyeSim VR environment generates multi-modal inputs, highlighting multi-modal data integration's benefits in enhancing robotic capabilities [42].

These advancements significantly enhance robotic systems' operational capabilities, allowing effective interaction in complex, dynamic environments. Multi-modal models promise to revolutionize robotics, facilitating intelligent, adaptive, and resilient autonomous systems across diverse fields, from military to medical applications [56, 57, 25].

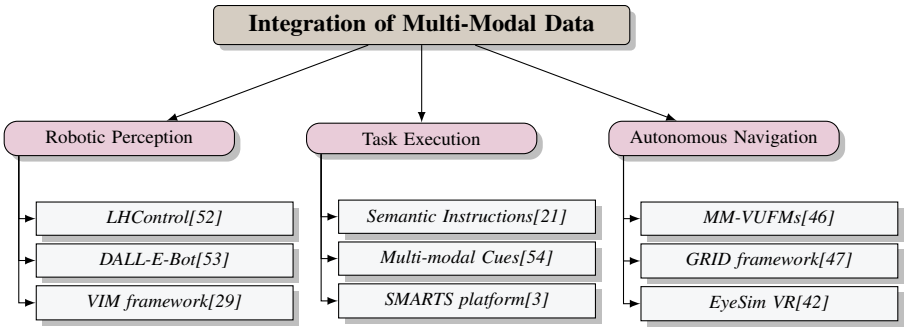

Figure 3: This figure illustrates the integration of multi-modal data in enhancing robotic capabilities, categorized into three main areas: Robotic Perception, Task Execution, and Autonomous Navigation. Each category highlights key frameworks and systems that leverage multi-modal data for improved performance.

3.2 Advancements in Robotic Perception

Recent advancements in robotic perception, driven by multi-modal models, have significantly enhanced robotic capabilities in complex environments. Mid-fusion strategies improve classification accuracy of container contents, highlighting the importance of sensory input combination [58].

Frameworks like gym-gazebo2 provide robust platforms for training modular robots, enhancing perception through reinforcement learning [59]. These frameworks enable better adaptation to dynamic environments, improving interaction and decision-making.

Meta-Learning Shared Hierarchies (MLSH) demonstrate the potential of hierarchical models in advancing perception and adaptability [60]. Transformer-based models in learning robot dynamics show promise in enhancing capabilities in complex environments [61].

These advancements illustrate the transformative impact of integrating diverse data sources and learning techniques, laying the groundwork for intelligent, resilient autonomous systems. Overcoming challenges like scalable data collection and robust learning algorithms enables rapid deployment of robots capable of learning from their environments, expanding applications in manufacturing, service, and healthcare [12, 13].

3.3 Role of Large Language Models (LLMs)

Integrating Large Language Models (LLMs) into robotics significantly enhances decision-making, adaptability, and human-robot interaction. LLMs improve mission planning and control, enhancing resilience in dynamic environments [11].

In autonomous driving, the ADriver-I model uses LLMs and video diffusion models to predict control signals and future frames, improving decision-making in driving scenarios [50]. LLMs enhance human-robot collaboration by improving task planning and collaboration [20]. The RAG-Driver model showcases LLMs' impact on contextualized decision-making [2].

LLMs enhance adaptability in robotic planning and reasoning [27]. The LARM method facilitates faster task execution by generating action tokens directly [5].

LLMs revolutionize robotics by enabling context-aware responses and improving task planning. They address safety and reliability challenges through teleoperation and Dynamic Movement Primitives (DMP) for action correction. Ongoing research into securing LLM-integrated systems highlights the importance of robust defense strategies [26, 42, 20]. LLMs promise to revolutionize robotics, facilitating intelligent, adaptive, and resilient autonomous systems in complex environments.

3.4 Applications in Complex Environments

Multi-modal models in complex environments significantly advance autonomous systems, particularly in navigation and decision-making. Multi-agent communication frameworks enhance task planning for multiple robots, improving collaboration and coordination [65].

Method Name	Methodologies	Application Domains	Capabilities Enhancement
GCL[62]	Grounded Curriculum Learning	Autonomous Navigation	Improves Learning Efficiency
MPC-PEARL[63]	Mpc-PEARL	Robot Navigation	Improve Learning Efficiency
RAMRL[64]	Ramrl	Autonomous Driving	Improve Decision-making
GRID[47]	Foundation Models	Robotics Applications	Perception State Estimation

Table 1: Summary of methodologies, application domains, and capabilities enhancement of selected methods in autonomous systems. The table highlights Grounded Curriculum Learning (GCL), MPC-PEARL, RAMRL, and GRID, focusing on their contributions to improving learning efficiency, decision-making, and perception state estimation in robotics and autonomous navigation.

In autonomous navigation, the Grounded Curriculum Learning (GCL) method improves navigation performance, emphasizing structured learning approaches [62]. The MPC-PEARL method combines predictive control with adaptive learning, enhancing navigation in dynamic environments [63].

In autonomous driving, the RAMRL method addresses on-ramp merging, showcasing reinforcement learning's potential in complex scenarios [64]. World models and anomaly detection methods offer insights into enhancing detection techniques [66].

MM-VUFMs demonstrate the potential of multi-modal models in enhancing perception and interaction [46]. Experiments with the AirGen simulator highlight multi-modal models' scalability and versatility in diverse challenges [47].

As illustrated in Figure 4, this figure highlights the applications of multi-modal models in complex environments, showcasing advancements in autonomous navigation, autonomous driving, and the use of multi-modal models. Specifically, it underscores the potential of methods such as GCL, MPC-PEARL, and RAMRL, along with the application of world models and simulators like AirGen, in enhancing perception, interaction, and decision-making in dynamic scenarios. Embodied Vision-Language Planning organizes research into tasks, approaches, and evaluation, highlighting the interplay between visual and linguistic data. The Grid-based Robotic Operations System streamlines operations through a grid-based approach, facilitating efficient communication and data processing. These examples underscore the potential of multi-modal models to enhance robotic systems' capabilities in complex environments, paving the way for more intelligent and adaptable systems [35, 47]. In addition, Table 1 provides a comprehensive overview of various methodologies applied in complex environments, detailing their specific application domains and the enhancements they bring to autonomous systems.

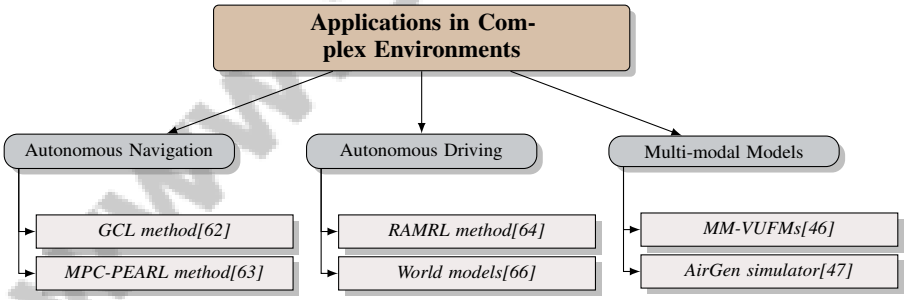

Figure 4: This figure illustrates the applications of multi-modal models in complex environments, highlighting advancements in autonomous navigation, autonomous driving, and multi-modal models. It underscores the potential of methods like GCL, MPC-PEARL, RAMRL, and the use of world models and simulators such as AirGen in enhancing perception, interaction, and decision-making in dynamic scenarios.

4 Reinforcement Learning and Embodied Intelligence

4.1 Adaptive Behaviors and Decision-Making

Reinforcement learning (RL) and embodied intelligence are pivotal in cultivating adaptive behaviors and enhancing decision-making within robotic systems. Their integration allows robots to effectively navigate dynamic environments, enabling autonomous learning and real-time decision-making from raw sensory inputs, essential for complex tasks [30]. The Multi-Fidelity Reinforcement Learning

(MFRL) approach exemplifies RL's role in adaptive behaviors, refining decision-making through real-time trajectory planning and multi-fidelity evaluations to adjust actions according to changing conditions [1].

In autonomous driving, the RAG-Driver model showcases RL's contribution by providing action justifications, enhancing decision-making in complex navigation scenarios [2]. Multi-Agent Reinforcement Learning (MARL) algorithms, such as PPO and MADDPG, demonstrate effectiveness in coordinating agents, thus improving adaptability in uncertain environments [3]. The integration of affordances into RL, as proposed by Khetarpal et al., allows agents to identify feasible actions, thereby enhancing planning and learning outcomes [67]. Decentralized architectures in multi-agent autonomous mobile robots further promote operational efficiency through real-time communication and coordination, facilitating effective collaboration in complex settings [7].

The MAZero algorithm enhances sample efficiency and policy learning through structured planning in multi-agent contexts [28], while the VIM framework emphasizes simultaneous learning of multiple agile locomotion skills, improving sample efficiency and task transitions [29]. Integrating RL and embodied intelligence significantly augments adaptive behaviors and decision-making capabilities, paving the way for intelligent, adaptable robots capable of intricate tasks in unstructured environments. Leveraging large-scale fleets for autonomous data collection enhances instruction-following capabilities without costly human interventions, streamlining learning processes and developing robust models for multifaceted tasks [12, 13].

4.2 Integration with Large Language Models

The synergy between reinforcement learning (RL) and large language models (LLMs) represents a transformative advancement in robotic intelligence, enhancing adaptability and decision-making. This integration leverages the strengths of both paradigms, enabling robots to perform complex tasks with increased efficiency. A notable innovation is the combination of Successor Features Generalized Policy Improvement (SF-GPI) and value composition methods, facilitating real-time policy composition and improving sample efficiency and task transferability [68].

The Meta-Learning Shared Hierarchies (MLSH) framework integrates gradient signals between policies, enhancing RL's sample efficiency when combined with LLMs, thus achieving robust learning across diverse task distributions [60]. Incorporating domain knowledge into RL, as demonstrated by C-PPO, reduces communication load while maintaining constraints on long-term modeling error, ensuring efficient operation in dynamic environments [69].

Secure prompting techniques for LLMs, evaluated for detecting and mitigating prompt injection attacks, highlight the importance of robustness in LLM-based models for safe and reliable operation in adversarial environments [42]. Exploring task embodiment using Kullback-Leibler divergence quantifies design complexity through graph entropy, enhancing robots' ability to process complex tasks and adapt to new situations when integrated with LLMs [34].

Bi-level optimization approaches, such as combining Multi-Objective Bayesian Optimization (MOBO) with Multi-Agent Reinforcement Learning (MARL), optimize fleet composition and performance evaluation in complex environments, leveraging LLM capabilities for improved system performance [70]. The integration of RL with LLMs fosters adaptive, efficient, and robust autonomous systems. By combining diverse learning paradigms and utilizing LLMs, particularly through multimodal approaches like GPT-4V, these methodologies enhance robots' ability to process natural language instructions and visual data for improved task execution. This integration addresses security vulnerabilities and generalization challenges, enhancing the safety, reliability, and overall performance of robotic operations across various mission-oriented contexts [71, 42, 44].

5 Hybrid Model of Experts and Sensor Fusion

5.1 Sensor Fusion and Data Quality

Sensor fusion is critical in enhancing data quality for robotics, enabling systems to operate efficiently in complex environments. By integrating multiple sensory inputs, robots achieve a comprehensive understanding of their surroundings, which bolsters decision-making and operational efficacy. For example, merging Basic Safety Messages (BSM) with surveillance images significantly enhances

data quality for Connected and Autonomous Vehicles (CAVs), resulting in safer merging strategies [64]. Additionally, audio integration with other sensory data further enriches the contextual understanding, underscoring the role of sensor fusion in developing reliable autonomous systems capable of navigating dynamic environments [72, 15].

In robotics, sensor fusion facilitates task execution by integrating diverse skills and environmental parameters [73]. This is crucial for grounding language models in real-world contexts, enhancing perception and action [21]. Understanding pedestrian dynamics is vital for navigation tasks, as demonstrated by the MR-LSP framework, which predicts action outcomes to improve data quality [74]. LiDAR-driven reinforcement learning also emphasizes the importance of integrating diverse data sources to enhance data quality [75, 1].

Sensor fusion is integral to improving data quality in robotics, facilitating accurate perception, interpretation, and interaction with the environment. By processing asynchronous sensory data—such as video, proprioceptive states, and force-torque measurements—robots can respond in real-time to dynamic conditions. Techniques like Decentralized Distributed Expert-Assisted Learning (D2EAL) enhance cooperative target tracking among heterogeneous robots through information sharing and adaptive learning, optimizing predictive accuracy [12, 76, 6, 77].

5.2 Applications in Navigation and Task Execution

Hybrid models and sensor fusion significantly enhance navigation and task execution capabilities in robotic systems, particularly in dynamic environments. These models utilize diverse sensory inputs and expert knowledge to improve decision-making and operational performance. In autonomous driving, integrating probabilistic safety guarantees with belief updates and scene decomposition enhances navigation safety and efficiency by managing interactions among multiple traffic participants [78]. The advancement of CAVs illustrates how hybrid models and sensor fusion improve traffic flow and reduce accidents, fostering cooperation among CAVs and surpassing traditional ego-driving models for safer navigation [79]. Figure 5 illustrates the application of hybrid models and sensor fusion in navigation and task execution, highlighting their roles in autonomous driving, cooperation among CAVs, and task efficiency through multimodal data integration.

Beyond autonomous driving, sensor fusion optimizes task execution in robotic systems. The integration of multi-modal data, including visual and auditory inputs, enhances the robot's ability to perceive and interpret environmental cues, improving task execution efficiency. This capability is crucial in high-precision contexts, such as search and rescue operations, where accurate localization and manipulation are essential, and in industrial automation, where robots must reliably perform complex tasks in unstructured environments. Enhanced learning techniques and multimodal perception approaches significantly improve performance in these scenarios, enabling robots to adapt to varying conditions [12, 58, 80].

The application of hybrid models and sensor fusion in navigation and task execution underscores their critical role in advancing autonomous systems. By seamlessly integrating diverse data sources, including Vision-Language Models (VLMs) and advanced reinforcement learning techniques, these robotic models enhance accuracy, reliability, and efficiency. This integration allows robots to navigate complex environments and perform intricate tasks, such as open-ended pick-and-drop operations, with high success rates—demonstrated by the OK-Robot framework achieving up to 82

6 Applications in Robot Navigation and Autonomous Systems

The practical deployment of multi-modal models has significantly advanced robot navigation and autonomous systems, enhancing capabilities and human-robot interactions. This section explores their transformative impact, with case studies illustrating effectiveness in robot-assisted therapy and human-robot interaction, particularly in personalized therapeutic experiences.

6.1 Case Studies in Robot-Assisted Therapy and Human-Robot Interaction

Multi-modal models using Large Language Models (LLMs) like GPT-2 and GPT-4V have shown great promise in robot-assisted therapy for children with Autism Spectrum Disorder (ASD). These models facilitate relevant verbal interactions and perspective-taking, enhancing robotic autonomy

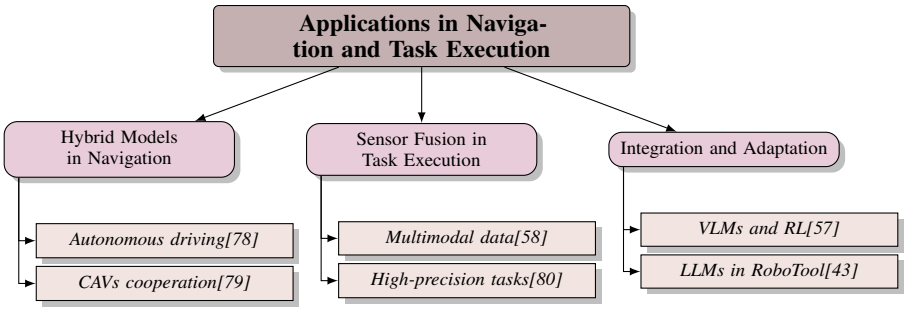

Figure 5: This figure illustrates the application of hybrid models and sensor fusion in navigation and task execution, highlighting their roles in autonomous driving, cooperation among CAVs, and task efficiency through multimodal data integration.

beyond pre-defined scripts. They improve therapy sessions by initiating conversations, prompting responses, and providing reinforcement, reducing demands on children and therapists. The integration of natural language processing with robotic visual perception enriches navigation and interaction within complex environments, tailoring therapy to individual needs and improving outcomes [81, 44].

Frameworks like FlowAct enhance human-robot interactions by enabling robots to provide proactive, multimodal assistance in environments like hospital waiting rooms. Continuous perception and dynamic coordination of actions improve robot responsiveness and patient experiences by ensuring timely, relevant support [56, 82]. This underscores the potential of multi-modal models in fostering effective therapeutic interactions.

Research shows a preference for multimodal assistance, with visual trajectory guidance favored over haptic or visual-only modalities. A study involving remote operators navigating a virtual telepresence robot demonstrated that combining visual and haptic feedback enhanced user experience, although not navigation performance [83, 54, 84]. This highlights the importance of integrating sensory inputs for optimized human-robot interactions.

Kstner et al. introduced a semantic deep reinforcement learning (DRL) agent for guiding humans in crowded environments, crucial for developing robots that navigate social settings and provide real-time assistance [18]. In manufacturing, Sousa et al.'s multi-agent system (MAS) coordinates autonomous mobile robots (AMRs) for efficient task execution, showcasing adaptability in simulated environments [7].

These case studies illustrate the transformative potential of multi-modal models in robot-assisted therapy and interaction. As depicted in Figure 6, which illustrates the hierarchical structure of key concepts in robot-assisted therapy and human-robot interaction, these models encompass various applications and studies, highlighting the interconnections between multi-modal models and human-robot interaction methods. By integrating diverse data modalities and advanced learning techniques, these models enhance robotic system effectiveness and adaptability in therapeutic contexts, fostering personalized interventions and enabling autonomous learning in unstructured environments [12, 81, 43, 13].

6.2 Optimization in Autonomous Driving and Traffic Systems

Multi-modal models optimize autonomous driving and traffic systems by enhancing vehicular efficiency and safety. These models integrate diverse data modalities and advanced learning techniques, improving decision-making and performance in dynamic traffic environments. Vision-based learning in drones showcases their versatility across sectors like agriculture and industrial inspection [4].

In autonomous driving, integrating sensory data from vision, lidar, and proprioception enables vehicles to interpret complex road scenarios, essential for motion planning and trajectory prediction. Studies show these models excel with all modalities and maintain performance with missing data, enhancing adaptability and safety [46, 85]. This is vital for developing robust navigation systems that adapt to real-time traffic changes, optimizing decisions related to speed, lane changes, and obstacle avoidance.

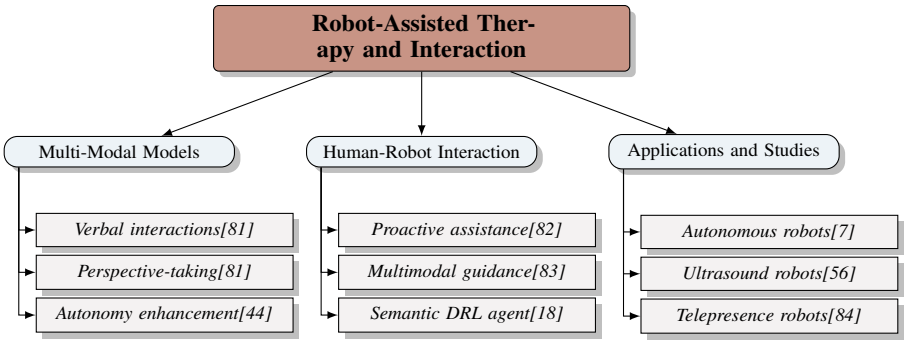

Figure 6: This figure illustrates the hierarchical structure of key concepts in robot-assisted therapy and human-robot interaction, highlighting multi-modal models, human-robot interaction methods, and various applications and studies.

Multi-agent reinforcement learning (MARL) frameworks enhance coordination among autonomous vehicles, improving communication and collaboration in shared environments. This optimizes system performance by fostering cooperative behavior among vehicles, addressing challenges like interactions with human-driven traffic. Techniques like shared policies and attention mechanisms capture agent dynamics, leading to safer transportation systems [86, 79, 28].

Integrating large language models (LLMs) into autonomous systems enhances their ability to interpret complex instructions and execute nuanced maneuvers, essential for navigating intricate scenarios. By improving context-aware decision-making, LLMs increase safety and reliability. Challenges like security risks necessitate robust defense mechanisms to ensure accurate navigation. Advances in path-planning algorithms leveraging LLMs optimize route efficiency while maintaining precision [2, 87, 88, 42].

The application of multi-modal models in autonomous driving and traffic systems offers improvements in efficiency, safety, and adaptability. By integrating various data sources and employing advanced learning techniques, these models transform vehicular operations. Innovations like the SMARTS simulation platform enhance perception and decision-making, and multi-agent graph reinforcement learning improves cooperation among connected and automated vehicles (CAVs), addressing challenges of interaction with diverse road users and system-level optimization [3, 17, 79, 89].

6.3 Real-World Implementations and Industrial Applications

Real-world implementations of multi-modal models in industry have demonstrated their transformative potential, improving operational efficiency and task execution. This progress is attributed to the integration of large language models with multimodal learning techniques, enabling AI to interpret complex queries. Innovations like Multimodal Chain of Thought (M-COT), Multimodal Instruction Tuning (M-IT), and Multimodal In-Context Learning (M-ICL) enhance model functionality, facilitating a unified approach to applications. Collaborative frameworks for large multimodal agents (LMAs) improve decision-making and reasoning in industrial environments [55, 45].

A learning-based hierarchical control framework validated in simulated environments with a quadruped robot shows potential for real-world robotic navigation applications [52]. This framework enables robots to adapt locomotion strategies based on feedback, enhancing navigation capabilities.

The Language Frontier Guide (LFG) has been evaluated in complex environments, demonstrating its effectiveness in improving robot navigation through semantic knowledge [23]. This enhances robot interaction with surroundings, paving the way for intelligent navigation systems.

In service robots, integrating safety control frameworks improves task execution rates, achieving a 95

Comparative experiments using a mobile robot simulator highlight the effectiveness of multi-modal models in optimizing control strategies [90]. These findings underscore the potential of advanced learning techniques to enhance decision-making in robotic systems.

Identifying gaps in research and potential for integrating external rule systems emphasize embodied intelligence's importance for developing robust robotic systems capable of effective operation in dynamic environments [55].

These implementations underscore the transformative impact of multi-modal models on industrial operations. By enhancing adaptability, precision, and reliability in robotic systems, advancements in robot learning and open-knowledge models are set to revolutionize sectors like manufacturing, service, and healthcare. Intelligent instruction-following capabilities and creative tool use, facilitated by vision-language models and LLMs, allow for efficient autonomous systems in intricate settings. This evolution promises to optimize efficiency and expand robots' task range, addressing complex real-world challenges [12, 6, 13, 57, 43].

6.4 Advancements in Multi-Robot Systems and Collaborative Tasks

Advancements in multi-robot systems have significantly improved collaborative task execution, enabling efficient operation in complex environments. The integration of Multi-Agent Reinforcement Learning (MARL) with model predictive control, as demonstrated by the MPC-PEARL framework, shows improved learning efficiency and navigation quality, reducing travel time and collision rates [63]. This highlights the potential of combining predictive control with adaptive learning to enhance coordination and performance.

The GRID framework emphasizes advancements in multi-robot systems by incorporating various robotic form factors and enhancing safety mechanisms. Future developments will explore new methodologies for simulation feedback, crucial for refining interaction and coordination among robotic agents [47]. This approach promises to enhance adaptability and safety in diverse environments.

Integrating affordances into reinforcement learning improves agents' planning and learning efficiency, enhancing performance in complex environments. This allows robots to better understand and predict action feasibility based on their current state [67]. By leveraging affordances, multi-robot systems achieve more effective collaboration and precision.

Advancements in multi-robot systems and collaborative applications highlight the transformative potential of integrating advanced learning techniques and control strategies. Ongoing research focuses on enhancing robot capabilities for complex manipulation tasks in unstructured environments, ensuring value alignment with human users, and enabling autonomous improvement of instruction-following skills through large-scale data collection and analysis. These efforts aim to make robots more accessible and effective in real-world scenarios, enhancing utility across sectors like manufacturing, service, and healthcare [12, 13, 91]. By improving coordination, adaptability, and efficiency, these developments promise to revolutionize robotics, paving the way for intelligent, resilient, and versatile autonomous systems capable of thriving in complex environments.

7 Challenges and Future Directions

7.1 Challenges in Multi-Modal Integration and Generalization

The integration and generalization of multi-modal models across diverse applications face significant challenges, particularly in complex environments. A major issue is the reliance on accurate models and predefined scenarios, limiting adaptability to novel tasks without extensive pre-training. This highlights the need for flexible frameworks that can accommodate new scenarios with minimal retraining [10]. The high sample complexity of learning from raw sensory inputs further complicates deployment, posing a barrier to effective application.

In autonomous driving, these challenges are intensified by the substantial computational resources needed to manage diverse scenarios. Data scarcity, due to costly annotation processes, complicates the development of robust systems capable of generalizing across environments. Current models often require extensive retraining to maintain performance. Innovations like the RAG-Driver model, which employs retrieval-augmented learning, enhance explainability and performance, while architectures such as ADriver-I integrate visual and control inputs for improved decision-making. However, transparency and system-level optimization remain critical as the industry moves towards integrated, data-driven approaches [50, 17, 2, 89]. These factors hinder the deployment of Multi-Modal Visual

Understanding Foundation Models (MM-VUFMs), essential for interpreting complex road scenes. Additionally, rapid fluctuations in control signals can lead to suboptimal frame generation, adversely affecting predictions and decision-making.

The limitations of Multi-Fidelity Reinforcement Learning (MFRL) underscore the complexities of effectively integrating and generalizing multi-modal models across applications, as researchers strive to develop unified models capable of processing diverse data inputs and enhancing reasoning in complex environments [55, 92, 83, 45]. The absence of performance guarantees in certain scenarios, coupled with the risk of model collapse due to reliance on learned verifiers for data generation, underscores the need for robust verification mechanisms. These challenges are compounded by the requirement for improved simulation environments to effectively generalize reinforcement learning (RL) policies.

In human-robot interaction, adapting agents to real-world complexities remains a significant challenge. Current studies often inadequately address critical issues in real-time processing, environmental adaptability, and the robustness of learning models in dynamic scenarios, particularly regarding embodied vision-language planning and the deployment of large language models in autonomous systems. Key challenges include ensuring model generalizability in real-world applications, effectively integrating prior knowledge for resilience against unforeseen circumstances, and interpreting complex natural language inputs for mission planning. Existing research often prioritizes architectural methods over a comprehensive understanding of these high-level challenges, highlighting the need for a holistic approach that considers all dimensions of multimodal machine learning and artificial intelligence [11, 35, 4]. Furthermore, the robustness of communication infrastructure in multi-agent systems affects coordination, limiting scalability. The ambiguous nature of large language model (LLM) outputs, which often generate descriptive sentences lacking clear actionable steps, along with slow inference speeds due to iterative token predictions, presents additional obstacles to effective multi-modal integration.

Despite these challenges, the potential of multi-modal models to transform autonomous systems across various domains is substantial. Addressing the obstacles in developing advanced AI systems requires creating flexible and adaptive models capable of seamlessly integrating diverse data sources and learning paradigms, akin to those utilized in large multimodal agents (LMAs) and multimodal large models (MLMs). These models must interpret and respond to complex multimodal inputs, enhancing decision-making and reasoning abilities similar to human cognition. Additionally, they should address challenges related to model generalizability and real-world deployment, as identified in embodied vision-language planning tasks and the evolving landscape of multimodal foundation models aiming to transition from specialized applications to generalized assistant roles [55, 93, 83, 45, 35]. By overcoming these challenges, the promise of multi-modal models to enhance the intelligence, resilience, and versatility of robotic systems can be fully realized, paving the way for advanced autonomous systems.

7.2 Scalability and Computational Complexity

The scalability and computational complexity of deploying multi-modal models in robotics and autonomous systems pose significant challenges, particularly in dynamic and complex environments. A notable limitation arises from the computational demands associated with training Multi-Agent Reinforcement Learning (MARL) agents, which can be resource-intensive and time-consuming [70]. This issue is exacerbated by the exponential growth of possible multi-robot actions as team size increases, complicating planning and scalability [74].

In safety-critical learning for robot control, the necessity of solving quadratic programs for safety checks at each step introduces substantial computational complexity [94]. This requirement challenges the maintenance of real-time performance while ensuring safety, especially in scenarios demanding rapid decision-making.

The computational complexity related to real-time nonmyopic planning, as highlighted in robotic planning under spatiotemporal uncertainty, indicates significant scalability and computational demands [95]. Such complexity can hinder the deployment of multi-modal models in environments requiring quick adaptation and decision-making.

Moreover, the computational overhead linked to Monte Carlo Tree Search (MCTS) planning in MAZero may necessitate considerable resources in highly complex environments [28]. This overhead can restrict the applicability of such models in scenarios with limited computational resources.

The scalability of models like DERL is challenged by the substantial computational resources required to evaluate diverse agent morphologies, emphasizing the need for efficient resource allocation and management [19]. While some approaches perform well in various scenarios, they may falter in environments with unpredictable dynamics or where the model fails to capture true underlying processes [14].

To effectively address the challenges of scalability and computational complexity in deploying multi-modal models, it is essential to explore innovative strategies that enhance resource utilization and optimize learning algorithm efficiency. This includes leveraging advanced techniques such as Multimodal Chain of Thought (M-COT) and Multimodal Instruction Tuning (M-IT) to improve model performance across diverse data modalities, alongside developing standardized evaluation frameworks for meaningful comparisons among multimodal agents. Additionally, integrating embodied intelligence and 3D generation can further enhance model capabilities, facilitating their application in real-world scenarios and advancing the pursuit of artificial general intelligence [55, 83, 93, 45, 35]. By overcoming these challenges, the potential of multi-modal models to enhance the intelligence, adaptability, and resilience of robotic systems can be fully realized, paving the way for advanced autonomous systems.

7.3 Future Research Directions

Future research in multi-modal models for robotics and autonomous systems should prioritize addressing current challenges to enhance model robustness, adaptability, and scalability across diverse applications. A key focus will be on scaling models and improving retrieval strategies to bolster the robustness and generalization of autonomous driving systems, particularly in light of data scarcity issues [2]. Enhancing collision detection mechanisms and integrating advanced computer vision techniques will be crucial for improving interactions and testing in more complex real-world scenarios [7].

Exploring novel methods for data collection and enhancing real-time performance are essential avenues for future research, alongside developing robust safety evaluation frameworks [27]. The application of LARM in real-world scenarios warrants further exploration, emphasizing the importance of training with larger and more diverse datasets to improve robustness and adaptability [5].

Research may also focus on enhancing the adaptability of the Versatile Instructable Motion (VIM) framework to more complex terrains and tasks, thereby improving the robustness of learned skills in real-world applications [29]. Developing versatile robotic systems capable of learning in various environments and tasks, along with establishing robust evaluation frameworks, will be critical for advancing the field [6].

Furthermore, future research should aim to create more interactive environments, refine navigation tasks, and enhance the understanding of natural language in navigation contexts to improve the effectiveness of multi-modal models in real-world applications [10]. By addressing these areas, future research will play a pivotal role in overcoming existing challenges and enhancing the capabilities of multi-modal models, paving the way for more intelligent, resilient, and versatile autonomous systems.

8 Conclusion

Multi-modal large models represent a pivotal advancement in robotics and autonomous systems, offering enhanced capabilities in decision-making, perception, and interaction through the integration of diverse data modalities and learning techniques. These models exemplify their transformative potential by effectively managing complex environments, as demonstrated in autonomous driving tasks where they mitigate uncertainty and enhance performance. The development of platforms like SMARTS for multi-agent reinforcement learning further underscores the importance of these models in shaping future research trajectories.

In the realm of human-robot interaction, the integration of large language models into collaborative frameworks marks a significant leap forward in task planning and execution, with ongoing efforts

aimed at refining teleoperation and improving task success rates. Such advancements highlight the promise of multi-modal large models in creating more robust and efficient human-robot collaboration scenarios.

The conclusion emphasizes the critical need for continued research into multi-modal large models, as they are set to revolutionize autonomous systems across a range of applications. By addressing current challenges and embracing innovative approaches, future research has the potential to fully harness these models, leading to the development of more intelligent, adaptable, and resilient robotic systems capable of operating effectively in complex and dynamic environments.

References

[1] Gilhyun Ryou, Geoffrey Wang, and Sertac Karaman. Multi-fidelity reinforcement learning for time-optimal quadrotor re-planning, 2024.

[2] Jianhao Yuan, Shuyang Sun, Daniel Omeiza, Bo Zhao, Paul Newman, Lars Kunze, and Matthew Gadd. Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model. *arXiv preprint arXiv:2402.10828*, 2024.

[3] Ming Zhou, Jun Luo, Julian Villella, Yaodong Yang, David Rusu, Jiayu Miao, Weinan Zhang, Montgomery Alban, Iman Fadakar, Zheng Chen, Aurora Chongxi Huang, Ying Wen, Kimia Hassanzadeh, Daniel Graves, Dong Chen, Zhengbang Zhu, Nhat Nguyen, Mohamed Elsayed, Kun Shao, Sanjeevan Ahilan, Baokuan Zhang, Jiannan Wu, Zhengang Fu, Kasra Rezaee, Peyman Yadmellat, Mohsen Rohani, Nicolas Perez Nieves, Yihan Ni, Seyedershad Banijamali, Alexander Cowen Rivers, Zheng Tian, Daniel Palenicek, Haitham bou Ammar, Hongbo Zhang, Wulong Liu, Jianye Hao, and Jun Wang. Smarts: Scalable multi-agent reinforcement learning training school for autonomous driving, 2020.

[4] Jiaping Xiao, Rangya Zhang, Yuhang Zhang, and Mir Feroskhan. Vision-based learning for drones: A survey, 2024.

[5] Zhuoling Li, Xiaogang Xu, Zhenhua Xu, SerNam Lim, and Hengshuang Zhao. Larm: Large auto-regressive model for long-horizon embodied intelligence, 2025.

[6] Nicholas Roy, Ingmar Posner, Tim Barfoot, Philippe Beaudoin, Yoshua Bengio, Jeannette Bohg, Oliver Brock, Isabelle Depatie, Dieter Fox, Dan Koditschek, Tomas Lozano-Perez, Vikash Mansinghka, Christopher Pal, Blake Richards, Dorsa Sadigh, Stefan Schaal, Gaurav Sukhatme, Denis Therien, Marc Toussaint, and Michiel Van de Panne. From machine learning to robotics: Challenges and opportunities for embodied intelligence, 2021.

[7] Norberto Sousa, Nuno Oliveira, and Isabel Praça. A multi-agent system for autonomous mobile robot coordination, 2021.

[8] Amrith Setlur, Katie Kang, Aviral Kumar, Feryal Behbahani, Roberta Raileanu, and Rishabh Agarwal. Self-improving foundation models without human supervision. In *ICLR 2025 Workshop Proposals*.

[9] Tianchen Ji, Roy Dong, and Katherine Driggs-Campbell. Traversing supervisor problem: An approximately optimal approach to multi-robot assistance, 2022.

[10] Fengda Zhu, Yi Zhu, Vincent Lee, Xiaodan Liang, and Xiaojun Chang. Deep learning for embodied vision navigation: A survey. *arXiv preprint arXiv:2108.04097*, 2021.

[11] Andrea Tagliabue, Kota Kondo, Tong Zhao, Mason Peterson, Claudius T. Tewari, and Jonathan P. How. Real: Resilience and adaptation using large language models on autonomous aerial robots, 2023.

[12] S. Reza Ahmadzadeh. Research report – persistent autonomy and robot learning lab, 2023.

[13] Zhiyuan Zhou, Pranav Atreya, Abraham Lee, Homer Walke, Oier Mees, and Sergey Levine. Autonomous improvement of instruction following skills via foundation models, 2024.

[14] Danijar Hafner. *Embodied Intelligence Through World Models*. PhD thesis, University of Toronto (Canada), 2024.

[15] Jithin Jagannath, Anu Jagannath, Sean Furman, and Tyler Gwin. Deep learning and reinforcement learning for autonomous unmanned aerial systems: Roadmap for theory to deployment, 2020.

[16] Ralf Der and Georg Martius. Self-organized control for musculoskeletal robots, 2016.

[17] Fei Ye, Shen Zhang, Pin Wang, and Ching-Yao Chan. A survey of deep reinforcement learning algorithms for motion planning and control of autonomous vehicles, 2021.

[18] Linh Kästner, Bassel Fatloun, Zhengcheng Shen, Daniel Gawrisch, and Jens Lambrecht. Human-following and -guiding in crowded environments using semantic deep-reinforcement-learning for mobile service robots, 2022.

[19] Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei. Embodied intelligence via learning and evolution, 2021.

[20] Haokun Liu, Yaonan Zhu, Kenji Kato, Izumi Kondo, Tadayoshi Aoyama, and Yasuhisa Hasegawa. Llm-based human-robot collaboration framework for manipulation tasks, 2023.

[21] Jin Wang and Nikos Tsagarakis. Grounding language models in autonomous loco-manipulation tasks, 2024.

[22] Yuxiang Cui, Haodong Zhang, Yue Wang, and Rong Xiong. Learning world transition model for socially aware robot navigation, 2020.

[23] Dhruv Shah, Michael Equi, Blazej Osinski, Fei Xia, Brian Ichter, and Sergey Levine. Navigation with large language models: Semantic guesswork as a heuristic for planning, 2023.

[24] Lili Fan, Chao Guo, Yonglin Tian, Hui Zhang, Jun Zhang, and Fei-Yue Wang. Sora for foundation robots with parallel intelligence: Three world models, three robotic systems. *Frontiers of Information Technology & Electronic Engineering*, 25(7):917–923, 2024.

[25] Sheuli Paul. A survey of technologies supporting design of a multimodal interactive robot for military communication. *Journal of Defense Analytics and Logistics*, 7(2):156–193, 2023.

[26] Jinzhou Lin, Han Gao, Xuxiang Feng, Rongtao Xu, Changwei Wang, Man Zhang, Li Guo, and Shibiao Xu. Advances in embodied navigation using large language models: A survey, 2024.

[27] Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, et al. Foundation models in robotics: Applications, challenges, and the future. *The International Journal of Robotics Research*, page 02783649241281508, 2023.

[28] Qihan Liu, Jianing Ye, Xiaoteng Ma, Jun Yang, Bin Liang, and Chongjie Zhang. Efficient multi-agent reinforcement learning by planning, 2024.

[29] Ruihan Yang, Zhuoqun Chen, Jianhan Ma, Chongyi Zheng, Yiyu Chen, Quan Nguyen, and Xiaolong Wang. Generalized animal imitator: Agile locomotion with versatile motion prior, 2024.

[30] Henry Zhu, Justin Yu, Abhishek Gupta, Dhruv Shah, Kristian Hartikainen, Avi Singh, Vikash Kumar, and Sergey Levine. The ingredients of real-world robotic reinforcement learning, 2020.

[31] Rachmad Vidya Wicaksana Putra, Alberto Marchisio, Fakhreddine Zayer, Jorge Dias, and Muhammad Shafique. Embodied neuromorphic artificial intelligence for robotics: Perspectives, challenges, and research development stack, 2024.

[32] Carlotta Langer and Nihat Ay. Outsourcing control requires control complexity, 2024.

[33] Yang Liu, Xinshuai Song, Kaixuan Jiang, Weixing Chen, Jingzhou Luo, Guanbin Li, and Liang Lin. Meia: Multimodal embodied perception and interaction in unknown environments, 2024.

[34] Ana Pervan and Todd D. Murphey. Algorithmic design for embodied intelligence in synthetic cells, 2020.

[35] Jonathan Francis, Nariaki Kitamura, Felix Labelle, Xiaopeng Lu, Ingrid Navarro, and Jean Oh. Core challenges in embodied vision-language planning. *Journal of Artificial Intelligence Research*, 74:459–515, 2022.

[36] Shuyuan Liu, Jiawei Chen, Shouwei Ruan, Hang Su, and Zhaoxia Yin. Exploring the robustness of decision-level through adversarial attacks on llm-based embodied models, 2024.

[37] Janderson Ferreira, Agostinho A. F. Júnior, Letícia Castro, Yves M. Galvão, Pablo Barros, and Bruno J. T. Fernandes. Analysis of social robotic navigation approaches: Cnn encoder and incremental learning as an alternative to deep reinforcement learning, 2020.

[38] Alan F. T. Winfield. How intelligent is your intelligent robot?, 2017.

[39] Boris Ivanovic, Amine Elhafsi, Guy Rosman, Adrien Gaidon, and Marco Pavone. Mats: An interpretable trajectory forecasting representation for planning and control, 2021.

[40] Ziyuan Liu and Georg von Wichert. A generalizable knowledge framework for semantic indoor mapping based on markov logic networks and data driven mcmc, 2020.

[41] Youzhi Qu, Chen Wei, Penghui Du, Wenxin Che, Chi Zhang, Wanli Ouyang, Yatao Bian, Feiyang Xu, Bin Hu, Kai Du, Haiyan Wu, Jia Liu, and Quanying Liu. Integration of cognitive tasks into artificial general intelligence test for large models, 2024.

[42] Wenxiao Zhang, Xiangrui Kong, Conan Dewitt, Thomas Braunl, and Jin B. Hong. A study on prompt injection attack against llm-integrated mobile robotic systems, 2024.

[43] Mengdi Xu, Peide Huang, Wenhao Yu, Shiqi Liu, Xilun Zhang, Yaru Niu, Tingnan Zhang, Fei Xia, Jie Tan, and Ding Zhao. Creative robot tool use with large language models, 2023.

[44] Jiaqi Wang, Zihao Wu, Yiwei Li, Hanqi Jiang, Peng Shu, Enze Shi, Huawen Hu, Chong Ma, Yiheng Liu, Xuhui Wang, Yincheng Yao, Xuan Liu, Huaqin Zhao, Zhengliang Liu, Haixing Dai, Lin Zhao, Bao Ge, Xiang Li, Tianming Liu, and Shu Zhang. Large language models for robotics: Opportunities, challenges, and perspectives, 2024.

[45] Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116*, 2024.

[46] Sheng Luo, Wei Chen, Wanxin Tian, Rui Liu, Luanxuan Hou, Xiubao Zhang, Haifeng Shen, Ruiqi Wu, Shuyi Geng, Yi Zhou, et al. Delving into multi-modal multi-task foundation models for road scene understanding: From learning paradigm perspectives. *IEEE Transactions on Intelligent Vehicles*, 2024.

[47] Sai Vemprala, Shuhang Chen, Abhinav Shukla, Dinesh Narayanan, and Ashish Kapoor. Grid: A platform for general robot intelligence development. *arXiv preprint arXiv:2310.00887*, 2023.

[48] Lingyu Xiao, Jiang-Jiang Liu, Sen Yang, Xiaofan Li, Xiaoqing Ye, Wankou Yang, and Jingdong Wang. Learning multiple probabilistic decisions from latent world model in autonomous driving, 2024.

[49] Yong Qi, Gabriel Kyebambo, Siyuan Xie, Wei Shen, Shenghui Wang, Bitao Xie, Bin He, Zhipeng Wang, and Shuo Jiang. Safety control of service robots with llms and embodied knowledge graphs, 2024.

[50] Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Yuqing Wen, Chi Zhang, Xiangyu Zhang, and Tiancai Wang. Adriver-i: A general world model for autonomous driving, 2023.

[51] Luca Carlone, Kasra Khosoussi, Vasileios Tzoumas, Golnaz Habibi, Markus Ryll, Rajat Talak, Jingnan Shi, and Pasquale Antonante. Visual navigation for autonomous vehicles: An open-source hands-on robotics course at mit, 2022.

[52] Ge Sun, Milad Shafiee, Peizhuo Li, Guillaume Bellegarda, Auke Ijspeert, and Guillaume Sartoretti. Learning-based hierarchical control: Emulating the central nervous system for bio-inspired legged robot locomotion, 2024.

[53] Ivan Kapelyukh, Vitalis Vosylius, and Edward Johns. Dall-e-bot: Introducing web-scale diffusion models to robotics, 2023.

[54] Gopika Ajaykumar and Chien-Ming Huang. Multimodal robot programming by demonstration: A preliminary exploration, 2023.

[55] Xinji Mai, Zeng Tao, Junxiong Lin, Haoran Wang, Yang Chang, Yanlan Kang, Yan Wang, and Wenqiang Zhang. From efficient multimodal models to world models: A survey, 2024.

[56] Huan Xu, Jinlin Wu, Guanglin Cao, Zhen Chen, Zhen Lei, and Hongbin Liu. Transforming surgical interventions with embodied intelligence for ultrasound robotics, 2024.

[57] Peiqi Liu, Yaswanth Orru, Jay Vakil, Chris Paxton, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. Ok-robot: What really matters in integrating open-knowledge models for robotics, 2024.

[58] Josua Spisak, Matthias Kerzel, and Stefan Wermter. Clarifying the half full or half empty question: Multimodal container classification, 2023.

[59] Nestor Gonzalez Lopez, Yue Leire Erro Nuin, Elias Barba Moral, Lander Usategui San Juan, Alejandro Solano Rueda, Víctor Mayoral Vilches, and Risto Kojcev. gym-gazebo2, a toolkit for reinforcement learning using ros 2 and gazebo, 2019.

[60] Kevin Frans, Jonathan Ho, Xi Chen, Pieter Abbeel, and John Schulman. Meta learning shared hierarchies, 2017.

[61] Manuel Bianchi Bazzi, Asad Ali Shahid, Christopher Agia, John Alora, Marco Forgione, Dario Piga, Francesco Braghin, Marco Pavone, and Loris Roveda. Robomorph: In-context meta-learning for robot dynamics modeling, 2024.

[62] Linji Wang, Zifan Xu, Peter Stone, and Xuesu Xiao. Grounded curriculum learning, 2024.

[63] Jaeuk Shin, Astghik Hakobyan, Mingyu Park, Yeoneung Kim, Gihun Kim, and Insoon Yang. Infusing model predictive control into meta-reinforcement learning for mobile robots in dynamic environments, 2022.

[64] Gaurav Bagwe, Jian Li, Xiaoyong Yuan, and Lan Zhang. Towards robust on-ramp merging via augmented multimodal reinforcement learning, 2022.

[65] Yongchao Chen, Jacob Arkin, Yang Zhang, Nicholas Roy, and Chuchu Fan. Scalable multi-robot collaboration with large language models: Centralized or decentralized systems?, 2024.

[66] Daniel Bogdoll, Lukas Bosch, Tim Joseph, Helen Gremmelmaier, Yitian Yang, and J. Marius Zöllner. Exploring the potential of world models for anomaly detection in autonomous driving, 2023.

[67] Khimya Khetarpal, Zafarali Ahmed, Gheorghe Comanici, David Abel, and Doina Precup. What can i do here? a theory of affordances in reinforcement learning, 2020.

[68] Yu Tang Liu and Aamir Ahmad. Multi-task reinforcement learning in continuous control with successor feature-based concurrent composition, 2024.

[69] Zhen Meng, Kan Chen, Yufeng Diao, Changyang She, Guodong Zhao, Muhammad Ali Imran, and Branka Vucetic. Task-oriented cross-system design for timely and accurate modeling in the metaverse, 2023.

[70] David Molina Concha, Jiping Li, Haoran Yin, Kyeonghyeon Park, Hyun-Rok Lee, Taesik Lee, Dhruv Sirohi, and Chi-Guhn Lee. Bayesian optimization framework for efficient fleet design in autonomous multi-robot exploration, 2024.

[71] Xuan Xiao, Jiahang Liu, Zhipeng Wang, Yanmin Zhou, Yong Qi, Qian Cheng, Bin He, and Shuo Jiang. Robot learning in the era of foundation models: A survey, 2023.

[72] Xuan Zhong. Building ears for robots: Machine hearing in the age of autonomy, 2023.

[73] Yutao Ouyang, Jinhan Li, Yunfei Li, Zhongyu Li, Chao Yu, Koushil Sreenath, and Yi Wu. Long-horizon locomotion and manipulation on a quadrupedal robot with large language models, 2024.

[74] Abhish Khanal and Gregory J. Stein. Learning augmented, multi-robot long-horizon navigation in partially mapped environments, 2023.

[75] Meraj Mammadov. End-to-end lidar-driven reinforcement learning for autonomous racing, 2023.

[76] Sumeet Singh, Francis McCann Ramirez, Jacob Varley, Andy Zeng, and Vikas Sindhwani. Multiscale sensor fusion and continuous control with neural cdes, 2022.

[77] Shubhankar Gupta and Suresh Sundaram. Decentralized distributed expert assisted learning (d2eal) approach for cooperative target-tracking, 2022.

[78] Maxime Bouton, Alireza Nakhaei, Kikuo Fujimura, and Mykel J. Kochenderfer. Safe reinforcement learning with scene decomposition for navigating complex urban environments, 2019.

[79] Tianyu Shi, Jiawei Wang, Yuankai Wu, Luis Miranda-Moreno, and Lijun Sun. Efficient connected and automated driving system with multi-agent graph reinforcement learning, 2021.

[80] Arturo Gomez Chavez, Qingwen Xu, Christian A. Mueller, Sören Schwertfeger, and Andreas Birk. Adaptive navigation scheme for optimal deep-sea localization using multimodal perception cues, 2019.

[81] Ruchik Mishra, Karla Conn Welch, and Dan O Popa. Human-mediated large language models for robotic intervention in children with autism spectrum disorders, 2024.

[82] Timothée Dhaussy, Bassam Jabaian, and Fabrice Lefèvre. Flowact: A proactive multimodal human-robot interaction system with continuous flow of perception and modular action subsystems, 2025.

[83] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, et al. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision*, 16(1-2):1–214, 2024.

[84] Kenechukwu C. Mbanisi and Michael A. Gennert. Multimodal shared autonomy for social navigation assistance of telepresence robots, 2022.

[85] Jean-François Tremblay, Travis Manderson, Aurélio Noca, Gregory Dudek, and David Meger. Multimodal dynamics modeling for off-road autonomous vehicles, 2021.

[86] Ruiqi Zhang, Jing Hou, Florian Walter, Shangding Gu, Jiayi Guan, Florian Röhrbein, Yali Du, Panpan Cai, Guang Chen, and Alois Knoll. Multi-agent reinforcement learning for autonomous driving: A survey. *arXiv preprint arXiv:2408.09675*, 2024.

[87] Hourui Deng, Hongjie Zhang, Jie Ou, and Chaosheng Feng. Can llm be a good path planner based on prompt engineering? mitigating the hallucination for path planning, 2024.

[88] Jianlin Ye. Llms-enhanced multi-modal navigation in uav systems. Master's thesis, Πανεπιστήμιο Κύπρου, Σχολή Θετικών και Εφαρμοσμένων Επιστημών/University of . . . , 2024.

[89] Naveen Mohan, Martin Törngren, Viacheslav Izosimov, Viktor Kaznov, Per Roos, Johan Svahn, Joakim Gustavsson, and Damir Nesic. Challenges in architecting fully automated driving; with an emphasis on heavy commercial vehicles, 2019.

[90] Pavel Osinenko, Grigory Yaremenko, Roman Zashchitin, Anton Bolychev, Sinan Ibrahim, and Dmitrii Dobriborsci. Critic as lyapunov function (calf): a model-free, stability-ensuring agent, 2024.

[91] Jaime F. Fisac, Monica A. Gates, Jessica B. Hamrick, Chang Liu, Dylan Hadfield-Menell, Malayandi Palaniappan, Dhruv Malik, S. Shankar Sastry, Thomas L. Griffiths, and Anca D. Dragan. Pragmatic-pedagogic value alignment, 2018.

[92] Yuji Cao, Huan Zhao, Yuheng Cheng, Ting Shu, Yue Chen, Guolong Liu, Gaoqi Liang, Junhua Zhao, Jinyue Yan, and Yun Li. Survey on large language model-enhanced reinforcement learning: Concept, taxonomy, and methods. *arXiv preprint arXiv:2404.00282*, 2024.

[93] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.

[94] Mingyu Cai and Cristian-Ioan Vasile. Safety-critical learning of robot control with temporal logic specifications, 2022.

[95] Victoria Preston, Genevieve Flaspohler, Anna P. M. Michel, John W. Fisher III au2, and Nicholas Roy. Robotic planning under uncertainty in spatiotemporal environments in expeditionary science, 2022.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.