Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Empowering Internet of Vehicles with Embodied Intelligence: A Survey of Technologies, Evaluation, and Outlook

Qiang Yu
Guizhou University
Kebing Jin
Guizhou University
Jianhang Tang
Guizhou University
Xuanyu Wu
Guizhou University
Yang Zhang
Nanjing University of Aeronautics and Astronautics
Celimuge Wu
The University of Electro-Communications
Yixiong Feng
yxfeng@gzu.edu.cn

Guizhou University

Research Article

Keywords: Internet of Vehicles, Embodied intelligence, V2X communication, Intelligent transportation

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11344222/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Empowering Internet of Vehicles with Embodied Intelligence: A Survey of Technologies, Evaluation, and Outlook

Qiang Yu[1], Kebing Jin[1], Jianhang Tang[1], Xuanyu Wu[2], Yang Zhang[3], Celimuge Wu[4], Yixiong Feng[2*]

[1]Guizhou Provincial Laboratory of Big Data, State Key Laboratory of Public Big Data, Guizhou University, Guiyang, 550025, Guizhou, China.
[2]School of Mechanical Engineering, Guizhou University, Guiyang, 550025, Guizhou, China.
[3]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China.
[4]Meta-Networking Research Center, The University of Electro-Communications, Tokyo, Japan.

*Corresponding author(s). E-mail(s): yxfeng@gzu.edu.cn;

Abstract

The Internet of Vehicles (IoV) advances intelligent transportation by allowing vehicles, roadside infrastructure, pedestrian devices, and edge and cloud resources to share observations, computing resources, and traffic information. However, the traditional IoV systems often process perception, scene understanding, driving decisions, and vehicle control sequentially without continuously using feedback from vehicle motion execution to update later understanding and decisions, allowing state estimation and decision errors to persist across driving cycles and resulting in outdated vehicle states, decisions inconsistent with traffic conditions, and infeasible or unsafe maneuvers. Embodied intelligence forms a closed-loop in which vehicles perceive traffic conditions, understand scene changes, make decisions, execute control actions, and use environmental feedback to update later perception and decisions. This closed-loop reduces error accumulation, aligns traffic understanding with current road conditions, and improves vehicle action feasibility and adaptability. Existing surveys have reviewed embodied intelligence and related autonomous driving technologies. However, the closed-loop relationships among vehicle-to-everything (V2X) observations, traffic understanding, driving decision-making, vehicle execution, and environmental feedback have not

been systematically examined. This survey organizes relevant studies around intelligent vehicle perception, traffic scene understanding, driving policy generation, and vehicle motion control, and clarifies the closed-loop process in which vehicles integrate multi-source observations collected by sensors to understand the current traffic scene, generate driving policies, control vehicle motion, and incorporate vehicle motion feedback into the next cycle of perception and decision making. The survey also summarizes representative datasets, metrics, challenges, and future directions for connected vehicles.

Keywords: Internet of Vehicles, Embodied intelligence, V2X communication, Intelligent transportation

1 Introduction

The Internet of Vehicles (IoV) underpins safer and more efficient intelligent transportation by enabling coordinated mobility across connected traffic entities. IoV is a heterogeneous cyber–physical networking paradigm connecting vehicles, roadside infrastructure, pedestrian devices, edge–cloud resources, and traffic management services through vehicle-to-everything (V2X) communication (Contreras-Castillo et al. 2018). Cross-entity information exchange extends sensing beyond line of sight, reveals occluded hazards, enables cooperative perception and computation offloading, supports coordinated maneuvers, and improves network-wide traffic management (Zhou et al. 2020). However, packet loss and asynchronous updates delay hazard detection in dynamic traffic environments. Short observation windows erase interaction history, weakening trend estimation and decision consistency. Limited execution feedback reduces coordination reliability and motion feasibility during closed-loop operation.

Embodied intelligence is essential for autonomous systems because reliable adaptation requires continuous coupling between environmental understanding and physical action. Embodied intelligence characterizes physically grounded agents capable of perceiving changing environments, reasoning over observations, executing actions, and learning from subsequent feedback (Sun et al. 2024). Closed-loop interaction connects perception, cognition, decision-making, execution, and feedback rather than treating individual functions as isolated stages. Task-directed perception prioritizes critical changes, while persistent memory links current observations with previous interactions and action outcomes. Physical grounding incorporates body dynamics, actuation limits, task constraints, and safety requirements into behavior generation, reducing infeasible actions. Execution feedback supports online error correction and adaptation to unexpected environmental responses. However, isolated embodied agents remain constrained by local sensing, occlusion, computing capacity, and individual experience, limiting environmental coverage, long-horizon prediction, decision consistency, and multi-agent coordination.

Embodied intelligence in IoV integrates networked traffic information with closed-loop perception, cognition, decision-making, execution, and environmental feedback. IoV provides distributed observations, V2X communication, and edge–cloud resources,

while embodied intelligence organizes networked information around goal-directed physical interaction (Ignatious et al. 2023; Wang et al. 2022). Vehicle-centered embodied agents can extend perception beyond onboard sensors and interpret shared observations in relation to current traffic tasks (Zia et al. 2026). Persistent memory connects current traffic states with previous interactions and execution outcomes, supporting trend estimation and decision consistency. Physical grounding incorporates vehicle dynamics, actuator limits, traffic rules, and safety requirements into behavior generation, reducing infeasible or unsafe maneuvers. Execution feedback updates state estimates and behavior strategies after environmental changes, limiting error accumulation across repeated interaction cycles. The integration links system-level connectivity with agent-level closed-loop intelligence, converting shared traffic information into coordinated actions and execution outcomes into updated environmental understanding.

Our survey is positioned as a structured review of embodied intelligence in IoV. Existing studies are distributed across vehicular networking, cooperative perception, traffic cognition, decision-making, and vehicle control, while related surveys generally focus on individual technical directions or adjacent application domains. Our goal is to connect relevant studies within a unified embodied process and clarify the technical scope of embodied intelligence in IoV. The main contributions are summarized as follows.

- **IoV-Oriented Integration.** We clarify the relationship among vehicles, IoV environments, and embodied interaction. Vehicles operate as embodied agents in dynamic traffic environments, while IoV provides the networked context for information exchange, multi-agent interaction, edge-cloud support, and traffic coordination. The resulting perspective connects physical vehicle intelligence with the broader IoV system.
- **Embodied Process Taxonomy.** We organize existing methods according to embodied perception, cognition, decision-making, and execution. The review covers multimodal and cooperative perception, traffic-scene understanding, intention reasoning, memory and world models, driving and coordination decisions, action-reference generation, and vehicle motion control. The taxonomy describes the complete process from environmental observation to physical vehicle action.
- **Datasets, Metrics, and Outlook.** We summarize representative datasets for perception, cognition, decision-making, and execution tasks. Quantitative metrics are further reviewed for perception accuracy, reasoning accuracy, decision quality, and execution quality. Major challenges and future research directions are discussed to identify unresolved issues and potential developments in embodied intelligence for IoV.

The overall organization of the survey is illustrated in Fig. 1. The survey begins with the conceptual foundations, proceeds to the core technical process, and then summarizes supporting resources, evaluation methods, and future research issues. Section II reviews related work and presents the organization of this survey. Section III introduces the foundations of IoV and embodied intelligence and clarifies the relationship between embodied intelligence and networked traffic environments. Section IV reviews

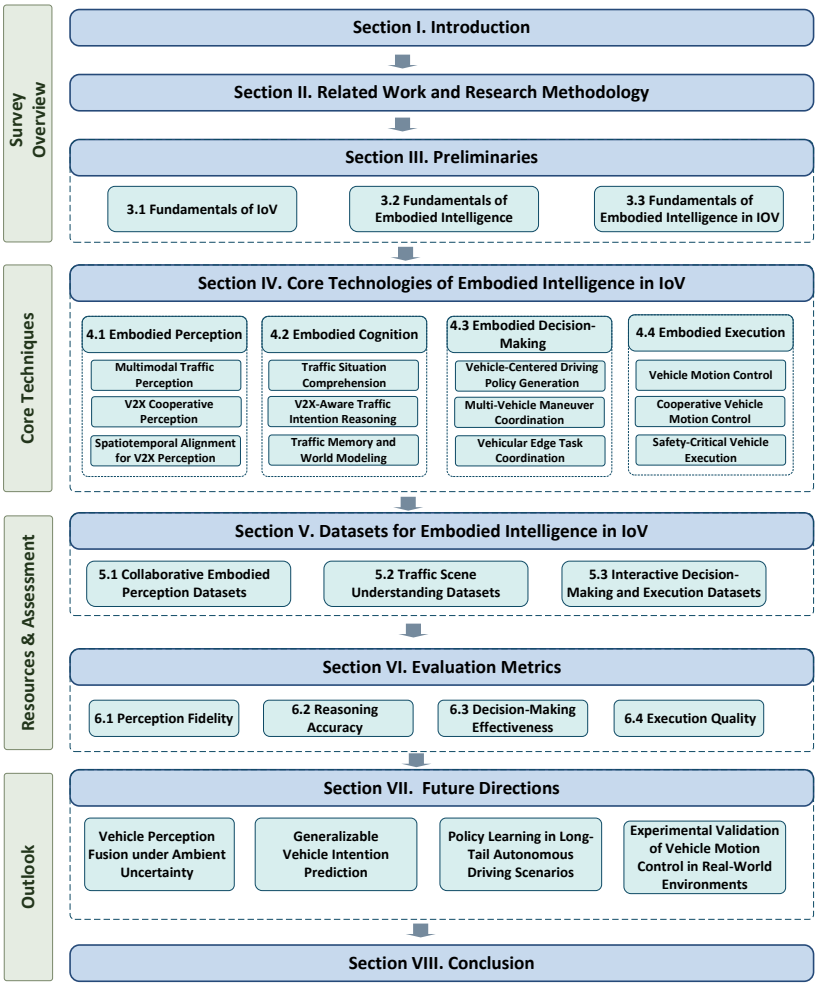


Fig. 1 Overall organization and structure of this survey.

the core techniques according to embodied perception, cognition, decision-making, and execution. Section V summarizes representative datasets for the corresponding embodied tasks. Section VI reviews quantitative metrics for evaluating perception accuracy, reasoning accuracy, decision quality, and execution quality. Section VII discusses major challenges and future research directions. Finally, Section VIII concludes the survey.

2 Related Work and Research Methodology

2.1 Related Work

Existing surveys have examined embodied intelligence from general, technique-oriented, task-oriented, multi-agent, and domain-specific perspectives, as summarized in Table 1. Duan et al. (2022) reviewed embodied artificial intelligence (AI) simulators and representative research tasks, including visual exploration, visual navigation, and embodied question answering, together with associated datasets and evaluation metrics. Sun et al. (2024) summarized embodied intelligence through perception, cognition, and behavior, while discussing representative architectures, development challenges, and future research directions. Liu et al. (2025c) provided a comprehensive survey of embodied robots, simulators, embodied perception, embodied interaction, embodied agents, and sim-to-real adaptation in the era of multimodal foundation models. Xu et al. (2024) reviewed foundation models for robotics, with emphasis on autonomous manipulation, high-level planning, low-level control, datasets, simulators, and benchmarks. Ma et al. (2026) surveyed vision-language-action models by analyzing model architectures, training paradigms, task planning, action generation, and representative embodied applications. Zheng et al. (2025c) examined object-centric robotic manipulation from the perspectives of embodied perception learning, policy learning, task learning, datasets, and evaluation protocols. Wong et al. (2025) reviewed physics simulators for robotic navigation and manipulation, covering simulation platforms, hardware requirements, benchmark tasks, evaluation metrics, and sim-to-real transfer. Liu et al. (2025b) surveyed embodied AI in healthcare across clinical intervention, daily care, infrastructure support, and biomedical research, while discussing enabling technologies, datasets, safety, and ethical concerns. Lisondra et al. (2026) reviewed foundation-model-enabled embodied AI for mobile service robots, focusing on multimodal perception, language-conditioned control, adaptive task execution, and human–robot interaction. Feng et al. (2026) summarized multi-agent embodied AI from the perspectives of environment interaction, agent communication, collaborative learning, joint decision-making, benchmarks, and evaluation methods. Xu et al. (2025) organized embodied intelligence for flexible manufacturing through the industrial eye, industrial hand, and industrial brain, covering environmental perception, precision manipulation, process planning, and production scheduling.

Among the existing surveys, Hu et al. (2026) is the most closely related to the present work. Hu et al. organize multi-agent embodied autonomous driving around the transition from V2X information exchange to shared world models, emphasizing state, intent, and plan alignment for coordinated action. Different from this representation-centered perspective, the present survey takes the closed-loop embodied process as its primary organizing framework. The distinction lies in the level of analysis: a shared world model focuses on the common representation through which multiple agents align states, intentions, and plans for coordination, whereas the closed-loop perspective follows how environmental information is transformed through perception, cognition, and decision-making into physical vehicle actions, and how the resulting environmental changes influence subsequent interaction. This difference is further reflected in the technical organization of the two surveys. In the present taxonomy, world modeling is

positioned within embodied cognition, while physical execution is treated as a peer-level technical stage covering vehicle motion control, cooperative motion control, and safety-critical control.

Overall, existing surveys have established broad foundations across embodied perception, cognition, decision-making, physical interaction, multi-agent collaboration, and domain-specific applications. However, a process-level synthesis of embodied intelligence in IoV remains limited: the connections through which networked perception and traffic cognition lead to vehicle decision-making and physical execution, and through which execution outcomes feed back into subsequent interaction, have not been systematically organized within a unified vehicle-centered taxonomy. The present survey is positioned around this closed-loop perspective and organizes the relevant literature according to perception, cognition, decision-making, execution, and environmental feedback. The differences between the reviewed surveys and the present survey across major dimensions are summarized in Table 1.

2.2 Research methodology

(1)Search strategy: To identify studies relevant to embodied intelligence in the Internet of Vehicles (IoV), we conducted a literature search covering publications from 2018 to 2026 in Web of Science, IEEE Xplore, SpringerLink, and Google Scholar. Broad search terms related to embodied intelligence, Internet of Vehicles, V2X, connected and autonomous vehicles, and autonomous driving were first used to define the research scope. The search was subsequently refined using representative terms corresponding to the four technical stages considered in this survey, including cooperative perception and sensor fusion for perception, scene understanding and intention prediction for cognition, driving decision-making, behavior planning, and maneuver coordination for decision-making, and motion control, trajectory tracking, and cooperative control for execution. These stage-specific terms were further cross-combined with established AI and machine-learning terms, such as deep learning, reinforcement learning, deep reinforcement learning, and multi-agent reinforcement learning, to identify AI-driven studies relevant to each stage.

(2)Eligibility criteria: Studies were included when they contributed to perception, cognition, decision-making, execution, or interactions among these stages in IoV and related traffic environments. Eligible studies reported technical methods, theoretical formulations, experimental evaluations, datasets, benchmarks, or evaluation approaches relevant to the survey scope. Peer-reviewed journal and conference papers were prioritized, while recent preprints were retained for rapidly developing topics with limited peer-reviewed literature. Earlier seminal studies were additionally considered when needed to establish foundational concepts, classical methods, representative datasets, or historical development. Studies were excluded when they were unrelated to vehicular or traffic environments, focused solely on communication, networking, computation, or resource management without a direct connection to the embodied process, lacked sufficient technical relevance, or duplicated a more complete version of the same work.

Table 1 Comparison of the present survey with representative embodied-intelligence surveys

Related work	Year	Domain	Main focus	Organization	DS	EM
Duan et al. (2022)	2022	General	Embodied AI simulators and research tasks	Simulators; visual exploration; visual navigation; embodied QA	✓	✓
Sun et al. (2024)	2024	General	Perception, cognition, and behavior integration	PCB-body and environment framework	–	✓
Liu et al. (2025c)	2025	General	Embodied AI across cyber and physical environments	Embodied perception; interaction; agents; sim-to-real adaptation	✓	–
Xu et al. (2024)	2024	Robotics	Foundation models for robotic manipulation	High-level planning; low-level control	✓	–
Ma et al. (2026)	2026	General	VLA models for embodied AI	VLA components; control policies; task planners	✓	–
Zheng et al. (2025c)	2025	Robotics	Embodied learning for object-centric robotic manipulation	Perceptual learning; policy learning; task-oriented learning	✓	✓
Wong et al. (2025)	2025	Robotics	Physics simulators for robotic navigation and manipulation	Navigation; manipulation	✓	✓
Liu et al. (2025b)	2025	Healthcare	Embodied AI capabilities and applications in healthcare	Brain capabilities; healthcare applications; intelligence levels	✓	–
Lisondra et al. (2026)	2026	Robotics	Foundation models for mobile service robots	Language-to-action; multimodal perception; uncertainty; computation	–	–
Feng et al. (2026)	2026	General	Multi-agent embodied AI	Single-agent foundations; multi-agent methods; benchmarks	–	–
Xu et al. (2025)	2025	Manufacturing	Embodied intelligence for flexible manufacturing	Industrial Eye; Industrial Hand; Industrial Brain	–	–
Hu et al. (2026)	2026	Autonomous driving	Shared world models for multi-agent embodied autonomous driving	V2X exchange → shared world models → coordinated action	✓	–
This work	2026	Internet of Vehicles	**Closed-loop mechanisms of embodied intelligence for IoV**	**Perception → cognition → decision-making → execution in a closed loop**	✓	✓

Note: ✓ indicates that the corresponding topic is systematically reviewed, whereas – indicates that it is not systematically covered. "General" denotes surveys that are not restricted to a specific application domain. QA: question answering; PCB: perception–cognition–behavior; VLA: vision-language-action; V2X: vehicle-to-everything; IoV: Internet of Vehicles; DS: Datasets; EM: Evaluation Metric.

(3)Screening and literature organization: The retrieved literature was screened according to the above criteria. Duplicate records were first removed, followed by title and abstract screening to exclude studies outside the survey scope. The remaining studies were assessed through full-text review to confirm their relevance to the four embodied stages or their cross-stage interactions. When necessary, backward and forward citation checking was used to identify representative foundational studies and recent related work not captured by the initial search. Finally, the retained literature was organized according to the four-stage taxonomy adopted in this survey, namely perception, cognition, decision-making, and execution, while datasets and evaluation methods were reviewed separately. Studies involving multiple stages were classified according to their primary technical focus, with relevant cross-stage connections discussed where appropriate.

3 Preliminaries

3.1 Fundamentals of Internet of Vehicles

1) Internet of Vehicles Background: Vehicular ad hoc networks (VANETs) are generally regarded as an early vehicular communication paradigm extended from mobile ad hoc networks (MANETs) to road transportation scenarios. In VANETs, vehicles act as highly mobile nodes to exchange local road information and support short-range traffic coordination. However, high mobility, dynamic topology, intermittent connectivity, and limited coverage restrict their support for large-scale and persistent intelligent transportation services (Hartenstein and Laberteaux 2008). To overcome these limitations, Internet of Things (IoT) technologies extend vehicular systems with broader sensing, communication, and computing capabilities, while edge and cloud computing further provide distributed storage, data processing, and service support (Lin et al. 2017; Liu et al. 2021).

2) Internet of Vehicles Definition: The IoT evolution has promoted the emergence of the Internet of Vehicles (IoV) as a heterogeneous cyber–physical networking paradigm for intelligent transportation. IoV connects vehicles with roadside infrastructure, pedestrian devices, edge–cloud platforms, and traffic management systems, integrating communication, sensing, computation, and service coordination rather than merely extending vehicular communication protocols (Contreras-Castillo et al. 2018). Vehicle-to-everything (V2X) communication provides the fundamental connectivity among these heterogeneous entities and is commonly categorized into vehicle-to-vehicle (V2V), vehicle-to-infrastructure (V2I), vehicle-to-pedestrian (V2P), and vehicle-to-network (V2N) communication, corresponding to interactions among vehicles, roadside infrastructure, vulnerable road users, and network-side platforms, respectively (Garcia-Roger et al. 2020). Other application-oriented expressions, such as vehicle-to-roadside (V2R), vehicle-to-sensor (V2S), and vehicle-to-cloud (V2C), can generally be incorporated into V2I or V2N according to the communicating entity.

3) Internet of Vehicles Architecture: Beyond individual communication modes, IoV also requires a layered architecture to organize heterogeneous traffic entities, network resources, computing capabilities, and intelligent services. Existing studies have proposed different architectural abstractions for IoV. Kaiwartya et al. (2016) describe a five-layer architecture consisting of a perception layer, coordination layer, artificial intelligence layer, application layer, and business layer. Contreras-Castillo et al. (2018) analyze IoV from connected vehicles, roadside infrastructure, communication protocols, cloud services, and security mechanisms. Vehicular edge computing (VEC) studies commonly use a vehicle-edge-cloud architecture to describe distributed computation, storage, and service support (Liu et al. 2021). Drawing on the studies, the survey adopts a synthesized IoV architecture with five functional components: perception, coordination and communication, edge-cloud intelligence, application and service, and management and governance. The overall structure is illustrated in Fig. 2, and the functions of the five layers are summarized as follows.

- **Perception layer:** The perception layer consists of vehicles, onboard sensors, positioning devices, roadside sensors, roadside units, and pedestrian terminals, providing physical-world information for environmental perception and vehicle-state monitoring.
- **Coordination and communication layer:** The coordination and communication layer connects heterogeneous traffic entities through V2X communication and supports message delivery, network access, and information exchange.
- **Edge-cloud intelligence layer:** The edge-cloud intelligence layer provides distributed storage, computation, and data analysis, supporting latency-sensitive edge processing and large-scale cloud services.
- **Application and service layer:** The application and service layer transforms sensed and processed information into intelligent transportation services, including cooperative perception, route planning, traffic management, and autonomous driving assistance.
- **Management and governance layer:** The management and governance layer provides cross-layer support for resource management, security, privacy protection, service management, and reliable system operation.

Through the integration of V2X communication modes and layered IoV architecture, IoV provides the networking and system foundation for large-scale intelligent transportation services. However, IoV environments are characterized by high mobility, dynamic topology, heterogeneous sensing and communication resources, time-sensitive safety requirements, and direct coupling with physical traffic processes (Alshammari and Mahgoub 2025). These characteristics require vehicles and infrastructure to continuously perceive changing road conditions, interpret interactive traffic contexts, and generate adaptive behaviors through AI-enabled perception, reasoning, and decision-making. Reliable vehicle intelligence further requires these intelligent processes to be continuously coupled with physical execution and environmental feedback, which motivates the following discussion on embodied intelligence.

3.2 Fundamentals of Embodied Intelligence

1) Embodied Intelligence Background: Intelligent agents have progressively evolved from internal reasoning toward interaction with physical environments. Early agents mainly relied on logical reasoning and symbolic computation, enabling rule-based task processing but lacking real-time environmental interaction (Newell and Simon 1976). Behavior-based robotics subsequently emphasized perception–action coupling through physical embodiment and environmental feedback, but remained constrained by handcrafted rules and reactive control (Brooks 1991). Cognitive reasoning and feedback learning further supported environmental understanding, behavior adjustment, and experience accumulation, contributing to the development of embodied intelligence (Wilson 2002).

2) Embodied Intelligence Definition: Embodied intelligence is conceptualized as a class of physically grounded agents that continuously interact with external environments through perception and action (Pfeifer and Iida 2004; Sun et al. 2024). A typical embodied intelligent system consists of three fundamental elements: brain,

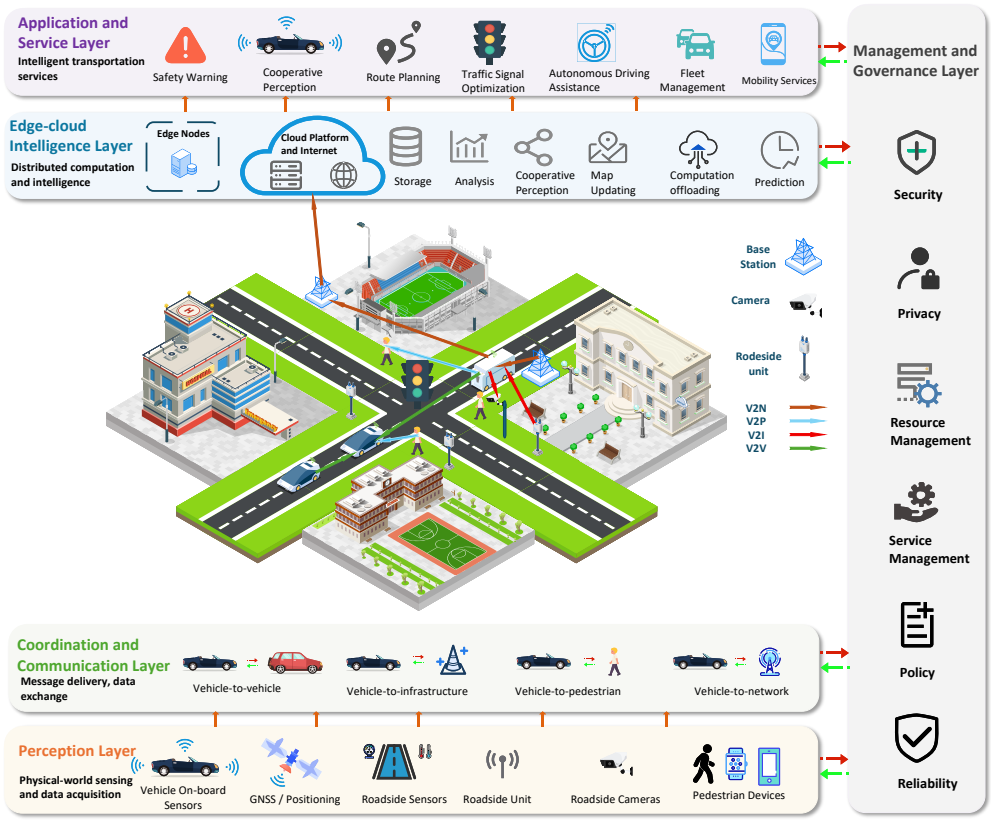


Fig. 2 Layered system architecture of IoV.

body, and environment. The brain represents the internal intelligence mechanism that processes task objectives, environmental states, memory, and feedback for understanding, reasoning, and decision-making. The body integrates sensors and actuators to acquire external information and execute physical actions, while the environment provides dynamic scenarios, interaction targets, and operational constraints.

The three elements do not operate independently but are continuously coupled through perceptual and execution interfaces. The body first acquires multimodal observations from the environment, which are processed by the brain to update the current state understanding and generate a decision strategy. The resulting strategy is then transformed by the body into executable motion or control actions that alter the environment. The resulting environmental changes generate new observations and feedback, which are incorporated into subsequent state updating, strategy adjustment, and behavior generation (Yu et al. 2025b). Through this continuous interaction, embodied intelligence forms a perception–cognition–decision-making–execution closed loop driven by environmental feedback. Compared with non-embodied agents that mainly rely on internal computation, static data processing, or symbolic outputs, embodied agents continuously adapt their understanding and behavior according to

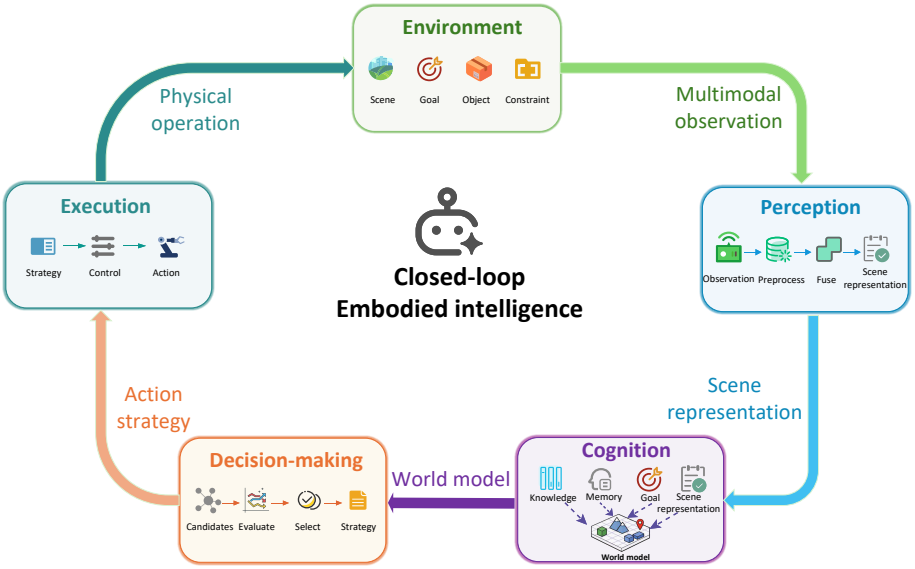


Fig. 3 Closed-loop workflow of embodied intelligence.

physical execution outcomes, thereby improving contextual understanding, execution feasibility, and environmental adaptability (Anderson 2003).

3) Embodied Intelligence Architecture: At the architectural level, a growing body of studies has examined the internal organization and functional coordination mechanisms of embodied intelligence. Specifically, Feng et al. (2026) conceptualize embodied systems as tightly coupled perception-cognition-action modules. Similarly, Sun et al. (2024) characterize dynamic environmental couplings through integrated perception-cognition-action and feedback learning loops. Zhang et al. (2025c) propose a three-layer architecture encompassing multimodal perception and alignment, world modeling and task graph generation, and policy adaptation and execution scheduling. Despite discrepancies regarding architectural nomenclature and partition granularitie across preceding studies, the core logic encompasses environmental perception, internal comprehension, decision-making, and physical execution. Guided by the shared conceptual baseline, this survey formalizes embodied intelligence capabilities into four modules comprising perception, cognition, decision-making, and execution. The perception, cognition, decision-making, and execution modules operate according to the closed-loop workflow shown in Fig. 3, with module-specific functions summarized as follows.

- **Perception module:** Through multi-source sensing, the agent acquires environmental, internal, and interactive states. Multimodal observations are processed and fused into task-oriented scene representations that support subsequent cognition.

- **Cognition module:** Building on perceptual representations, cognition incorporates prior knowledge, memory, task objectives, and physical constraints. Environmental states, scene semantics, interaction relations, and possible future evolution are interpreted to establish task-oriented scene understanding for decision-making.
- **Decision-making module:** Based on the scene understanding established by cognition, candidate behaviors are evaluated under task objectives, operational constraints, and interaction requirements. The evaluation results support the selection of an action strategy for downstream execution.
- **Execution module:** Action strategies are converted into control directives and physical operations. Physical actions change the agent state and surrounding environment, generating feedback for subsequent perception, cognition, and decision-making.

The above four modules outline the basic functional structure of embodied intelligence, in which perception, cognition, decision-making, and execution jointly support the closed-loop transformation from environmental observation to physical operation. However, when this closed-loop mechanism is extended to IoV, embodied intelligence is no longer limited to a single agent-environment interaction, but operates within a distributed traffic ecosystem involving vehicles, roadside infrastructure, traffic participants, communication networks, and cloud-edge resources (Danba et al. 2022). Therefore, the next section further discusses how embodied intelligence can be formulated in IoV environments, where its functional modules are reshaped by V2X connectivity, traffic dynamics, safety constraints, and multi-agent collaboration.

3.3 Fundamentals of Embodied Intelligence in Internet of Vehicles

Through V2X communication, IoV connects vehicles, infrastructure, pedestrians, and networks, providing support for cross-entity information sharing and cooperative services (Zhou et al. 2020). However, as traffic complexity and intelligence demands increase, IoV research has expanded from information transmission and resource sharing to continuous intelligent processes, including environmental understanding, interaction judgment, behavioral decision-making, and physical execution (Malik et al. 2021). The processes show closed-loop coupling among perception, cognition, decision-making, execution, and feedback, consistent with embodied intelligence's agent-environment interaction mechanism. Therefore, extending embodied closed-loop interaction to IoV scenarios helps characterize dynamic cooperation among vehicles, infrastructure, platforms, pedestrians, and environments, promoting IoV evolution from information sharing to closed-loop cooperative intelligence (Wang et al. 2022).

1)Definition of Embodied Intelligence in IoV: This study defines embodied intelligence in IoV as a vehicle-centered intelligent agent that takes the vehicle as the core embodied entity and continuously interacts with dynamic traffic environments. Supported by onboard sensors and V2X communication, the agent acquires ego states, surrounding environments, and external traffic information, understands road structures, traffic states, intentions, and risks, generates and executes safe, rule-compliant

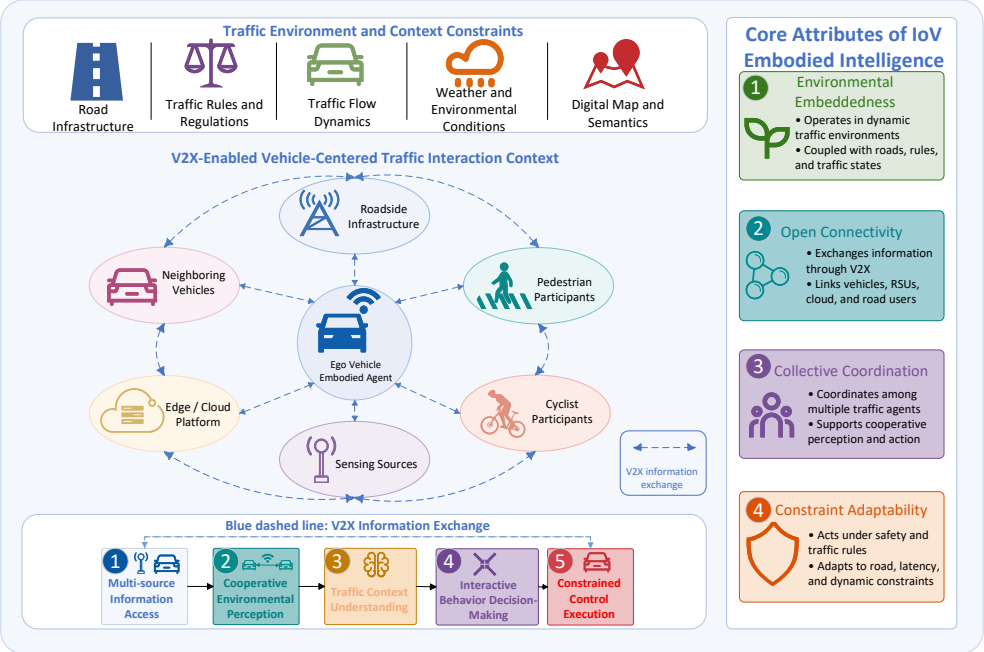


Fig. 4 Vehicle-Centered Embodied Intelligence Framework in IoV.

driving behaviors, and adjusts subsequent decisions through environmental feedback (Ignatious et al. 2023).

IoV scenarios further endow the functional modules of embodied intelligence with specific traffic-system connotations. Compared with general embodied intelligent systems, perception, cognition, decision-making, and execution in IoV operate under the joint influence of V2X connectivity, multi-agent interaction, traffic rules, safety requirements, and latency constraints. As illustrated in Fig. 4, the vehicle-centered embodied agent is embedded in a V2X-enabled traffic interaction context and is shaped by environmental, connectivity, coordination, and constraint-related attributes. Such conditions lead to more distinct scenario-oriented characteristics. Specifically, the scenario-oriented characteristics can be summarized into the following four aspects.

- **Cooperative perception (Cui et al. 2022):** Perception extends beyond local observations from onboard sensors by integrating multi-source information from other vehicles, roadside infrastructure, and network-side resources. Cross-entity information sharing mitigates limited sensing range and occlusion, forming a more complete and stable representation of the traffic environment.
- **Traffic cognition (Guan et al. 2024):** Traffic-oriented cognition organizes and reasons over road structures, traffic rules, agent states, interaction intentions, and potential risks. The resulting traffic-scene understanding provides the cognitive basis for subsequent interaction judgment and behavioral decision-making.

- **Interactive decision-making (Schwarting et al. 2019):** Decision-making extends beyond single-vehicle behavior selection by considering state relations, intention differences, and cooperation requirements among surrounding vehicles, pedestrians, roadside infrastructure, and traffic management systems. Multi-agent interaction judgment supports safer and more cooperative behavior strategies.
- **Traffic-constrained execution (Paden et al. 2016):** Execution converts decision strategies into vehicle motion under real traffic conditions. Physical execution must satisfy vehicle dynamics, road structures, traffic rules, safety distances, and communication latency.

Embodied intelligence in IoV operates as a vehicle-centered closed loop. Onboard sensing and V2X communication continuously provide local and shared traffic information, supporting traffic-scene understanding and behavior decision-making. The resulting decisions are translated into physical vehicle actions that change both the vehicle state and the surrounding traffic environment. These changes are captured again through sensing and communication as feedback for the next interaction cycle. Based on this closed-loop process, Section 4 further reviews the key technologies supporting embodied perception, cognition, decision-making, and execution in IoV, and clarifies how these technical components jointly enable continuous interaction between vehicle-centered embodied agents and dynamic traffic environments.

4 Core Technologies of Embodied Intelligence in Internet of Vehicles

Embodied intelligence in IoV is formed through the continuous coupling of vehicle-centered embodied agents with dynamic traffic environments. To describe the technical realization of the embodied perception–cognition–decision-making–execution closed loop, this section reviews embodied perception, embodied cognition, embodied decision-making, and embodied execution, which respectively address environmental state acquisition, traffic-context understanding, decision strategy generation, and physical execution.

4.1 Embodied Perception in Internet of Vehicles

In IoV, embodied perception constructs the task-oriented perceptual state required by a vehicle-centered embodied agent for embodied closed-loop interaction. Rather than focusing only on object detection, embodied perception transforms onboard observations and V2X messages into representations that support embodied scene cognition, intention reasoning, decision-making, and execution. Accordingly, this section reviews embodied perception from three levels: local multimodal representation, V2X cooperative information acquisition, and spatiotemporal alignment-fusion. Local multimodal representation, V2X cooperative information acquisition, and spatiotemporal alignment jointly convert heterogeneous and distributed traffic observations into an ego-centric scene representation for subsequent embodied cognition and behavior generation. The overall workflow of embodied perception in IoV is illustrated in Fig. 5.

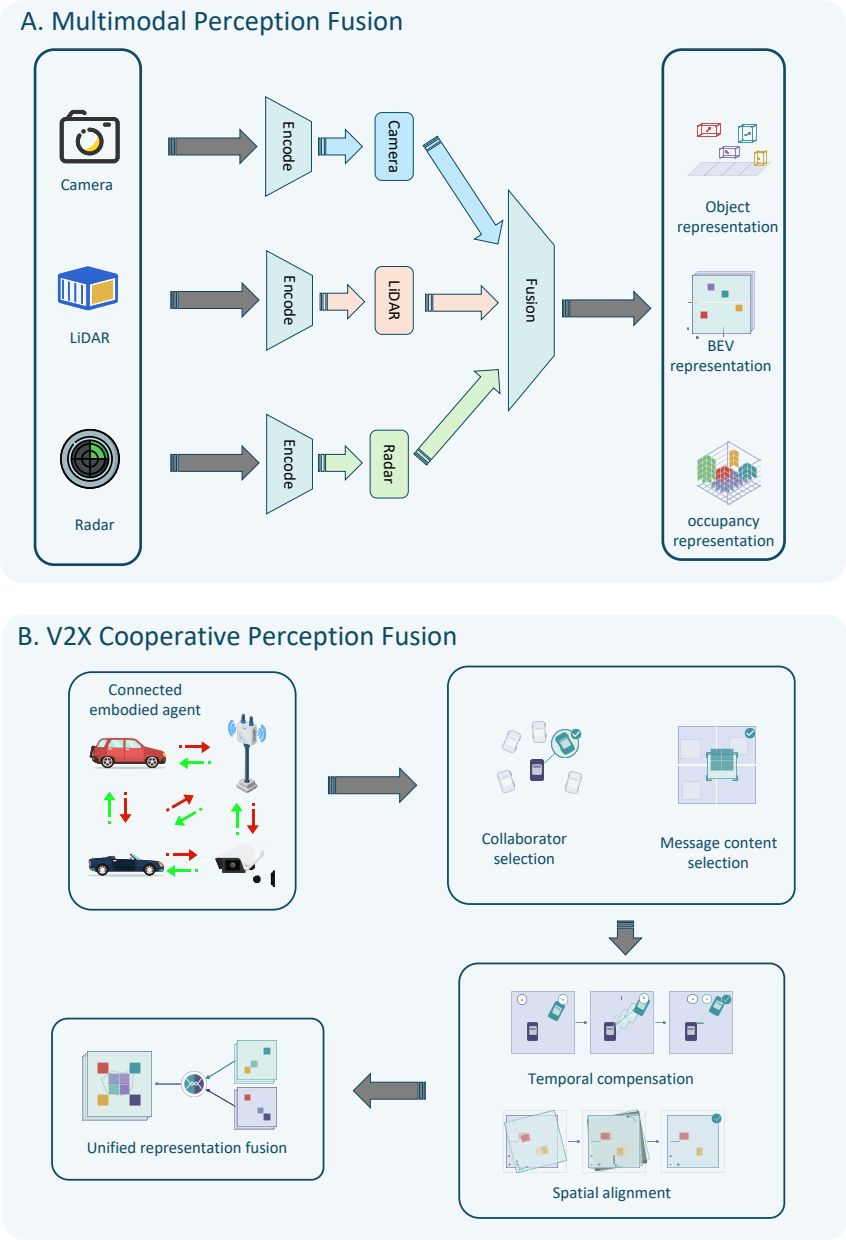


Fig. 5 Embodied Perception Workflow in IoV.

4.1.1 Multimodal Traffic Perception

In embodied perception for IoV, a single sensing modality is insufficient for reliable traffic scene understanding because each sensor has inherent limitations. Embodied multimodal fusion integrates complementary cues from cameras, light detection and ranging (LiDAR), and radar to construct more robust environmental representations (Paek and Kong 2025; Łach 2026). According to the output form of fused representations, existing methods can be grouped into object-level representation, bird's-eye view (BEV) spatial representation, and semantic occupancy representation.

A. Object-level representation focuses on instance-level recognition and state estimation of traffic participants and obstacles. FUTR3D (Chen et al. 2023) uses a modality-agnostic feature sampler to aggregate camera, LiDAR, and radar features according to each query's 3D reference point for 3D object detection. The object feature fusion, with positional encoding (PE), can be summarized as:

$$q_i^{\text{new}} = q_i + PE(c_i) + \Phi_{\text{fus}}\left(s_i^{\text{lid}} \oplus s_i^{\text{cam}} \oplus s_i^{\text{rad}}\right), \tag{1}$$

where q_i denotes the i-th object query, and q_i^{new} denotes the updated query after multimodal fusion. c_i is the 3D reference point of the query, and $PE(c_i)$ provides the corresponding encoding. s_i^{lid}, s_i^{cam}, and s_i^{rad} denote the sampled LiDAR, camera, and radar features, respectively. $\oplus$ denotes feature concatenation, and Φ_{fus} is the fusion multilayer perceptron (MLP) used to obtain the unified query-level representation.

In the query-level fusion pipeline of FUTR3D, multimodal features are sampled around 3D reference points and subsequently integrated into object queries. CMT (Yan et al. 2023a) instead encodes 3D positional information into image and point-cloud tokens, enabling direct multimodal interaction without explicit view transformation. SparseFusion (Xie et al. 2023) further shifts multimodal fusion toward sparse candidates and sparse representations, avoiding the noise and inefficiency of dense candidates and scene representations.

B. BEV spatial representation organizes multimodal information into an ego-centric BEV space. BEVGuide (Man et al. 2023) uses BEV-guided multi-sensor attention to learn a unified BEV representation from sensor-specific features. The process of BEVGuide can be summarized as:

$$\mathbf{z}_i = \text{Fuse}_{m\in\mathcal{M}}\left(\sum_{j=1}^{S_m} \text{Softmax}_j\left(\mathbf{q}_i^{\top}(\mathbf{f}_{m,j} + \mathbf{p}_{m,j})\right)\mathbf{f}_{m,j}\right), \tag{2}$$

where $\mathbf{z}_i$ denotes the fused feature of the i-th BEV grid, $\mathcal{M}$ denotes the sensor set, and $\mathbf{f}_m = E_m(X_m)$ denotes the feature map extracted from the m-th sensor input by a sensor-specific encoder. $\mathbf{f}_{m,j}$, S_m, $\mathbf{q}_i$, and $\mathbf{p}_{m,j}$ denote the j-th feature position, the number of feature positions, the BEV query, and the PE, respectively. The softmax term calculates attention weights between the BEV query and sensor features, while Fuse aggregates attended features from different modalities into a unified BEV representation.

Subsequent studies have further refined multimodal BEV fusion. BEV-MOSNet (Cong et al. 2025) employs deformable cross-attention-guided camera–LiDAR–radar fusion for BEV moving-object segmentation, whereas ASF (Paek and Kong 2025) combines unified canonical projection with patch-wise cross-sensor attention to improve feature consistency and robustness under sensor degradation or failure.

C. Semantic occupancy representation discretizes the space surrounding a vehicle into three-dimensional voxel grids and predicts the occupancy state and semantic category of each voxel, thereby enabling dense 3D modeling of the external environment. OccFusion (Ming et al. 2025) integrates multiscale camera, LiDAR, and radar features through dynamic 2D/3D fusion and global-local attention for 3D semantic occupancy prediction. The process of OccFusion can be summarized as:

$$\mathrm{Occ} = H\Big(\{\mathrm{GLA}(D_{\mathrm{2D}}([F_g^{m,l}]_{m\in\mathcal{M}}), D_{\mathrm{3D}}([F_v^{m,l}]_{m\in\mathcal{M}}))\}_{l=1}^{L}\Big), \tag{3}$$

where Occ denotes the final 3D semantic occupancy grid, $\mathcal{M}$ denotes the set of sensor modalities, m denotes a specific modality, l denotes the feature scale, and L denotes the total number of scales. $F_g^{m,l}$ and $F_v^{m,l}$ represent the BEV global feature and 3D voxel feature of modality m at scale l, respectively. $[\cdot]$ denotes channel-wise concatenation across modalities. D_{2D} and D_{3D} denote the Dynamic Fusion 2D and Dynamic Fusion 3D modules, which combine convolutional feature mixing with SENet-based channel recalibration. GLA denotes global-local attention fusion, and H denotes the final occupancy prediction head.

MS-Occ (Wei et al. 2026) further extends multimodal occupancy fusion across multiple stages by combining middle-stage geometry–semantic enhancement with adaptive and confidence-guided late-stage voxel fusion.

Embodied multimodal perception fusion integrates complementary camera, LiDAR, and radar cues to provide a more reliable local scene representation. However, local fusion is still limited by sensing range, occlusion, and incomplete observations. Therefore, embodied cooperative perception should be further considered to compensate for occlusion and local sensing limitations.

4.1.2 V2X Cooperative Perception

V2X cooperative perception extends vehicle-centered embodied perception by incorporating complementary observations from connected embodied vehicles and roadside infrastructure. Existing studies mainly address two communication-aware problems: collaborator selection and message content selection (Huang et al. 2025b). The former identifies useful external agents, while the latter determines task-relevant perceptual information under bandwidth and latency constraints.

A. Traffic-aware collaborator selection determines which external nodes should communicate with the ego embodied vehicle, thereby reducing the bandwidth cost and redundant interference caused by fully connected collaboration. Who2com (Liu et al. 2020c) uses a learnable handshake mechanism in which a target

agent sends a compact request and selects a collaborator based on returned matching scores. The most valuable collaborator is selected through attention-based matching:

$$\hat{i} = \arg\max_{i} \left(G_m^j(x_j)\right)^T W_a G_k^i(x_i), \tag{4}$$

where x_j and x_i denote the observations of the target embodied vehicle and the candidate node, respectively. G_m^j and G_k^i denote the request generator and the key generator, respectively. W_a is a learnable weight matrix, and $\hat{i}$ denotes the selected collaborator.

However, Who2com assumes that the target vehicle requires external communication. When2com (Liu et al. 2020b) further learns when to communicate and dynamically constructs communication groups. Select2Col (Liu et al. 2024) further incorporates the spatial–temporal importance of semantic information and uses a lightweight graph neural network to identify contributive collaborators.

B. Traffic-aware message content selection determines which perceptual representations should be transmitted to the ego vehicle after collaborators are identified, thereby avoiding redundant message exchange and maintaining task-related scene information under bandwidth and latency constraints. V2VNet (Wang et al. 2020) shares compressed intermediate representations rather than raw sensor data or final outputs, balancing perception–prediction performance with communication bandwidth.

Based on intermediate feature sharing, subsequent studies further reduce redundant feature transmission. To reduce redundancy in intermediate feature sharing, Where2comm (Hu et al. 2022) uses spatial confidence maps to select spatially sparse yet perceptually critical features for communication. How2comm (Yang et al. 2023) further refines feature sparsification through mutual-information-aware spatial–channel filtering to retain informative shared features.

Although spatial and channel selection reduces redundancy, cooperative messages remain tied to BEV feature maps. SparseCoop (Wang et al. 2026) replaces intermediate BEV representations with kinematic-grounded instance queries that encode explicit 3D geometry and velocity for fully sparse cooperative perception. Overall, message content selection shows a progressive refinement from compressed intermediate representations to spatially selected BEV features, spatial-channel semantic features, and sparse instance queries for efficient vehicle-centered embodied perception.

Collaborator selection and message content selection provide the communication basis for V2X cooperative perception by identifying useful agents and selecting compact perceptual messages. However, the selected messages are still generated from different viewpoints, timestamps, and coordinate frames (Lu et al. 2023; Wang et al. 2025c). Therefore, spatiotemporal alignment and fusion are further required to transform distributed messages into a unified ego-centric scene representation (Xu et al. 2022).

4.1.3 Spatiotemporal Alignment for V2X Perception

Distributed observations usually differ in spatial reference, sensing time, and communication delay, resulting in potential misalignment during fusion. Reliable integration

therefore requires spatial calibration and temporal compensation before heterogeneous observations are fused. The alignment-fusion process can be organized into three modules: spatial alignment, temporal compensation, and unified representation fusion.

A. Spatial alignment of traffic representations

Spatial alignment maps collaborative messages from different embodied agents into the ego-vehicle coordinate system before fusion. V2VNet (Wang et al. 2020) aligns received intermediate features to the receiver frame through relative-pose-based spatial warping and bilinear resampling before graph aggregation. For heterogeneous V2X settings, V2X-ViT (Xu et al. 2022) uses a spatial-temporal correction module to warp delayed agent features to the current ego-vehicle pose before Transformer fusion.

The direct pose-based spatial alignment process can be abstracted at both coordinate and feature-map levels:

$$\begin{aligned} p_e &= R_{k\to e} p_k + t_{k\to e}, \\ F_{k\to e}(u) &= F_k\left(G_{e\to k}(u;\xi_{k\to e})\right), \end{aligned} \tag{5}$$

where p_k and p_e denote the same spatial point in the coordinate systems of agent k and the ego vehicle, respectively. $R_{k\to e}$ and $t_{k\to e}$ denote the rotation matrix and translation vector from agent k to the ego vehicle. F_k and $F_{k\to e}$ denote the original and aligned BEV feature maps, respectively. u denotes an ego-view BEV grid location, $\xi_{k\to e}$ denotes the initial relative pose, and $G_{e\to k}$ denotes the pose-induced grid mapping function used for feature warping and resampling.

Direct geometric transformation is sensitive to localization noise, heading error, and calibration drift. Building on the V2VNet communication setting, Vadivelu et al. (2021) estimate pairwise pose errors, enforce global pose consistency, and attenuate unreliable messages during aggregation. Object-structure-based methods further exploit shared objects for pose correction. OptiMatch (Song et al. 2023) estimates correction transforms through optimal-transport-based co-visible object matching, while CoAlign (Lu et al. 2023) optimizes cross-agent pose consistency through an agent-object pose graph.

B. Temporal compensation of traffic representations

Temporal compensation aligns collaborative messages generated at earlier sensing times with the current ego-vehicle time before fusion. DAIR-V2X (Yu et al. 2022a) introduces Time Compensation Late Fusion, which estimates object motion from adjacent infrastructure frames to compensate infrastructure detections to the vehicle timestamp before late fusion. Beyond object-level compensation, SyncNet (Lei et al. 2022) performs feature-level synchronization through feature-attention symbiotic estimation and time modulation.

The temporal compensation process can be summarized as:

$$\tilde{Z}_k^{t_e} = \mathcal{A}\left(Z_k^{t_k}, \Phi_\theta\left(Z_k^{t_k}, \mathcal{H}_k, t_e - t_k\right)\right), \tag{6}$$

where $Z_k^{t_k}$ denotes the collaborative information generated by agent k at sensing time t_k, and t_e denotes the current ego-vehicle fusion time. The term $t_e - t_k$ represents the temporal delay between information generation and fusion. $\mathcal{H}_k$ denotes historical

information used to infer temporal changes, such as adjacent frames, region-of-interest (ROI) tracklets, or cached BEV features. Φ_θ estimates how the collaborative information changes during the delay. Depending on the compensation design, Φ_θ models object motion, feature evolution, BEV flow, or reference-point offsets. $\mathcal{A}$ then applies the estimated change to $Z_k^{t_k}$, producing the temporally compensated representation $\tilde{Z}_k^{t_e}$ at the ego-vehicle time.

Recent studies further model temporal displacement in BEV space. CoBEVFlow (Wei et al. 2023) further models temporal displacement through BEV flow, which reassigns asynchronous perceptual features to appropriate spatial locations under irregular and continuous delays. LRCP (Wang and Nordström 2025) instead directly fuses asynchronous BEV features by using latency-aware flow from cached history to calibrate deformable-attention sampling locations.

C. Unified traffic representation fusion

Unified representation fusion integrates spatially aligned and temporally compensated collaborative features into an ego-centric perceptual representation. Early feature-level methods mainly rely on direct aggregation. F-Cooper (Chen et al. 2019) performs voxel- and spatial-feature fusion using maxout over overlapping features. V2VNet (Wang et al. 2020) further formulates fusion as graph-based message passing, aggregating spatially aligned inter-vehicle features through message aggregation and convolutional gated recurrent unit (ConvGRU)-based updates.

Subsequent studies introduce more adaptive and attention-based fusion mechanisms. DiscoNet (Li et al. 2021) introduces a distilled collaboration graph with matrix-valued edge weights for spatially adaptive inter-agent feature aggregation. For heterogeneous V2X agents, V2X-ViT (Xu et al. 2022) combines type-aware multi-agent self-attention with multi-scale window attention for robust feature fusion. CoBEVT (Xu et al. 2023a) uses fused axial attention to capture sparse local and global interactions across agents for BEV-level fusion. Overall, unified representation fusion has evolved from maxout-based feature aggregation to graph message passing, cell-level adaptive weighting, and attention-based contextual interaction.

By integrating multimodal observations, cooperative messages, spatiotemporal alignment, and representation fusion, embodied perception provides the vehicle-centered embodied agent with a unified perceptual representation of the traffic environment. Such representations form the basis for embodied closed-loop interaction, while embodied cognition further interprets scene semantics, interactive intentions, and traffic knowledge under rule constraints.

4.2 Embodied Cognition in Internet of Vehicles

Embodied cognition in IoV builds on the unified scene representation produced by embodied perception and further converts observable traffic states into task-oriented understanding. In dynamic traffic environments, a vehicle-centered embodied agent must not only recognize objects and spatial layouts, but also interpret road structures, task constraints, interaction intentions, historical experiences, and possible future evolution. The conversion of observable traffic states into task-oriented understanding provides the semantic and temporal basis for subsequent embodied decision-making by organizing perceptual information into structured traffic knowledge. Accordingly, this

section discusses traffic situation comprehension, traffic intention reasoning, memory reflection, and world modeling.

4.2.1 Traffic Situation Comprehension

Traffic situation comprehension serves as the initial embodied cognitive process that converts ego-centric perceptual representations into traffic-oriented world understanding. Beyond object recognition and spatial representation, traffic situation comprehension interprets the road structure in which traffic participants move, the semantic relations among vehicles, pedestrians, road elements, signals, obstacles, and infrastructure, and the rule-related constraints that shape the current traffic state. Existing studies mainly focus on road structure cognition and task-constrained traffic situation assessment.

A. Road structure cognition

Road structure cognition converts perceptual representations into structured descriptions of road geometry and traffic organization. HDMapNet (Li et al. 2022a) provides an early instance-level representation by learning local vectorized semantic maps from onboard observations through semantic segmentation, instance embedding, and direction prediction.

Beyond explicit map elements, GeMap (Zhang et al. 2025e) introduces G-Representation and geometry-decoupled attention to learn Euclidean shapes and relations among map instances. LaneSegNet (Li et al. 2024) further advances road representation to lane segments that jointly encode lane geometry and topology within a unified structure.

On this basis, topology-oriented studies further incorporate structural relations. OpenLane-V2 (Wang et al. 2023) formulates scene structure perception and reasoning through 3D lane detection, traffic-element recognition, and lane–lane and lane–traffic topology recognition. TopoNet (Li et al. 2026b) performs end-to-end topology reasoning using a scene graph neural network (SGNN) and scene knowledge graph to model lane connectivity and lane–traffic-element associations.

B. Task-constrained traffic situation assessment

Task-constrained situation assessment characterizes the current driving condition with respect to navigation commands or ongoing maneuvers. Conditional Affordance Learning for Driving in Urban Environments (CAL) (Sauer et al. 2018) predicts six intermediate affordances from monocular video, with road-relative cues conditioned on directional commands and the remaining traffic-related cues estimated independently for downstream control.

Maneuver-level risk assessment further incorporates spatial relations and temporal evolution. Scene-Graph Augmented Data-Driven Risk Assessment of Autonomous Vehicle Decisions (Yu et al. 2022b) models subjective driving risk from multi-relational scene graphs using a multi-relational graph convolutional network (MR-GCN), attention, and long short-term memory (LSTM)-based temporal modeling.

RS2G (Wang et al. 2024a) further replaces predefined scene-graph relations with data-driven edge extraction and variationally regularized relation learning for

subjective-risk assessment. By contrast, risk-oriented methods additionally use spatiotemporal traffic graphs or counterfactual object removal to identify road agents that influence driver behavior (Suman et al. 2023; Li et al. 2023).

Existing embodied traffic situation comprehension methods provide a structured basis for interpreting road organization, task constraints, and risk-relevant conditions. However, geometric structure, topology relations, affordance states, and risk estimates are commonly modeled through separate task-specific formulations, limiting the integration of geometric structure, topology relations, affordance states, and risk estimates into a coherent understanding of complex traffic situations. Beyond understanding the current scene, embodied cognition must further infer the behavioral intentions of surrounding embodied agents.

4.2.2 V2X-Aware Traffic Intention Reasoning

Embodied traffic situation comprehension establishes a structured account of road organization, task constraints, and risk-related conditions, providing the contextual basis for interpreting surrounding embodied traffic agents. V2X-aware traffic intention reasoning extends the structured understanding of road organization, task constraints, and risk-related conditions from current scene states to potential future behaviors by integrating historical motion, scene context, agent attributes, and available V2X information (Casas et al. 2018; Salzmann et al. 2020). Existing studies mainly focus on individual traffic-agent intention understanding and multi-agent traffic interaction reasoning. The former identifies the latent intentions and motion tendencies of individual vehicles, pedestrians, and cyclists, whereas the latter models dynamic influences, yielding relations, priority conflicts, and socially responsive behaviors among multiple embodied traffic agents (Gupta et al. 2018; Sun et al. 2022).

A. Individual Traffic-Agent Intention Understanding

Individual traffic-agent intention understanding extends traffic situation comprehension from observable states to latent actions and motion tendencies. Existing studies mainly examine vehicle maneuver intention and vulnerable-road-user behavior. Vehicle intention is typically constrained by road geometry and traffic control, whereas pedestrian and cyclist intention additionally depends on body motion, posture, and local visual context.

For embodied vehicle agents, IntentNet (Casas et al. 2018) jointly detects vehicles and predicts discrete driving intentions and future trajectories from sequential LiDAR sweeps and rasterized dynamic high-definition (HD) maps, with intention scores conditioning trajectory forecasting. Rather than predicting predefined semantic classes, MTR (Shi et al. 2022) represents multimodal futures with learnable motion-query pairs that jointly perform global intention localization and local movement refinement. In IoV settings, V2X-Graph (Ruan et al. 2024) further associates cross-view trajectories of the same agent and fuses cooperative motion, agent–lane, and inter-agent interaction features for multimodal motion forecasting.

Vulnerable-road-user intention understanding relies more strongly on action-specific bodily cues. Pedestrian crossing prediction with attention (PCPA) (Kotseruba et al. 2021) combines local visual context, bounding-box trajectories, 2D poses, and ego-vehicle speed through temporal and modality attention for pedestrian crossing

prediction. For cyclists, Bieshaar et al. (2018) combine infrastructure-camera and smart-device starting detectors through an extreme-gradient-boosting stacking ensemble for cooperative starting-movement detection. Collectively, the reviewed vehicle and vulnerable-road-user-intention methods extend intention representation from explicit maneuvers to latent motion modes and connected multi-source behavioral inference. Most methods nevertheless remain centered on individual-agent outcomes, motivating subsequent reasoning about directional influence and reciprocal response among agents.

B. Multi-Agent Traffic Interaction Reasoning

Multi-agent traffic interaction reasoning extends individual intention understanding by modeling how the prospective future of one embodied traffic agent conditions the behavior of others. M2I (Sun et al. 2022) models directional interactions by identifying influencer–reactor relations and combining marginal influencer prediction with conditional reactor prediction.

FJMP (Rowe et al. 2023) extends factorized interaction reasoning to multiple agents by representing future interactions as a sparse directed acyclic graph and decomposing joint prediction into ordered marginal and conditional predictions. Following FJMP, the resulting directed acyclic interaction graph supports the following joint factorization:

$$P(Y \mid X, C) = \prod_{n=0}^{N-1} P\left(Y^{(n)} \mid Y_{\mathrm{pa}_G(n)}, X, C\right), \tag{7}$$

where X and Y denote the observed and future trajectories of all agents, respectively, and C denotes the encoded HD-map context. G represents the predicted directed acyclic interaction graph, while $\mathrm{pa}_G(n)$ denotes the parent agents whose predicted future trajectories are used to condition the future prediction of agent n.

Beyond graph-ordered prediction, GameFormer (Huang et al. 2023) models hierarchical future responses through level-k decoding, with each higher level conditioned on other agents' predictions from the preceding level.

IoV connectivity further introduces externally shared temporal information into interaction reasoning. With V2X connectivity, Co-MTP (Zhang et al. 2025b) further fuses vehicle- and infrastructure-side histories with ego planning and infrastructure-side future predictions through a heterogeneous graph transformer for cooperative trajectory forecasting.

Taken together, existing studies have progressively extended interactive intention reasoning from explicit maneuver recognition and multimodal motion inference to vulnerable-road-user behavior anticipation, structured joint prediction, and hierarchical response modeling (Huang et al. 2025a). These developments enrich the representation of heterogeneous traffic-agent intentions and improve the consistency of multi-agent future prediction. Nevertheless, current approaches remain limited by task-specific intention definitions, insufficient pedestrian and cyclist interaction data, and sensitivity to occlusion, tracking errors, and perception–prediction error propagation (Dal'Col et al. 2025). V2X-assisted future conditioning further

increases dependence on communication quality, with latency, bandwidth limitations, localization errors, and communication disturbances introducing uncertainty into collaborative reasoning. Heuristic relation labels, acyclic influence structures, and fixed-level reasoning also simplify reciprocal and long-range interactions, while dense traffic scenarios impose additional computational and training costs.

4.2.3 Traffic Memory and World Modeling

Traffic situation comprehension and interactive intention reasoning provide task-oriented interpretations of current traffic conditions and surrounding-agent behaviors. Sustained embodied cognition further requires historical states and driving experience to remain available across perception–action cycles, while the evolution of the traffic environment must be represented and predicted beyond the current observation. Accordingly, existing studies mainly investigate reflective driving memory for temporal continuity and experience reuse, together with traffic-dynamics-aware world modeling for representing traffic evolution and generating possible future states.

A. Reflective Driving Memory

Reflective driving memory preserves task-relevant information across successive perception–execution cycles, allowing historical states and experiences to inform subsequent embodied reasoning. BridgeAD (Zhang et al. 2025a) reformulates motion and planning queries as multi-step queries and integrates historical prediction and planning information into the corresponding perception and motion-planning stages.

Beyond short-term query caching, explicit memory supports the retrieval of experience acquired from previous driving situations. DiLu (Wen et al. 2024) retrieves similar driving experiences from memory for few-shot reasoning and uses reflection to revise unsafe decisions before updating the memory.

LeapVAD (Ma et al. 2025) improves experience retrieval with learned scene tokens and combines an Analytic Process for experience accumulation and reflection with a Heuristic Process for efficient memory-based decision-making. For multi-agent driving, KoMA (Jiang et al. 2025) introduces shared memory and ranking-based reflection to accumulate, evaluate, and refine driving experiences across multiple agents.

B. Traffic-Dynamics-Aware World Modeling

Traffic-dynamics-aware world models construct internal representations of traffic environments and learn how the internal traffic-environment representations evolve over time, enabling future scene states to be generated from historical observations and ego-motion conditions (Guan et al. 2024).

Moving beyond appearance-level prediction, structured world models represent traffic evolution in three-dimensional spatial states. OccWorld (Zheng et al. 2025a) tokenizes 3D semantic occupancy into discrete scene representations and uses a spatial–temporal generative Transformer to predict future occupancy and ego trajectories.

Action-conditioned world models further characterize the conditional evolution of traffic environments under different ego behaviors. Drive-WM (Wang et al. 2024d) employs latent video diffusion with temporal and multiview factorization to generate spatially and temporally consistent future observations conditioned on ego-motion trajectories. Drive-OccWorld (Yang et al. 2025b) instead forecasts future occupancy and

3D motion flow from historical BEV features using semantic- and motion-conditioned normalization and flexible action conditions.

Taken together, existing studies have progressively expanded vehicular memory from short-term query retention to experience retrieval, reflection-based revision, and multi-agent sharing, improving temporal continuity and the reuse of prior driving knowledge. Nevertheless, most methods still depend on fixed memory lengths or embedding similarity, while experience validation, forgetting, and conflict resolution remain insufficiently addressed. World models have similarly advanced toward action-conditioned visual and occupancy forecasting, yet the performance of action-conditioned visual and occupancy world models is still limited by short prediction horizons, accumulated errors, rare interactions, and substantial computational overhead.

4.3 Embodied Decision-Making in Internet of Vehicles

Based on the traffic understanding established by embodied cognition, embodied decision-making generates feasible embodied behavior strategies for vehicle-centered embodied agents operating in dynamic IoV environments. As a central stage of the embodied perception–cognition–decision-making–execution loop, embodied decision-making evaluates candidate behaviors according to task objectives, traffic regulations, safety constraints, surrounding-agent intentions, and anticipated environmental evolution. In IoV scenarios, embodied decision-making extends beyond ego-vehicle policy generation, because connected embodied agents must negotiate conflicting maneuvers and coordinate distributed computation and communication tasks to support timely embodied interaction. Through ego-policy generation, multi-agent maneuver negotiation, and distributed task coordination, embodied decision-making links embodied cognition with embodied execution and enables vehicle-centered agents to interact safely and cooperatively with the physical traffic environment. Accordingly, this section discusses driving policy generation, multi-vehicle maneuver coordination, and vehicular edge task coordination.

4.3.1 Vehicle-Centered Driving Policy Generation

Following traffic situation comprehension, intention reasoning, memory retrieval, and world modeling, embodied decision-making further determines appropriate embodied action policies under task objectives, traffic regulations, safety requirements, and anticipated environmental evolution (Li et al. 2026a). Vehicle-centered driving policy generation focuses on the algorithmic derivation of behavior-level policies that provide high-level guidance for subsequent embodied execution, rather than directly producing trajectories or control commands. Existing approaches can be broadly organized by policy-generation mechanism into simulation-based driving policy reasoning and data-driven driving policy learning.

A. Simulation-Based Driving Policy Reasoning

Simulation-based driving policy reasoning selects an ego embodied vehicle policy by explicitly representing admissible closed-loop policies and evaluating the predicted consequences of the admissible closed-loop policies. Cunningham et al. (2015) propose

Multipolicy Decision-Making (MPDM), which evaluates a finite set of closed-loop ego policies by forward-simulating interactions with surrounding vehicles and selecting the policy with the best predicted outcome. The basic policy-election process can be summarized as

$$\begin{aligned} \Psi_\pi &= \mathrm{Sim}\left(x_t, \pi, \hat{\pi}_t^{-e}, H\right), \\ S(\pi \mid x_t) &= \sum_{q=1}^{Q} w_q \bar{m}_q\left(\Psi_\pi\right), \\ \pi_t^* &= \arg\max_{\pi \in \Pi_{\mathrm{app}}(x_t)} S(\pi \mid x_t), \end{aligned} \tag{8}$$

Here, x_t denotes the current world-state estimate, and $\Pi_{\mathrm{app}}(x_t)$ is the set of applicable ego policies. The term π denotes a candidate ego policy, whereas $\hat{\pi}_t^{-e}$ contains the most likely policies assigned to surrounding vehicles. H is the simulation horizon, and Ψ_π is the resulting multi-vehicle state sequence. Q denotes the number of evaluation metrics. $\bar{m}_q(\Psi_\pi)$ denotes the normalized performance of the simulated policy outcome under the q-th evaluation criterion, while w_q specifies the relative contribution of that criterion to the overall policy score. $S(\pi \mid x_t)$ denotes the overall policy score, and π_t^* is the selected ego policy.

Galceran et al. (2017) subsequently augment MPDM with Bayesian changepoint-based behavior prediction, sampling surrounding-vehicle policy hypotheses for closed-loop simulation and selecting the ego policy with the highest expected reward.

Building on MPDM's semantic closed-loop policies, Zhang et al. (2020) propose Efficient Uncertainty-Aware Decision-Making (EUDM), which combines a domain-specific closed-loop policy tree with conditional focused branching to support multi-stage ego decisions while limiting action- and intention-space branching.

B. Data-Driven Driving Policy Learning

Data-driven driving policy learning derives driving policies from data-driven optimization, where the decision rule is obtained through environment interaction or expert demonstrations rather than explicit online simulation of predefined policies. Reinforcement learning provides a basic formulation for learning-based policy generation. Mirchevska et al. (2021) couple amortized Q-learning with model-based action proposals from a trajectory planner to select long-term highway driving strategies from locally optimal maneuver candidates. The learning and decision process can be summarized as

$$\begin{aligned} y_t &= r_{t+1} + \gamma \max_{a'} Q_{\theta^-}(s_{t+1}, a'), \\ L(\theta) &= \mathbb{E}\left[\left(y_t - Q_\theta(s_t, a_t)\right)^2\right], \\ a_t^* &= \arg\max_{a \in A(s_t)} Q_\theta(s_t, a), \end{aligned} \tag{9}$$

Here, s_t and a_t denote the current traffic state and candidate driving action, while r_{t+1} and s_{t+1} denote the immediate reward and next state. Q_θ is the action-value network, Q_{θ^-} is the target network used to compute the Bellman target y_t, γ is the discount factor, and $A(s_t)$ denotes the available action set. The learned policy selects the action with the maximum estimated value.

Subsequent reinforcement learning studies further improve safety during policy optimization. Wang et al. (2019) improve safety by combining DQN-based lane-change

decision-making with rule-based constraints and low-level trajectory modification before execution.

Imitation learning provides another route by reducing dependence on manually specified rewards. Wang et al. (2021) augment AIRL with semantic rewards to jointly learn driving rewards and policies from expert demonstrations in interactive driving scenarios. Recent studies further introduce generative artificial intelligence into vehicle-centered driving policy generation. DriveMLM (Cui et al. 2025) aligns a multimodal large language model with behavioral planning states to predict path and speed decisions from multimodal observations, traffic rules, and user instructions, whereas Diffusion Planner (Zheng et al. 2025b) models closed-loop planning as diffusion-based conditional trajectory generation for multimodal driving behaviors.

Overall, vehicle-centered driving policy generation has enabled traffic cognition to be transformed into candidate driving strategies and executable behavior policies. Existing studies are particularly effective in structured policy evaluation, uncertainty-aware strategy selection, and data-driven policy acquisition, improving both the controllability and adaptability of autonomous driving decisions. However, current methods remain constrained by predefined policy spaces, simplified interaction assumptions, training-data dependence, reward-design sensitivity, and limited safety generalization. These limitations suggest that vehicle-centered driving policy generation has made substantial progress, but still faces challenges in achieving robust, interpretable, and generalizable policy selection under complex IoV traffic dynamics.

4.3.2 Multi-Vehicle Maneuver Coordination

Multi-vehicle maneuver coordination further addresses how candidate maneuvers generated by embodied decision policy generation can be accepted, rejected, modified, or committed in embodied multi-agent traffic conflicts. Different from ego-centered maneuver selection, the key problem is not to generate another driving policy, but to determine whether multiple candidate maneuvers can coexist under shared road-space, priority, and reservation constraints. In lane-changing, merging, and unsignalized-intersection scenarios, multi-vehicle maneuver negotiation involves conflict-area allocation, spatiotemporal reservation, maneuver-coordination message exchange, and right-of-way assignment (Manzinger and Althoff 2017).

A basic formulation is the negotiation of conflicting road areas. Manzinger and Althoff (2018) use reachable-set computation to determine road cells requested by multiple cooperative embodied vehicles and formulate the allocation of the requested conflict cells among cooperative vehicles as a winner determination problem:

$$\begin{aligned} &\max_{\delta(t,C_j)} \quad \sum_{C_j \in P(t)} \delta(t,C_j) b(t,C_j), \\ &\text{s.t. } \forall g_i \in G^C(t): \quad \sum_{C_j : g_i \in C_j} \delta(t,C_j) \leq 1, \quad \forall C_j \in P(t) : \delta(t,C_j) \in \{0,1\}, \end{aligned} \tag{10}$$

Here, $G^C(t)$ denotes the set of conflicting road cells, C_j is a candidate package of conflict cells, $P(t)$ is the allowed package set, $b(t, C_j)$ denotes the bid for selecting C_j, and

$\delta(t, C_j)$ indicates whether the package is assigned. The constraint prevents a conflict cell from being allocated more than once. The winner-determination formulation shifts decision-making from selecting an isolated ego maneuver to allocating conflicting road areas and deriving an individual driving corridor for each cooperative vehicle.

Subsequent studies refine the object and communication process of multi-vehicle maneuver negotiation. Heß et al. (2018) introduce an application-level vehicle-to-vehicle (V2V) protocol that coordinates cooperative maneuvers by negotiating space-time reservations in conflict areas. Nichting et al. (2020) further generalize the Space Time Reservation Procedure (STRP) to coordinate cooperative vehicles across diverse conflict scenarios.

Maneuver coordination protocols further organize negotiation through explicit message types. Mizutani et al. (2021) develop AutoMCM with seven maneuver coordination message (MCM) types and event-driven state management to support general maneuver coordination. Rather than introducing another coordination protocol, Molina-Masegosa et al. (2023) propose Risk and Tracking Trajectories rules that adapt MCM generation to potential safety risks or significant planned-trajectory changes.

At Stop Controlled-Intersections (SC-Is), Shah et al. (2024) propose the distributed Automated Right-of-Way (AROW) algorithm to assign explicit crossing priorities to connected and/or autonomous vehicles (CAVs) while preserving existing stop-control rules.

Overall, existing work has progressed from the allocation of conflicting road areas to space-time reservation, maneuver coordination protocols, message generation rules, and right-of-way assignment. Conflict-area allocation, space-time reservation, MCM-based coordination, and right-of-way assignment improve the consistency and interpretability of cooperative maneuver decisions, but still depend on predefined maneuver representations, reliable V2X communication, limited negotiation scale, and stable agreement formation in dense or mixed traffic. Failure-prevention studies further indicate that negotiation protocols require explicit handling of inconsistent intentions, communication failures, and unresolvable maneuver conflicts under complex traffic conditions (Häfner et al. 2022).

4.3.3 Vehicular Edge Task Coordination

Embodied decision-making in IoV involves not only selecting driving policies, but also making coordination decisions over distributed computation and communication resources. Perception updates, intention prediction, policy evaluation, and execution feedback are continuously generated during vehicle–environment interaction, while onboard processors and V2X links are constrained by computation limits, mobility, latency, and unstable connectivity. Maintaining decision validity requires vehicles, roadside units, and edge servers to decide where tasks are processed and how resources are assigned. Existing studies mainly address the coordination of distributed computation and communication resources through vehicular edge task scheduling and V2X edge resource allocation.

A. Vehicular Edge Task Scheduling

Vehicular edge task scheduling determines the embodied execution decisions of decision-related computation tasks, including the processing location and execution time of each task and how dependent subtasks are arranged across vehicles and edge servers. In VEC, offloading studies usually describe each task through input data size, required computation, and latency constraints, and then jointly determine whether computation should remain on the vehicle or be transferred to edge-side resources. The dependency-aware task model establishes a scheduling mechanism in which task order, execution cost, communication delay, and application deadline jointly determine the collaborative embodied execution path (Dai et al. 2018; Xu et al. 2023c).

Later studies refine dependency-aware vehicular edge task scheduling by considering heterogeneous V2X links, wireless interference, edge cooperation, and service-quality constraints. These factors make scheduling decisions more consistent with cellular vehicle-to-everything (C-V2X) link conditions and with the limited computing capacity of roadside or edge nodes. As a result, vehicular edge task scheduling is usually coupled with V2X edge resource allocation rather than treated as an isolated placement problem (Dai et al. 2018; Xu et al. 2023c).

B. V2X Edge Resource Allocation

V2X edge resource allocation determines the resource decisions associated with scheduled or offloaded tasks, including how communication bandwidth, transmission power, and edge computing capacity are assigned under latency, energy, and service-quality constraints. Joint Offloading and Resource Allocation in Vehicular Edge Computing and Networks (Dai et al. 2018) provides a representative model by describing each vehicular task through input data size, required computation, and maximum latency. Communication time and edge execution time are incorporated into the task-processing delay, and JOSC jointly optimizes VEC-server selection and computation-resource allocation to maximize system utility under latency constraints. The JOSC joint offloading and resource-allocation formulation shows that offloading decisions must be coupled with edge computing capacity rather than treated as independent task-placement choices.

Subsequent studies strengthen V2X edge resource allocation by incorporating wireless interference, edge cooperation, application-specific quality of service (QoS), and non-orthogonal multiple access (NOMA). Joint Task Offloading and Resource Optimization in NOMA-Based Vehicular Edge Computing (Xu et al. 2023c) introduces NOMA-based V2I transmission, models intra-edge and inter-edge interference, and allows heterogeneous edge nodes to cooperate for real-time task processing. Task offloading is formulated as an exact potential game and solved by MAD4PG, while resource allocation is decomposed into convex optimization problems solved through gradient-based iteration and Karush–Kuhn–Tucker (KKT) conditions.

Overall, existing methods have progressed from deadline-constrained vehicular edge task scheduling to joint server-selection and computation allocation, NOMA-aware edge cooperation, and QoS-aware resource optimization. The reviewed vehicular edge task-scheduling and V2X edge resource-allocation methods improve the real-time feasibility of decision-related computation, but most formulations still optimize delay, energy, service ratio, or utility at the system level. The effects of coordination

decisions on cognition updates, policy validity, execution feedback, and safety-critical closed-loop performance remain insufficiently integrated.

4.4 Embodied Execution in Internet of Vehicles

Embodied execution in IoV realizes the behavior strategies generated by embodied decision-making through physical vehicle motion. As the physical implementation stage of the embodied closed-loop, embodied execution maps reference paths, trajectories, desired velocities, and cooperative maneuver agreements into steering, acceleration, and braking commands while satisfying vehicle-dynamic and actuator constraints. In connected traffic environments, embodied execution further requires vehicle-centered embodied agents to execute coordinated motions and maintain safety and stability under interaction uncertainty and imperfect V2X communication. Accordingly, this section discusses vehicle motion control, cooperative vehicle motion control, and safety-critical vehicle control.

4.4.1 Vehicle Motion Control

Vehicle motion execution generates feasible trajectories and control commands based on driving objectives, vehicle states, and surrounding traffic information, providing the direct basis for physical vehicle operation. As the final stage connecting high-level driving decisions with actual vehicle motion, vehicle motion control determines whether planned behaviors can be accurately and reliably realized under vehicle dynamics and actuator constraints, and its execution results also provide physical feedback for subsequent perception and decision-making. Existing AI-based methods can be summarized into three directions: diffusion-based vehicle trajectory generation, learning-based vehicle trajectory tracking, and reinforcement learning (RL)-based vehicle motion control.

A. Diffusion-Based Vehicle Trajectory Planning

Diffusion-based vehicle trajectory planning generates future ego-vehicle trajectories from current traffic information, providing continuous motion references for subsequent vehicle control. Diffusion Policy (Chi et al. 2023) provides a key policy-learning foundation for diffusion-based action generation by representing visuomotor policies as conditional denoising diffusion processes and generating continuous action sequences through iterative denoising. Combined with receding-horizon control, the policy executes part of each predicted action sequence before replanning from updated observations.

In autonomous driving, Diffusion Planner (Zheng et al. 2025b) applies conditional diffusion modeling to closed-loop trajectory planning. Following its diffusion formulation, the forward diffusion process is expressed as

$$q_{t|0}\left(x^{(t)} \mid x^{(0)}\right) = \mathcal{N}\left(x^{(t)} \mid \alpha_t x^{(0)}, \sigma_t^2 \mathbf{I}\right), \quad t \in [0,1], \tag{11}$$

and the planner is trained to recover clean future trajectories through

$$\mathcal{L}_\theta = \mathbb{E}_{x^{(0)},\, t\sim U(0,1),\, x^{(t)}\sim q_{t|0}\left(x^{(t)}|x^{(0)}\right)} \left[\left\| \mu_\theta \left(x^{(t)}, t, C \right) - x^{(0)} \right\|_2^2 \right]. \tag{12}$$

Here, $x^{(0)}$ denotes the clean joint future trajectories used as the training target, and $x^{(t)}$ is their noisy representation at diffusion time t. C denotes the driving conditions, σ_t controls the Gaussian-noise variance, $\alpha_t = \sqrt{1-\sigma_t^2}$ controls the retained trajectory information, and $\mu_\theta(x^{(t)}, t, C)$ denotes the trajectory estimate produced by the model parameterized by θ. Equation (11) progressively adds Gaussian noise, whereas Eq. (12) trains the model to recover the clean trajectories from noisy samples under the driving conditions.

Subsequent studies improve the efficiency and formulation of diffusion-based planning. DiffusionDrive (Liao et al. 2025) introduces prior multi-mode trajectory anchors and a truncated diffusion schedule, reducing the 20-step vanilla diffusion baseline used in its experiments to two denoising steps. BridgeDrive (Liu et al. 2026) further addresses the asymmetry between the forward and denoising processes of truncated anchor diffusion by formulating a diffusion bridge from coarse anchor trajectories to refined context-aware plans.

B. Learning-Based Vehicle Trajectory Tracking

Learning-based vehicle trajectory tracking incorporates learned vehicle dynamics, model-error estimates, or control mappings into trajectory-tracking controllers to improve tracking performance under nonlinear dynamics and real-time requirements. Hu et al. (2025) employ a variational free-energy sparse Gaussian process within a two-layer model predictive control (MPC) framework to estimate the mismatch between the nominal model and actual vehicle dynamics during drifting. The next-step lower-layer vehicle state is calculated as

$$x_{k+1}^l = f_d\left(x_k^l, u_k^l\right) + d_{\mathrm{VFE}}\left(x_k^l, u_k^l\right), \tag{13}$$

where $x_k^l = [V_k, \beta_k, r_k]^{\mathrm{T}}$ denotes velocity, sideslip angle, and yaw rate, while $u_k^l = [\delta_k, F_{xr,k}]^{\mathrm{T}}$ contains steering angle and rear-tire longitudinal force. f_d is the nominal discrete single-track model, and d_{VFE} is the model error learned by the variational free-energy sparse Gaussian process.

Recent studies further enhance vehicle-dynamics modeling with neural networks. Chen et al. (2026) integrate a physics-based vehicle dynamics model into physics-informed neural network (PINN) training and constrain the learned physical parameters to remain valid. The resulting PINN model is embedded into nonlinear model predictive control (NMPC) to improve trajectory-tracking prediction accuracy. Gong et al. (2025) employ a Symplectic ODE-Net to learn lateral vehicle dynamics and construct a SymODEN-based MPC. A deep stabilizing controller further accounts for input constraints and learning-error bounds, enabling feasibility and stability analysis during trajectory tracking. Shi et al. (2025) focus on the high online computational cost of NMPC and propose an NN Optimizer. During training, the NMPC loss is

backpropagated through a differentiable vehicle model to learn the control policy. During online tracking, control inputs are obtained through network evaluation, avoiding repeated nonlinear optimization.

C. RL-Based Vehicle Motion Control

Reinforcement learning (RL)-based vehicle motion control learns control policies through interaction with the driving environment and directly participates in vehicle motion control. Compared with conventional controllers that rely on predefined models or control laws, RL can improve adaptability to nonlinear vehicle dynamics and changing driving conditions. Xia et al. (2025) combine a preview model with TD3 to generate the desired yaw rate for lateral tracking. An adaptive corrector further refines the yaw-rate command based on predicted lateral deviation, while RI-SAC adjusts the desired longitudinal speed according to tracking errors.

To improve the stability and interpretability of learning-based tracking, Han et al. (2025) combine Pure Pursuit with PPO. Pure Pursuit provides the primary steering command with baseline stability and interpretability, while PPO generates a corrective steering command to compensate for tracking errors. PPO-generated coefficients also adjust the look-ahead distance according to vehicle speed and road curvature. Recent studies further extend RL to multiple vehicle control variables. Yang et al. (2026) develop a hierarchical deep reinforcement learning (DRL) framework in which the high-level controller determines acceleration and the low-level controller determines steering, reducing the complexity of joint control while maintaining accurate path tracking.

Constrained RL further incorporates explicit safety constraints into learned control policies. Yu et al. (2025a) formulate autonomous racing as a constrained Markov decision process (CMDP) and train agents using CPO and PPO-Lagrangian. A shielding framework based on vehicle rollover dynamics further projects unsafe speed commands onto a safe boundary, improving safety while maintaining competitive racing performance.

Overall, the introduction of diffusion models, learning-based dynamics models, and RL policies improves trajectory generation, tracking accuracy, and adaptive motion control. However, computational efficiency, model generalization, control stability, and reliable constraint satisfaction under unseen driving conditions still limit real-world deployment.

4.4.2 Cooperative Vehicle Motion Control

In Internet of Vehicles-based embodied intelligence, vehicle motion is often influenced by the actions of surrounding embodied vehicles. Platoon following, lane changing, and intersection crossing require coordinated motion among multiple vehicles, since the feasible motion of each vehicle depends on the speed, spacing, and trajectory adjustments of others. V2X communication supports such cooperative motion execution by providing vehicle-state and interaction information, allowing multiple vehicles to coordinate their acceleration, steering, spacing, and trajectories. Recent studies further introduce reinforcement learning and multi-agent learning to adapt these control actions to complex vehicle interactions and changing traffic conditions (Zhang et al.

2026). Existing approaches mainly focus on cooperative platoon control, cooperative lane-changing control, and cooperative intersection control.

A. Cooperative Platoon Control

Cooperative platoon control uses inter-vehicle information to coordinate the motion of multiple vehicles in a platoon, maintaining desired spacing, velocity consistency, and stable collective motion. Recent studies increasingly employ reinforcement-learning-based cooperative control to learn vehicle control commands from platoon interactions, improving adaptability to curved roads, mixed traffic, and non-ideal V2V communication. Li et al. (2025a) develop a serial distributed PPO framework for CAV platooning on curved roads. The SDPPO agents generate optimal longitudinal acceleration for individual vehicles, while a nonlinear lateral controller produces steering adjustments for trajectory tracking. V2V communication supports distributed policy learning across platoon members and reduces the complexity of multi-vehicle control.

Wang et al. (2025b) further consider mixed platoons containing connected autonomous vehicles and human-driven vehicles. The proposed framework characterizes aggregated human-driving behavior through traffic-flow and car-following models, and incorporates previous control inputs into the SAC state representation to address communication delay. The learned policy directly generates longitudinal control inputs subject to vehicle acceleration limits. Xu et al. (2026) further address time-varying delay and packet loss through DCT-MARL. A gated communication mechanism dynamically adjusts inter-vehicle information exchange, while historical control and delay information are incorporated into the agent state. The resulting policy outputs commanded acceleration for following vehicles, improving platoon string stability and driving comfort under non-ideal V2V communication.

B. Cooperative Lane-Changing Control

Cooperative lane-changing control coordinates the longitudinal and lateral motions of the lane-changing vehicle and surrounding vehicles to create suitable gaps and complete lane changes smoothly. Recent studies increasingly employ reinforcement learning and learning-based control to generate coordinated vehicle motions from multi-vehicle interactions. Wang et al. (2024b) propose a multi-agent reinforcement learning framework for mandatory lane changing in bottleneck areas, where each CAV continuously generates longitudinal and lateral accelerations. A buffer-zone strategy maintains suitable spacing before the bottleneck, while a two-dimensional collision-avoidance mechanism masks unsafe actions during execution. Yao et al. (2024a) further consider cooperative lane changing in mixed traffic containing CAVs and human-driven vehicles. The TD3-based controller generates the longitudinal and lateral accelerations of the target vehicle and the longitudinal accelerations of the new leader and follower when they participate in cooperative control. A feedback module pre-computes the utilities of potential lane-changing executions in terms of safety, efficiency, comfort, and ecology to guide subsequent lane-changing decisions. Wan et al. (2024) further combine multi-agent reinforcement learning with model-based lateral control for platoon lane changing. A fully decentralized MA-DDPG framework generates longitudinal accelerations for platoon following under different communication topologies, while adaptive MPC tracks the lateral lane-changing trajectory and a fuzzy controller compensates the front-wheel steering angle.

C. Cooperative Intersection Control

Cooperative intersection control coordinates the motion of multiple connected vehicles approaching and traversing an intersection to avoid conflicts while maintaining traffic efficiency. Recent studies increasingly employ multi-agent reinforcement learning to generate cooperative vehicle motions at unsignalized intersections. Guillen-Perez and Cano (2022) develop an end-to-end multi-agent deep reinforcement learning framework for autonomous intersection management, where the learned policy determines the normalized speed that each controlled CAV should follow at the next time step, enabling collaborative intersection crossing without traffic signals. Gao et al. (2025) further propose a behavior-constrained MAPPO framework for cooperative vehicle control. The learned policy generates longitudinal acceleration commands and steering variations for CAVs. Formal safety verification introduces penalties for risky interactions, while behavior constraints are triggered when interactions become unsafe. Zhao et al. (2025) further extend safe multi-agent reinforcement learning to all-directional unsignalized intersections through a centralized MAPCPO framework. Approaching vehicles transmit their kinematic information to a central controller, while a queue-based dynamic state input handles changing traffic and LSTM units capture vehicle interactions and historical state information. The learned policy jointly performs lateral and longitudinal vehicle control, while MAPCPO constrains policy updates within explicit safety limits.

Overall, reinforcement learning and multi-agent learning enable cooperative controllers to directly optimize continuous vehicle motions from multi-vehicle interactions, improving adaptability to nonlinear coupling, mixed traffic, and changing communication conditions. However, these methods still depend heavily on training scenarios and communication quality, and their scalability, generalization to unseen vehicle behaviors, and reliable safety guarantees under large-scale interactions remain limited.

4.4.3 Safety-Critical Vehicle Execution

Safety-critical vehicle execution provides a protection layer between nominal motion commands and physical vehicle actuation, ensuring that acceleration, steering, and other control inputs satisfy collision-avoidance requirements, vehicle-dynamic limits, and actuator constraints. Conventional control barrier function (CBF)-based safety filters modify nominal commands when predefined safety constraints are close to being violation. However, manually constructing suitable barrier functions becomes difficult for nonlinear dynamics and complex traffic interactions, while learning-based controllers such as reinforcement learning do not inherently guarantee constraint satisfaction. Recent studies therefore increasingly learn barrier functions and integrate CBF constraints with learned control policies.

One line of recent work reduces the dependence on handcrafted safety functions by learning CBFs from data. So et al. (2024) propose a policy neural control barrier function (PNCBF), which constructs a CBF by learning the value function of a nominal policy and uses the learned certificate as a safety filter. Yin et al. (2025) further extend PNCBFs to discrete-time systems and learn an approximate discrete-time CBF for sampling-based model predictive control (MPC). The learned barrier function is incorporated into variational inference MPC, while a resampling strategy improves

sampling efficiency under safety constraints. The resulting Neural Shield-VIMPC is further validated in simulation and on the AutoRally autonomous-vehicle platform. A representative discrete-time CBF condition employed by Yin et al. (2025) can be expressed as

$$B(f(x_k, \pi(x_k))) - B(x_k) \leq -\alpha(B(x_k)), \qquad \mathcal{S} = \{x \mid B(x) \leq 0\}, \tag{14}$$

where $B(x)$ characterizes the safety boundary of the discrete-time system, $f(x_k, \pi(x_k))$ denotes the state reached after applying policy π, and $\alpha(\cdot)$ is an extended class-$\mathcal{K}$ function. The inequality restricts the evolution of the barrier value so that a state initially satisfying $B(x) \leq 0$ does not cross the prescribed safety boundary during subsequent control. Neural-CBF methods replace the manually constructed B with a learned approximation, but additional verification may still be required before the learned function can provide formal safety certification.

Beyond learning the barrier function itself, recent studies increasingly integrate barrier-function-based constraints with reinforcement learning to improve safety during policy learning and execution, as reviewed by Kushwaha and Biron (2026). Wang et al. (2025a) propose knowledge-enhanced safe reinforcement learning for autonomous driving, combining a high-order CBF-based model constraint with a model-free constraint that estimates the distribution of accumulated safety cost. Both constraints are incorporated into a Lagrangian framework for safe policy optimization. Zhou et al. (2026) further introduce cooperative CBF constraints into multi-agent reinforcement learning for mixed-autonomy platoon control. A differentiable quadratic programming layer incorporates CBF-based safety constraints into the learned controller, while conformal prediction quantifies uncertainties in surrounding-vehicle behaviors and introduces the resulting bounds into the safety constraints. Luo et al. (2026) further integrate CBF constraints into cooperative lane-changing multi-agent reinforcement learning. CBF-MAAC computes dynamic safe action ranges and maps actor outputs into these ranges, allowing safety constraints to restrict both policy exploration and vehicle-motion execution.

Overall, learning-based CBFs reduce reliance on manually designed safety functions, while CBF-integrated reinforcement learning introduces explicit safety constraints into learned vehicle control. However, verification of learned barriers, generalization to unseen traffic conditions, real-time computation, and feasibility under multiple simultaneous safety constraints remain important challenges.

More importantly, embodied intelligence in IoV does not terminate at vehicle execution. The executed acceleration, steering, and cooperative control commands change both the vehicle state and the surrounding traffic environment, and these physical changes are subsequently captured by onboard sensors and V2X-connected agents. This continuous interaction between vehicle actions and the evolving environment is a key characteristic of closed-loop autonomous driving (Li et al. 2026a). Based on the updated observations, the perception module reconstructs the current traffic scene, while the cognition module further interprets vehicle states, surrounding agents, and their interactions. The decision-making module then generates new driving decisions according to the updated scene understanding, and the execution module realizes these decisions through vehicle motion control. Accordingly, perception, cognition,

decision-making, and execution form a continuous feedback loop rather than an isolated sequential pipeline, enabling embodied vehicles to repeatedly sense, understand, decide, act, and adapt as the traffic environment evolves.

5 Datasets for Embodied Intelligence in Internet of Vehicles

Datasets provide the empirical foundation for developing and evaluating embodied intelligence in IoV. Different stages of the embodied closed loop require distinct forms of data supervision: distributed sensor observations support collaborative perception, intention and motion annotations facilitate cognitive reasoning, and trajectories, actions, rewards, and control signals enable decision-making and execution. Accordingly, this section organizes representative datasets into collaborative embodied perception datasets, traffic scene understanding datasets, and interactive decision-making and execution datasets, with the characteristics of the three dataset categories summarized in Tables 2, 3, and 4. The representative datasets in Tables 2–4 are compared in terms of data source, participating entities or cooperation modes, sensing modalities, task targets, and reported scale.

5.1 Collaborative Embodied Perception Datasets

Collaborative embodied perception datasets provide the observational basis for embodied agents to construct vehicle-centered representations of dynamic traffic environments. Unlike conventional single-vehicle datasets that describe the environment from an isolated sensing viewpoint, collaborative embodied perception datasets record complementary observations distributed across connected vehicles, roadside infrastructure, and other embodied sensing entities, including configurations based on infrastructure-to-infrastructure (I2I) cooperation. Cross-agent poses, calibration parameters, timestamps, and perception annotations are commonly provided to support embodied information sharing, spatial-temporal alignment, multimodal fusion, and collaborative scene construction. Table 2 compares representative datasets according to cooperation mode, data source, input modality, embodied representation type, and reported scale.

Existing datasets mainly support three levels of embodied scene representation: object-centric, BEV semantic, and 3D occupancy representations. Object-centric datasets constitute the dominant category and provide 3D bounding boxes, object attributes, and motion states for collaborative detection, tracking, and target-state estimation. Such representations allow embodied vehicles to identify surrounding traffic participants and establish basic interaction objects. BEV semantic datasets further organize vehicles, roads, lanes, and other traffic elements within a unified BEV space, enabling embodied agents to construct structured spatial representations beyond isolated objects. More recent 3D occupancy datasets discretize the surrounding environment into volumetric grids and annotate occupancy states or semantic categories, thereby extending embodied perception from sparse traffic entities to dense modeling of visible and occluded spaces.

Table 2 Representative datasets for collaborative embodied perception

Dataset	Year	Cooperation Mode	Source	Input modalities			Type	Dataset scale	
				Camera	LiDAR	Radar		Scenes	Frames / samples
VIC-Sync (Yu et al. 2022a)	2022	V2I	Real-world	✓	✓	–	Object-centric	–	9,311
V2V4Real (Xu et al. 2023b)	2023	V2V	Real-world	✓	✓	–	Object-centric	67	20,000
TUMTraf-V2X (Zimmer et al. 2024)	2024	V2I	Real-world	✓	✓	–	Object-centric	–	2,000
V2X-Real (Xiang et al. 2025)	2024	V2V / I2I	Real-world	✓	✓	–	Object-centric	–	33,000
RCooper (Hao et al. 2024)	2024	I2I	Real-world	✓	✓	–	Object-centric	–	30,000
DOLPHINS (Mao et al. 2023)	2022	V2V / V2I	Simulation	✓	✓	–	Object-centric	6	42,376
V2X-Radar-C (Yang et al. 2025a)	2025	V2I	Real-world	✓	✓	✓	Object-centric	50	20,000
V2X-R (Huang et al. 2025c)	2025	V2X	Simulation	✓	✓	✓	Object-centric	12,079	37,727
V2X-Sim (Li et al. 2022b)	2022	V2V / V2I	Simulation	✓	✓	–	BEV semantic	100	10,000
OPV2V (Xu et al. 2023a)	2023	V2V	Simulation	✓	–	–	BEV semantic	73	11,464
DeepAccident (Wang et al. 2024c)	2024	V2V / V2I	Simulation	✓	✓	–	BEV semantic	691	57,000
M2I (Zhang et al. 2025d)	2025	I2I	Simulation	✓	✓	–	BEV semantic	1,103	278,184
M^3CAD (Zhu et al. 2026)	2026	V2V	Simulation	✓	✓	–	BEV semantic	204	30,000
Semantic-OPV2V (Song et al. 2024)	2024	V2V	Simulation	✓	–	–	3D occupancy	73	11,464
UniOcc (OpenCOOD) (Wang et al. 2025d)	2025	V2V	Simulation	✓	–	–	3D occupancy	–	–

Note: ✓ indicates that the modality is available or explicitly used for the listed benchmark; – indicates unavailable or not reported. Annotation counts are excluded from dataset scale. Year denotes the dataset's first public release year and may differ from the publication year of the cited paper.

Substantial differences can also be observed in collection environments, sensing configurations, and embodied cooperation patterns. Real-world datasets mainly focus on V2V, V2I, or I2I cooperation and remain concentrated on object-centric embodied perception, whereas simulation facilitates large-scale generation of BEV semantic and 3D occupancy annotations. Camera–LiDAR configurations dominate current datasets because camera and LiDAR provide complementary semantic and geometric information for embodied scene construction, while radar-supported collaborative datasets remain comparatively limited. Existing benchmarks therefore support increasingly diverse forms of embodied perception, but large-scale real-world datasets covering heterogeneous sensors, asynchronous observations, dense spatial annotations, adverse conditions, and broader multi-entity embodied interaction remain insufficient.

5.2 Traffic Scene Understanding Datasets

Traffic scene understanding datasets provide the supervision required for vehicle-centered embodied agents to transform perceptual observations into task-oriented interpretations of traffic environments. Collaborative embodied perception describes observable objects, spatial structures, and environmental states, whereas traffic scene understanding further interprets the latent meanings and possible evolution underlying these observations. Historical motion, road structures, traffic controls, agent attributes, interaction relations, and ego-vehicle states are therefore organized to support the inference of behavioral intentions, future motion, potential risks, and scene semantics. Through such understanding, an embodied vehicle can move beyond static recognition and establish an internal representation of the potential behaviors of surrounding traffic agents and the subsequent evolution of the physical traffic environment.

Existing datasets support different levels of traffic scene understanding, ranging from intention and trajectory prediction to behavior understanding and scene-level cognition. Intention datasets focus on latent actions such as pedestrian crossing, vehicle turning, yielding, or lane changing, while trajectory datasets describe possible future motions under road and interaction constraints. Behavior datasets provide action- or event-level annotations for interpreting observable traffic activities. Traffic scene understanding datasets further incorporate risk objects, causal explanations, question answering, and language-based scene interpretation, enabling embodied agents to associate traffic observations with higher-level semantic judgments. Table 3 compares representative datasets according to data source, involved traffic-agent categories, scene-understanding task, and reported scale.

Table 3 Representative Datasets for Traffic Scene Understanding

Dataset	Year	Source	Ego / driver	Vehicle	Pedestrian	Cyclist	Task type	Dataset scale
JAAD (Rasouli et al. 2017)	2017	Real-world	–	–	✓	–	Intention	346
PIE (Rasouli et al. 2019)	2019	Real-world	–	–	✓	–	Intention	1,842
STIP (Liu et al. 2020a)	2020	Real-world	–	–	✓	–	Intention	1,108,176
Brain4Cars (Jain et al. 2015)	2015	Real-world	✓	–	–	–	Intention	700
LOKI (Girase et al. 2021)	2021	Real-world	–	✓	✓	✓	Intention	644
IDD-PeD (Bokkasam et al. 2025)	2025	Real-world	–	–	✓	–	Intention	5,000
PIE++ (Khindkar et al. 2024)	2024	Real-world	–	–	✓	–	Intention	5,607
DRAMA-X (Godbole et al. 2025)	2025	Real-world	✓	–	✓	✓	Intention	5,686
INTERACTION (Zhan et al. 2019)	2019	Real-world	–	✓	✓	✓	Trajectory	40,054
INT2 (Yan et al. 2023b)	2023	Real-world	–	✓	✓	✓	Trajectory	612,000
WOMD (Ettinger et al. 2021)	2021	Real-world	–	✓	✓	✓	Trajectory	103,354
V2X-Seq-TFD (Yu et al. 2023)	2023	Real-world	–	✓	✓	✓	Trajectory	210,000
V2X-Traj (Ruan et al. 2024)	2024	Real-world	–	✓	✓	✓	Trajectory	10,102
TITAN (Honda Research Institute USA 2020)	2020	Real-world	✓	✓	✓	✓	Behavior	700
ROAD (Singh et al. 2023)	2021	Real-world	–	✓	✓	✓	Behavior	22
WOMD-Reasoning (Li et al. 2025b)	2025	Real-world	–	✓	✓	✓	Cognition	3,000,000
BDD-X (Kim et al. 2018)	2018	Real-world	✓	–	–	–	Cognition	6,970
BDD-OIA (Xu et al. 2020)	2020	Real-world	✓	✓	✓	✓	Cognition	22,924
DRAMA (Malla et al. 2023)	2023	Real-world	–	✓	✓	✓	Cognition	17,785
DriveLM-nuScenes (Sima et al. 2025)	2024	Real-world	✓	✓	✓	✓	Cognition	4,871
Reason2Drive (Nie et al. 2025)	2024	Real-world	✓	✓	✓	✓	Cognition	600,000

Note: ✓ indicates that the corresponding agent category is included or explicitly considered; – indicates unavailable or not emphasized. Task labels denote intention prediction, trajectory reasoning, behavior understanding, and traffic scene understanding. Dataset scale follows each dataset's reported unit and is not directly comparable across all rows. Year denotes the dataset's first public release year and may differ from the publication year of the cited paper.

5.3 Interactive Decision-Making and Execution Datasets

Interactive decision-making and execution datasets support the transformation of traffic cognition into goal-directed behavior and physically feasible vehicle execution. After an embodied vehicle has interpreted road structures, surrounding-agent intentions, potential risks, and future traffic evolution, the embodied vehicle must determine how to respond within traffic rules, safety constraints, and vehicle dynamics. Relevant datasets therefore provide supervision in the form of behavior labels, planned trajectories, expert demonstrations, execution sequences, reward signals, and low-level control commands. Unlike embodied cognitive reasoning datasets that mainly predict or explain environmental evolution, decision-making and execution datasets directly describe how a vehicle-centered embodied agent should operate in the physical traffic environment.

Existing datasets cover multiple levels of embodied decision-making and execution. Behavior-decision and trajectory-planning datasets supervise the selection of driving maneuvers and the generation of future motion paths. Vision–language trajectory-planning datasets further associate multimodal scene understanding and linguistic instructions with embodied execution outputs. Offline reinforcement-learning datasets provide state, action, reward, and transition records for learning embodied policies without continuous online interaction. Planning-and-control datasets jointly evaluate trajectory generation and closed-loop vehicle operation, whereas steering-control datasets focus on converting perceptual observations into low-level actuation commands. Table 4 compares representative datasets according to data source, embodied target, task orientation, and reported scale, distinguishing datasets that primarily support embodied decision-making, embodied execution, or the integrated optimization of embodied decision-making and execution.

6 Evaluation Metrics for Embodied Intelligence in Internet of Vehicles

6.1 Perception Fidelity

Perception fidelity characterizes how faithfully an embodied vehicle converts multimodal observations into an internal representation of the surrounding traffic environment. In embodied intelligence, the internal representation of the surrounding traffic environment provides the perceptual basis for interaction reasoning, decision-making, and closed-loop execution. Perception-fidelity evaluation should therefore examine whether traffic participants are correctly detected and localized, whether scene semantics and spatial occupancy are accurately reconstructed, and whether dynamic states remain temporally consistent.

At the object level, precision is calculated as the ratio of correct detections to all reported detections and therefore quantifies the reliability of the reported detections. Recall divides the number of successfully detected objects by the total number of ground truth (GT) objects, indicating whether relevant traffic participants are omitted. Average precision (AP) summarizes the precision–recall relationship for one

Table 4 Representative embodied decision-making and execution datasets

Dataset	Year	Source	Embodied target		Task	Scale
			Decision-making	Execution		
M^3CAD (Zhu et al. 2026)	2026	Simulation	✓	–	TP	30,000
V2Xverse (Liu et al. 2025a)	2025	Simulation	✓	–	TP	36,130
VLA4CoDrive (Boroujeni and Razi 2026)	2026	Simulation	✓	–	TP	1,000,000
nuPlan (Karnchanachari et al. 2024)	2024	Real-world	✓	–	TP	1,282
CoVLA (Arai et al. 2025)	2025	Real-world	✓	–	VLTP	10,000
BDD-OIA (Xu et al. 2020)	2020	Real-world	✓	–	BD	22,924
AD4RL (Lee et al. 2024)	2024	Hybrid	✓	–	ORL	19,000,000
Bench2Drive (Jia et al. 2024)	2024	Simulation	✓	✓	PC	2,000,000
D4RL-CARLA (Fu et al. 2021)	2020	Simulation	✓	✓	ORL-C	2,000,000
DDD17 (Binas et al. 2017)	2017	Real-world	–	✓	SC	12
DDD20 (Hu et al. 2020)	2020	Real-world	–	✓	SC	51

Note: ✓ indicates that the dataset officially supports or explicitly targets the corresponding embodied decision-making or execution task. – indicates unavailable or not emphasized. Abbreviations: TP = trajectory planning, VLTP = vision–language trajectory planning, BD = behavior decision, ORL = offline reinforcement learning, PC = planning and control, ORL-C = offline reinforcement learning for control, SC = steering control. Dataset scales follow the original units reported by the corresponding publications and are not directly comparable across datasets. Year denotes the dataset's first public release year and may differ from the publication year of the cited paper.

category across confidence thresholds, while mean average precision (mAP) averages AP over all evaluated categories:

$$\mathrm{AP}_c = \int_0^1 p_c(r)\,\mathrm{d}r, \qquad \mathrm{mAP} = \frac{1}{C}\sum_{c=1}^{C}\mathrm{AP}_c. \tag{15}$$

Here, $p_c(r)$ denotes the precision of category c at recall r, AP_c is its average precision, and C is the number of evaluated categories.

For 3D detection, AP_{3D} determines matches using complete three-dimensional bounding boxes and evaluates volumetric localization, whereas AP_{BEV} performs matching on ground-plane projections and emphasizes horizontal position and footprint consistency (Geiger et al. 2012).

At the scene level, pixel accuracy is calculated as the proportion of correctly classified pixels or BEV cells, providing an overall measure of semantic classification. Because pixel accuracy is biased toward large-area classes, Intersection over Union

(IoU) is more commonly used to measure the overlap between the predicted and reference regions of each semantic category. Mean IoU (mIoU) averages the class-wise results using true positive (TP), false positive (FP), and false negative (FN) counts:

$$\mathrm{IoU}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c}, \qquad \mathrm{mIoU} = \frac{1}{C}\sum_{c=1}^{C} \mathrm{IoU}_c. \tag{16}$$

Here, TP_c, FP_c, and FN_c denote correctly classified, falsely assigned, and missed elements of category c, respectively.

In BEV segmentation, mIoU evaluates roads, lane boundaries, and pedestrian crossings. For 3D occupancy prediction, geometric IoU ignores category labels and compares occupied with free voxels, whereas semantic mIoU preserves voxel classes. RayIoU casts query rays into the predicted and reference volumes and evaluates the category and depth of the first occupied voxel, revealing depth inconsistencies not captured by voxel-wise overlap metrics (Tang et al. 2024).

Temporal fidelity examines whether embodied perception maintains stable identities and motion states across successive observations. Multiple Object Tracking Accuracy (MOTA) aggregates FN, FP, and identity switches (IDSW) relative to the number of GT objects:

$$\mathrm{MOTA} = 1 - \frac{\sum_t (\mathrm{FN}_t + \mathrm{FP}_t + \mathrm{IDSW}_t)}{\sum_t \mathrm{GT}_t}. \tag{17}$$

Here, FN_t, FP_t, IDSW_t, and GT_t represent the FN, FP, IDSW, and GT counts at time t, respectively.

Multiple Object Tracking Precision (MOTP) averages the localization discrepancy of matched tracks and measures spatial tracking precision (Bernardin and Stiefelhagen 2008). Identification F1 score (IDF1) combines identity precision and identity recall to assess whether the same physical agent retains a consistent identity throughout a sequence (Ristani et al. 2016).

Taken together, the object-level, scene-level, and temporal-fidelity metrics indicate whether the embodied world representation is sufficiently accurate, complete, and temporally coherent to support reliable cognition, decision-making, and execution. Because aggregate scores alone do not expose failures under safety-critical conditions, results should also be reported separately for different sensing distances, levels of occlusion, weather conditions, sensor-noise levels, and communication-quality settings.

6.2 Reasoning Accuracy

In vehicular embodied intelligence, reasoning accuracy measures whether an embodied agent can derive correct semantic conclusions and reliable future-state estimates from the internal traffic-environment representation constructed by the embodied agent. The derived semantic conclusions and future-state estimates connect environmental understanding with interaction, decision-making, and closed-loop execution. Because embodied cognition produces both discrete judgments, such as traffic-rule states and agent intentions, and continuous predictions, such as future trajectories, the evaluation of embodied cognition generally combines classification-based and displacement-based metrics.

For traffic-situation assessment and intention inference, accuracy is calculated as the proportion of samples assigned to the correct category and therefore measures the overall correctness of the inferred labels. Precision measures the proportion of predicted instances of a target state that are correct, whereas recall measures the proportion of the target state's GT instances that are successfully identified. F1 combines precision and recall through the harmonic mean of precision and recall:

$$\mathrm{Acc} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i), \qquad \mathrm{F1} = \frac{2PR}{P+R}. \tag{18}$$

Here, N is the number of evaluated samples, while $\hat{y}_i$ and y_i denote the predicted and GT labels of sample i. P and R represent precision and recall, respectively (Kotseruba et al. 2021).

Accuracy provides an intuitive overall score but is often dominated by frequent states, such as lane-keeping or non-crossing behavior. F1 jointly penalizes FP and FN reasoning results and is therefore more suitable for evaluating a safety-critical target state. For multi-class reasoning, Macro-F1 first computes F1 for each category and then averages the class-wise values with equal weight, reducing the influence of class imbalance.

Future-state reasoning is evaluated by comparing predicted trajectories with the observed future motion. Average Displacement Error (ADE) averages the Euclidean distance between predicted and GT positions over the entire prediction horizon, whereas Final Displacement Error (FDE) measures the position error at the final time step:

$$\mathrm{ADE} = \frac{1}{T}\sum_{t=1}^{T}\left\|\hat{x}_t - x_t\right\|_2, \qquad \mathrm{FDE} = \left\|\hat{x}_T - x_T\right\|_2. \tag{19}$$

Here, T denotes the number of predicted time steps, and $\hat{x}_t$ and x_t are the predicted and GT positions at time t, respectively. Lower ADE and FDE indicate more accurate trajectory inference.

When multiple futures are generated, minADE and minFDE select the smallest ADE and FDE among the K candidate trajectories according to the benchmark protocol. MinADE and minFDE therefore measure whether the candidate set contains at least one trajectory close to the realized future. Miss Rate (MR) calculates the proportion of samples for which the endpoint of every candidate trajectory exceeds a benchmark-defined distance threshold from the GT endpoint, thereby revealing whether the predicted modes fail to cover the actual future (Wilson et al. 2021).

Classification metrics and trajectory metrics assess complementary aspects of embodied reasoning: the former evaluate the correctness of semantic conclusions, whereas the latter quantify the geometric agreement between inferred and observed futures. However, fixed-dataset scores cannot fully capture how reasoning errors influence interaction, decision-making, and execution in closed-loop traffic environments. Results should therefore also be reported across interaction types, prediction horizons, class distributions, and safety-critical scenarios.

6.3 Decision-Making Effectiveness

In vehicular embodied intelligence, the decision-making module generates and selects driving strategies based on traffic cognition, including scene understanding, interaction-intention inference, and future-state prediction. Decision-making quality evaluates whether the selected strategy enables the embodied vehicle to accomplish the prescribed driving objective while maintaining safety, traffic-rule compliance, efficiency, and passenger comfort. Because each strategy alters subsequent observations and interactions, safety, traffic-rule compliance, efficiency, and passenger comfort are primarily assessed under closed-loop conditions.

Task accomplishment is commonly measured by route completion (RC) and success rate (SR). RC is the completed route length divided by the prescribed route length, whereas SR is the proportion of routes in which the vehicle reaches the prescribed destination within the allotted time and without benchmark-defined infractions:

$$\mathrm{RC}_i = \frac{L_i^{\mathrm{completed}}}{L_i^{\mathrm{route}}} \times 100\%, \qquad \mathrm{SR} = \frac{n_{\mathrm{success}}}{n_{\mathrm{total}}}. \tag{20}$$

Here, $L_i^{\mathrm{completed}}$ and L_i^{route} denote the completed and total lengths of route i, while n_{success} and n_{total} are the numbers of successful and evaluated routes. RC preserves partial progress when a route is interrupted, whereas SR applies a stricter criterion to complete task achievement (Jia et al. 2024).

Safety and rule compliance are evaluated through collisions, red-light and stop-sign violations, off-road driving, and related infractions. CARLA-based evaluation combines route progress with multiplicative penalties in the driving score (DS):

$$\mathrm{DS} = \frac{1}{N} \sum_{i=1}^{N} \left(\mathrm{RC}_i \prod_{j=1}^{n_i} p_{i,j} \right). \tag{21}$$

Here, N is the number of routes, n_i is the number of multiplicatively penalized infractions on route i, and $p_{i,j} \in [0, 1]$ is the corresponding penalty coefficient. DS decreases as infractions accumulate, preventing route progress from masking unsafe or illegal decisions. Individual infraction frequencies should also be reported because the aggregate score does not identify the source of performance degradation (Jia et al. 2024).

Decision-making efficiency measures whether the policy maintains traffic-compatible progress without unnecessary waiting, deadlock, or excessive conservatism. Bench2Drive defines speed percentage (SP) by comparing the ego-vehicle speed with the average speed of nearby traffic at each valid checkpoint and obtains driving efficiency (DE) by averaging the resulting ratios:

$$\mathrm{SP}_q = \frac{v_q^{\mathrm{ego}}}{\bar{v}_q^{\mathrm{nearby}}}, \qquad \mathrm{DE} = \frac{1}{Q} \sum_{q=1}^{Q} \mathrm{SP}_q. \tag{22}$$

Here, Q is the number of valid speed checks, while v_q^{ego} and $\bar{v}_q^{\mathrm{nearby}}$ denote the ego and nearby-traffic speeds at checkpoint q. A higher score indicates stronger traffic progress but should be interpreted together with safety performance (Jia et al. 2024).

Comfort characterizes the smoothness of the executed closed-loop trajectory. Fixed-length trajectory segments are evaluated using longitudinal and lateral acceleration, yaw rate, yaw acceleration, and jerk-related variables. A segment is considered smooth only when all variables remain within predefined bounds:

$$\text{Smoothness} = \frac{N_{\text{smooth}}}{N_{\text{total}}}. \tag{23}$$

Here, N_{smooth} and N_{total} denote the numbers of smooth and evaluated segments. A higher score indicates fewer abrupt motions and better passenger comfort (Jia et al. 2024).

Overall, closed-loop benchmarks extend decision-making evaluation beyond task completion by jointly considering safety, efficiency, and comfort. The joint assessment of task completion, safety, efficiency, and comfort provides a more comprehensive basis for evaluating whether an embodied vehicle can generate effective strategies from traffic cognition and sustain goal-directed interaction with the environment. Nevertheless, composite scores often obscure distinct failure modes; disaggregated results across interaction types and safety-critical scenarios remain necessary to identify unsafe, overly conservative, or dynamically unstable strategies.

6.4 Execution Quality

In vehicular embodied intelligence, execution quality evaluates the ability of the vehicle control module to accurately and promptly track the reference trajectory or achieve the target state generated by the planning layer. The execution module determines steering, driving, and braking commands, thereby completing the perception–cognition–decision-making–execution loop. The performance of the execution module is primarily examined through tracking accuracy, transient stability, computational timeliness, actuator feasibility, and robustness under varying operating conditions.

Tracking accuracy is quantified from the discrepancy between executed and reference states. Lateral error measures the vehicle's offset from the reference path, heading error describes the angular difference between the vehicle orientation and the path direction, and speed error quantifies longitudinal speed-tracking deviation. Depending on the evaluated state, the root mean square error (RMSE) and maximum absolute error can be expressed as:

$$\text{RMSE}_z = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(z_t - z_t^{\text{ref}}\right)^2}, \qquad E_z^{\max} = \max_{1\le t\le T}\left|z_t - z_t^{\text{ref}}\right|. \tag{24}$$

Here, T denotes the number of control samples, while z_t and z_t^{ref} represent the executed and reference values at sample t. Depending on the evaluation target, the state z denotes lateral position, heading angle, or longitudinal speed (Yang et al. 2020).

RMSE summarizes tracking deviation over the complete test and assigns greater weight to large errors, whereas the maximum absolute error records the worst observed deviation. The former characterizes overall execution accuracy, while the latter reveals

whether short-duration tracking errors approach operational limits. Existing vehicle-control experiments consequently report lateral and heading errors separately and examine the maximum lateral and heading errors under different trajectories.

Transient stability is further assessed through maximum overshoot, which records the peak deviation beyond the reference during a steering or motion transition. Smaller overshoot indicates that rapid reference changes are followed without excessive correction or oscillation. Average computation time measures the typical online computational demand per control cycle:

$$\bar{\tau}_c = \frac{1}{T} \sum_{t=1}^{T} \tau_{c,t}. \tag{25}$$

Here, $\tau_{c,t}$ denotes the time required to calculate the control command at step t (Yang et al. 2020).

Lower computation time indicates reduced computational demand, while maximum computation time should additionally be examined when control-cycle deadlines must be guaranteed. Original controller comparisons commonly report lateral RMSE, maximum overshoot, and average computation time together to reveal the trade-off between tracking precision, response stability, and real-time feasibility.

Actuator feasibility is examined from the magnitude and variation of executed steering, driving, and braking commands. Steering saturation, high-frequency oscillation, or abrupt command changes indicate that low tracking error is achieved at the expense of physically feasible or smooth control. Corresponding limits should therefore be evaluated according to the vehicle architecture, such as steering-angle, steering-rate, acceleration, and braking constraints (Liu et al. 2018).

Robustness is evaluated by repeating the above measurements under different speeds, road curvatures, low-friction surfaces, model mismatches, time delays, and external disturbances. Limited degradation in RMSE, maximum error, overshoot, and computation time indicates that the embodied execution module can continue realizing the selected motion under changing conditions (Yang et al. 2020; Liu et al. 2018).

Overall, execution quality requires a complementary assessment of tracking accuracy, transient response, computational feasibility, and actuator behavior. A low average tracking error alone does not rule out large error peaks, delayed commands, or unstable control actions.

7 Future Directions

Embodied intelligence for IoV integrates networked information exchange with the perception, cognition, decision-making, and execution of connected vehicles and roadside infrastructure in dynamic traffic environments. Recent studies have explored perception, cognition, decision-making, and execution capabilities across different functional stages, but the field remains at an early stage, with many fundamental questions yet to be resolved. Four directions are particularly relevant: perception fusion under ambient uncertainty, generalizable intention prediction, policy learning in long-tail autonomous driving scenarios, and experimental validation of vehicle motion control in real-world environments.

7.1 Vehicle Perception Fusion under Ambient Uncertainty

Existing embodied perception methods in IoV combine multimodal fusion, collaborator selection, perception information selection, spatial alignment, temporal compensation, pose error correction, and unified representation fusion. Multimodal fusion improves local object and occupancy perception, while V2X cooperation extends sensing beyond line of sight and supplements observations of occluded traffic participants. Spatial alignment and temporal compensation further reduce geometric inconsistency and information staleness before aggregation. However, ambient uncertainty remains after fusion because limited fields of view, unobserved regions, delayed observations, and conflicting cross-agent estimates leave the existence, location, and motion states of surrounding traffic participants uncertain. Confidence maps and attention weights mainly indicate information relevance or feature importance rather than calibrated residual uncertainty (Cho et al. 2025). Recent studies have therefore begun to introduce uncertainty quantification and uncertainty-aware fusion, while using quantified uncertainty to guide perception information selection (Gawlikowski et al. 2023; Chen et al. 2025).

Future research should develop a closed-loop perception-fusion process that continuously estimates and reduces ambient uncertainty. A probabilistic scene representation should first identify unobserved regions, low-confidence objects, and inconsistent state estimates. Residual uncertainty should then guide collaborator selection and perception information selection (Mukhopadhyay et al. 2026), allowing the ego vehicle to acquire observations that are most likely to reduce uncertainty in safety-critical regions. Selected observations should undergo spatial alignment, temporal compensation, and pose error correction before entering a common representation space. Uncertainty-aware fusion should update object states and occupancy representations according to cross-agent consistency, temporal freshness, and calibrated confidence, while unresolved conflicts should remain explicitly represented rather than being forced into single deterministic estimates. Residual uncertainty should subsequently be propagated to cognition and decision-making, where high-uncertainty states can trigger additional information requests or conservative behavior generation. The resulting framework should support selective information acquisition, iterative scene-state updating, calibrated uncertainty estimation, and real-time operation. Such development can expand effective perception coverage, reduce missed detections behind occlusions, improve cross-agent state consistency, and provide more accurate and reliable traffic representations for closed-loop embodied interaction.

7.2 Generalizable Vehicle Intention Prediction

Existing intention-prediction methods can infer vehicle maneuvers, pedestrian crossing intentions, and interaction-dependent behaviors from motion histories, road context, agent relations, and multimodal observations (Karuppasamy et al. 2025). Early intention estimates provide behavioral evidence for traffic-risk assessment and planning before the corresponding physical actions become fully observable. However, most reported performance gains are still obtained within fixed datasets or narrowly defined traffic domains. Cross-dataset evaluations reveal substantial accuracy degradation

when visual conditions, road layouts, traffic composition, or regional behavior patterns change (Yao et al. 2024b). Representative generalization studies also remain concentrated on limited tasks, including intersection turning, two-vehicle passing and yielding, and pedestrian crossing. Recent research has consequently begun to move beyond source-domain accuracy by conducting cross-dataset evaluations, learning causal time-series representations across driving domains, and incorporating scene classification to strengthen generalization.

Future research should develop intention-prediction models that remain reliable across unseen cities, road structures, traffic-agent populations, sensing configurations, and V2X deployment conditions without complete retraining for every environment. Developing such cross-domain capability first requires training data that cover different regions, road types, traffic rules, participant compositions, and connectivity settings. Intention-prediction models should then identify behavioral cues that remain informative across these environments, including motion evolution, gap acceptance, rule compliance, and responses to surrounding agents, while reducing dependence on background appearance, regional driving habits, and sensor-specific patterns. Road topology, lane permissions, and traffic-control states should subsequently exclude infeasible maneuvers and refine the candidate intention set for each traffic scene (Grimm et al. 2025). Evaluation should finally reserve complete cities, road categories, traffic-agent groups, and V2X configurations as unseen domains rather than drawing random test samples from familiar datasets. Domain-generalized intention prediction can reduce dependence on scenario-specific retraining, improve prediction reliability in unseen traffic domains, and provide more stable cognitive inputs for risk assessment, interaction negotiation, and behavior planning.

7.3 Policy Learning in Long-Tail Autonomous Driving Scenarios

Existing policy-learning methods can generate lane-following, lane-changing, merging, intersection-crossing, and other driving decisions from expert demonstrations, interaction rewards, and multimodal traffic observations (Jia et al. 2024). Learned policies have achieved strong decision performance in frequently observed traffic conditions and reduced dependence on manually specified driving rules. However, existing datasets provide insufficient coverage of long-tail scenarios involving rare road geometries, unusual agent behaviors, and complex multi-agent interactions. Limited exposure to long-tail conditions can lead to brittle or unsafe decisions when traffic situations extend beyond the training distribution. Recent research has begun to identify failure-prone long-tail events from driving logs, reconstruct each event into an interactive simulation, augment the reconstructed scenarios with diverse traffic variations, and refine driving policies through reinforcement post-training (Yan et al. 2025). Current methods nevertheless remain limited to failure modes already contained in collected logs. Imperfect world reconstruction, incomplete modeling of pedestrians, cyclists, and unstructured road users, manually defined rewards, and instability across multiple rounds of post-training further constrain policy improvement.

Future research should extend long-tail event discovery beyond logged data through closed-loop adversarial scenario generation and min–max optimization (Nie et al.

2026). Newly discovered conditions should be reconstructed into interactive worlds and augmented with realistic and controllable behaviors of surrounding agents. The resulting rollouts should support targeted reinforcement post-training, while learned or verifiable reward functions should guide policy improvement according to collision avoidance, lane adherence, route progress, and more complex driving norms. The improved policy should then be re-evaluated to discover unresolved or newly introduced failure modes, followed by successive rounds of world modeling and reinforcement post-training. Larger policy models and more sophisticated regularization strategies should support stable multi-round post-training without interfering with previously learned behaviors. Iterative refinement can expand decision coverage beyond recorded long-tail events, narrow the simulation-to-road transfer gap, reduce recurrent failures, and improve the safety and reliability of embodied decision-making in the Internet of Vehicles.

7.4 Experimental Validation of Vehicle Motion Control in Real-World Environments

Existing vehicle motion control methods can coordinate trajectory tracking, vehicle stability, and actuator allocation in high-fidelity simulations, hardware-in-the-loop platforms, and limited vehicle tests (Abtahi et al. 2025). Experimental validation further examines whether control algorithms can maintain real-time computation, accurate actuator execution, and closed-loop stability on physical vehicles. However, predefined vehicle models and standard test maneuvers cannot fully reproduce tire–road adhesion changes, actuator delays, sensor noise, onboard computing constraints, or V2X communication latency. Most evaluations also remain limited to fixed road conditions, bounded speed ranges, and specific vehicle configurations, leaving simulated control performance insufficient to demonstrate reliable operation in complex traffic environments (Zhao et al. 2024).

Recent studies have therefore shifted from simulation-based evaluation toward progressive experimental validation on real-time processors, prototype vehicles, closed tracks, and public roads. Future validation should first integrate onboard measurements, V2X information, and actuator feedback to estimate vehicle dynamics, road adhesion, and available steering, driving, and braking capabilities (Ning and Behl 2024). Updated estimates should then revise reference motions and safety constraints before control commands are issued. Recorded execution responses should support online model correction and controller retuning. Experimental validation should subsequently expand from hardware-in-the-loop tests to high-speed, low-adhesion, actuator-degraded, and open-road conditions. Such development can improve the real-time feasibility, stability, transferability, and safety of vehicle motion control across vehicle platforms and traffic environments.

8 Conclusion

This survey examined embodied intelligence for the Internet of Vehicles from a vehicle-centered and closed-loop perspective. Rather than treating vehicular networking, perception, reasoning, planning, and control as separate research topics, the survey

connected IoV information exchange and resource sharing with the continuous embodied process of perception, cognition, decision-making, execution, and environmental feedback. On this basis, existing techniques were organized according to their roles in the embodied loop, while representative datasets and quantitative metrics were reviewed for the corresponding perception, cognition, decision-making, and execution tasks.

The reviewed literature shows that IoV has substantially extended the observational and computational capabilities of intelligent vehicles, particularly through multimodal sensing, cooperative perception, and edge-cloud support. However, progress remains uneven across the embodied process. Generalizable intention prediction, interaction-aware decision-making, and physically grounded execution are still commonly studied under simplified communication conditions, short prediction horizons, cooperative-agent assumptions, or module-level evaluation. The resulting gap means that improvements in detection, prediction, or planning accuracy do not necessarily translate into safe and effective vehicle behavior during continuous interaction. Embodied intelligence in IoV therefore requires a shift from isolated module optimization and information sharing toward temporally consistent traffic understanding, interaction-aware coordination, and executable closed-loop behavior.

The synthesis is bounded by the rapid evolution of the field and by substantial differences among datasets, communication settings, vehicle platforms, task definitions, and evaluation protocols, which limit direct quantitative comparison across studies. Within these boundaries, four priorities emerge: perception fusion under ambient uncertainty, generalizable intention prediction, policy learning in long-tail autonomous driving scenarios, and experimental validation of vehicle motion control in real-world environments under communication, computing, and physical constraints. Progress in these directions will depend on tracing how uncertainty and errors propagate through the complete embodied loop and on evaluating whether networked intelligence improves task completion and safety in realistic traffic environments. By providing a unified technical organization, a task-oriented resource summary, and a closed-loop research agenda, this survey offers a foundation for developing embodied IoV systems that can move from connected information processing toward sustained and reliable interaction with the physical traffic environment.

Funding Declaration

This work was funded in part by National Natural Science Foundation of China (Grants No. 62466009 and 62562011), in part by Guizhou Provincial Basic Research Program (Grants No. ZK[2024]YiBan048 and No. ZK[2025]Mianshang627), in part by Guizhou Provincial Science and Technology Plan Project (Grants No. QKHRCK-JZY[2025]021 and No. QKHPT-KXJZ[2025]003), in part by Guizhou Provincial Science and Technology Projects (QKHRC-XKBF[2025]014, QKHRC-XKBF[2025]015, BQW[2024]010 QKHZD[2025]zhongda031, and KXJZ[2024]018), in part by National Key Research and Development Program of China (2024YFB3311500) and the National Science and Technology Major Project of China (2025ZD1607103).

References

Abtahi M, Motallebi Araghi F, Mojahed N, et al (2025) Deep bilinear koopman model for real-time vehicle control in Frenet frame. arXiv preprint arXiv:250712578 URL https://arxiv.org/abs/2507.12578

Alshammari T, Mahgoub I (2025) Nature-Inspired algorithms in internet of vehicles: A survey and analysis. IEEE Internet of Things Journal 12(6):6347–6370. https://doi.org/10.1109/jiot.2025.3528872

Anderson ML (2003) Embodied cognition: A field guide. Artificial Intelligence 149(1):91–130. https://doi.org/10.1016/S0004-3702(03)00054-7

Arai H, Miwa K, Sasaki K, et al (2025) CoVLA: Comprehensive Vision-Language-Action dataset for autonomous driving. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp 1933–1943, https://doi.org/10.1109/wacv61041.2025.00195

Bernardin K, Stiefelhagen R (2008) Evaluating multiple object tracking performance: The CLEAR MOT metrics. EURASIP Journal on Image and Video Processing 2008:1–10. https://doi.org/10.1155/2008/246309

Bieshaar M, Zernetsch S, Hubert A, et al (2018) Cooperative starting movement detection of cyclists using convolutional neural networks and a boosted stacking ensemble. IEEE Transactions on Intelligent Vehicles 3(4):534–544. https://doi.org/10.1109/tiv.2018.2873900

Binas J, Neil D, Liu SC, et al (2017) DDD17: End-to-end DAVIS driving dataset. In: ICML 2017 Workshop on Machine Learning for Autonomous Vehicles, Sydney, Australia, URL https://arxiv.org/abs/1711.01458, accessed 29 June 2026

Bokkasam R, Gangisetty S, Hafez AA, et al (2025) Pedestrian intention and trajectory prediction in unstructured traffic using IDD-PeD. In: 2025 IEEE International Conference on Robotics and Automation (ICRA), pp 763–770, https://doi.org/10.1109/icra55743.2025.11128395

Boroujeni SPH, Razi A (2026) VLA4CoDrive: Vision-Language-Action Dataset for Cooperative Autonomous Driving. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops, pp 1789–1799, URL https://openaccess.thecvf.com/content/WACV2026W/LLVM-AD/html/Boroujeni_VLA4CoDrive_Vision-Language-Action_Dataset_for_Cooperative_Autonomous_Driving_WACVW_2026_paper.html, accessed 29 June 2026

Brooks RA (1991) Intelligence without representation. Artificial Intelligence 47(1–3):139–159. https://doi.org/10.1016/0004-3702(91)90053-M

Casas S, Luo W, Urtasun R (2018) IntentNet: Learning to Predict Intention from Raw Sensor Data. In: Proceedings of the 2nd Conference on Robot Learning, Proceedings of Machine Learning Research, vol 87. PMLR, pp 947–956, URL https://proceedings.mlr.press/v87/casas18a.html, accessed 29 June 2026

Chen G, Zhao X, Gao Z, et al (2026) Experimental evaluation of trajectory tracking with nonlinear model predictive control enhanced by physics-informed neural network. IEEE Transactions on Vehicular Technology 75(6):8928–8939. https://doi.org/10.1109/TVT.2025.3635878

Chen L, Wang J, Mortlock T, et al (2025) Hyperdimensional uncertainty quantification for multimodal uncertainty fusion in autonomous vehicles perception. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 22306–22316, URL https://openaccess.thecvf.com/content/CVPR2025/html/Chen_Hyperdimensional_Uncertainty_Quantification_for_Multimodal_Uncertainty_Fusion_in_Autonomous_Vehicles_CVPR_2025_paper.html

Chen Q, Ma X, Tang S, et al (2019) F-Cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3D point clouds. In: Proceedings of the 4th ACM/IEEE Symposium on Edge Computing, pp 88–100, https://doi.org/10.1145/3318216.3363300

Chen X, Zhang T, Wang Y, et al (2023) FUTR3D: A unified sensor fusion framework for 3D detection. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp 172–181, https://doi.org/10.1109/cvprw59228.2023.00022

Chi C, Feng S, Du Y, et al (2023) Diffusion policy: Visuomotor policy learning via action diffusion. In: Proceedings of Robotics: Science and Systems, Daegu, Republic of Korea, https://doi.org/10.15607/RSS.2023.XIX.026

Cho M, Cao Y, Sun J, et al (2025) Cocoon: Robust multi-modal perception with uncertainty-aware sensor fusion. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=69b7p8IQ8A

Cong H, Sigatapu A, Das A, et al (2025) BEVMOSNet: Multimodal fusion for BEV moving object segmentation. In: Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, pp 863–872, https://doi.org/10.5220/0013383300003912

Contreras-Castillo J, Zeadally S, Guerrero-Ibanez JA (2018) Internet of vehicles: Architecture, protocols, and security. IEEE Internet of Things Journal 5(5):3701–3709. https://doi.org/10.1109/JIOT.2017.2690902

Cui E, Wang W, Li Z, et al (2025) DriveMLM: aligning multi-modal large language models with behavioral planning states for autonomous driving. Visual Intelligence 3:22. https://doi.org/10.1007/s44267-025-00095-w

Cui G, Zhang W, Xiao Y, et al (2022) Cooperative perception technology of autonomous driving in the internet of vehicles environment: A review. Sensors 22(15):5535. https://doi.org/10.3390/s22155535

Cunningham AG, Galceran E, Eustice RM, et al (2015) MPDM: Multipolicy decision-making in dynamic, uncertain environments for autonomous driving. In: 2015 IEEE International Conference on Robotics and Automation (ICRA), pp 1670–1677, https://doi.org/10.1109/icra.2015.7139412

Dai Y, Xu D, Maharjan S, et al (2018) Joint offloading and resource allocation in vehicular edge computing and networks. In: 2018 IEEE Global Communications Conference (GLOBECOM), pp 1–7, https://doi.org/10.1109/glocom.2018.8648004

Dal'Col L, Oliveira M, Santos V (2025) Joint perception and prediction for autonomous driving: A survey. IEEE Transactions on Intelligent Transportation Systems 26(12):21427–21452. https://doi.org/10.1109/tits.2025.3607743

Danba S, Bao J, Han G, et al (2022) Toward collaborative intelligence in IoV systems: Recent advances and open issues. Sensors 22(18):6995. https://doi.org/10.3390/s22186995

Duan J, Yu S, Tan HL, et al (2022) A survey of embodied AI: From simulators to research tasks. IEEE Transactions on Emerging Topics in Computational Intelligence 6(2):230–244. https://doi.org/10.1109/tetci.2022.3141105

Ettinger S, Cheng S, Caine B, et al (2021) Large scale interactive motion forecasting for autonomous driving: The Waymo Open Motion Dataset. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, pp 9690–9699, https://doi.org/10.1109/ICCV48922.2021.00957

Feng Z, Xue R, Yuan L, et al (2026) Multi-agent embodied AI: Advances and future directions. Science China Information Sciences 69(5):151202. https://doi.org/10.1007/s11432-025-4820-4

Fu J, Kumar A, Nachum O, et al (2021) D4RL: Datasets for Deep Data-Driven Reinforcement Learning. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=px0-N3_KjA, accessed 29 June 2026

Galceran E, Cunningham AG, Eustice RM, et al (2017) Multipolicy decision-making for autonomous driving via changepoint-based behavior prediction: Theory and experiment. Autonomous Robots 41(6):1367–1382. https://doi.org/10.1007/s10514-017-9619-z

Gao Z, Hao H, Gao F, et al (2025) Behavior-constrained multi-agent proximal policy optimization for cooperative vehicle control at future intersections. Automotive Innovation 8(4):896–912. https://doi.org/10.1007/s42154-024-00308-w

Garcia-Roger D, González EE, Martín-Sacristán D, et al (2020) V2X support in 3GPP specifications: From 4G to 5G and beyond. IEEE Access 8:190946–190963. https://doi.org/10.1109/ACCESS.2020.3028621

Gawlikowski J, Tassi CRN, Ali M, et al (2023) A survey of uncertainty in deep neural networks. Artificial Intelligence Review 56(Suppl 1):1513–1589. https://doi.org/10.1007/s10462-023-10562-9, URL https://doi.org/10.1007/s10462-023-10562-9

Geiger A, Lenz P, Urtasun R (2012) Are we ready for autonomous driving? the KITTI vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp 3354–3361, https://doi.org/10.1109/cvpr.2012.6248074

Girase H, Gang H, Malla S, et al (2021) LOKI: Long term and key intentions for trajectory prediction. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp 9783–9792, https://doi.org/10.1109/iccv48922.2021.00966

Godbole M, Gao X, Tu Z (2025) Drama-X: A Fine-Grained intent prediction and risk reasoning benchmark for driving. In: 2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp 826–831, https://doi.org/10.1109/iccvw69036.2025.00090

Gong H, Yan T, Xie H, et al (2025) Learning stabilizable symplectic ODE-net-based MPC for autonomous vehicle trajectory tracking. Neurocomputing 654:131222. https://doi.org/10.1016/j.neucom.2025.131222

Grimm D, Abouelazm A, Zöllner JM (2025) Goal-based trajectory prediction for improved cross-dataset generalization. arXiv preprint arXiv:250718196 URL https://arxiv.org/abs/2507.18196

Guan Y, Liao H, Li Z, et al (2024) World models for autonomous driving: An initial survey. IEEE Transactions on Intelligent Vehicles pp 1–17. https://doi.org/10.1109/TIV.2024.3398357

Guillen-Perez A, Cano MD (2022) Multi-agent deep reinforcement learning to manage connected autonomous vehicles at tomorrow's intersections. IEEE Transactions on Vehicular Technology 71(7):7033–7043. https://doi.org/10.1109/TVT.2022.3169907

Gupta A, Johnson J, Fei-Fei L, et al (2018) Social GAN: Socially acceptable trajectories with generative adversarial networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 2255–2264, https://doi.org/10.1109/cvpr.2018.00240

Häfner B, Jiru J, Schepker H, et al (2022) Preventing failures of cooperative maneuvers among connected and automated vehicles. Computer Communications 193:234–245. https://doi.org/10.1016/j.comcom.2022.07.013

Han Z, Chen P, Zhou B, et al (2025) Hybrid path tracking control for autonomous trucks: Integrating pure pursuit and deep reinforcement learning with adaptive look-ahead mechanism. IEEE Transactions on Intelligent Transportation Systems 26(5):7098–7112. https://doi.org/10.1109/TITS.2025.3530507

Hao R, Fan S, Dai Y, et al (2024) RCooper: A real-world large-scale dataset for roadside cooperative perception. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 22347–22357, https://doi.org/10.1109/cvpr52733.2024.02109

Hartenstein H, Laberteaux KP (2008) A tutorial survey on vehicular ad hoc networks. IEEE Communications Magazine 46(6):164–171. https://doi.org/10.1109/MCOM.2008.4539481

Heß D, Lattarulo R, Pérez J, et al (2018) Fast maneuver planning for cooperative automated vehicles. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), pp 1625–1632, https://doi.org/10.1109/ITSC.2018.8569791

Honda Research Institute USA (2020) TITAN: Traffic Interaction Dataset. Honda Research Institute USA, URL https://usa.honda-ri.com/titan, accessed 29 June 2026

Hu C, Xie Y, Xie L, et al (2025) Sparse gaussian process-based strategies for two-layer model predictive control in autonomous vehicle drifting. Transportation Research Part C: Emerging Technologies 174:105065. https://doi.org/10.1016/j.trc.2025.105065

Hu S, Fang Z, Tao Y, et al (2026) Multi-agent embodied autonomous driving: From V2X information exchange to shared world models. arXiv preprint arXiv:2606.13840, URL https://arxiv.org/abs/2606.13840, accessed 1 July 2026

Hu Y, Binas J, Neil D, et al (2020) DDD20 End-to-End event camera driving dataset: Fusing frames and events with deep learning for improved steering prediction. In: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), pp 1–6, https://doi.org/10.1109/itsc45102.2020.9294515

Hu Y, Fang S, Lei Z, et al (2022) Where2comm: Communication-Efficient collaborative perception via spatial confidence maps. In: Advances in Neural Information Processing Systems 35, pp 4874–4886, https://doi.org/10.52202/068431-0352

Huang R, Xue H, Pagnucco M, et al (2025a) Vision-Based Multi-Future trajectory prediction: A survey. IEEE Transactions on Neural Networks and Learning Systems 36(8):13691–13708. https://doi.org/10.1109/tnnls.2025.3550350

Huang T, Liu J, Zhou X, et al (2025b) Vehicle-to-Everything cooperative perception for autonomous driving. Proceedings of the IEEE 113(5):443–477. https://doi.org/10.1109/jproc.2025.3600903

Huang X, Wang J, Xia Q, et al (2025c) V2X-R: Cooperative LiDAR-4D radar fusion with denoising diffusion for 3D object detection. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 27390–27400, https://doi.org/10.1109/cvpr52734.2025.02551

Huang Z, Liu H, Lv C (2023) GameFormer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp 3880–3890, https://doi.org/10.1109/iccv51070.2023.00361

Ignatious HA, El-Sayed H, Khan MA, et al (2023) Analyzing factors influencing situation awareness in autonomous vehicles: A survey. Sensors 23(8):4075. https://doi.org/10.3390/s23084075

Jain A, Koppula HS, Raghavan B, et al (2015) Car That Knows Before You Do: Anticipating Maneuvers via Learning Temporal Driving Models. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp 3182–3190, https://doi.org/10.1109/ICCV.2015.364

Jia X, Yang Z, Li Q, et al (2024) Bench2Drive: Towards Multi-Ability benchmarking of Closed-Loop End-To-End autonomous driving. In: Advances in Neural Information Processing Systems 37, pp 819–844, https://doi.org/10.52202/079017-0025

Jiang K, Cai X, Cui Z, et al (2025) KoMA: Knowledge-Driven Multi-Agent framework for autonomous driving with large language models. IEEE Transactions on Intelligent Vehicles 10(10):4655–4668. https://doi.org/10.1109/tiv.2024.3488793

Kaiwartya O, Abdullah AH, Cao Y, et al (2016) Internet of vehicles: Motivation, layered architecture, network model, challenges, and future aspects. IEEE Access 4:5356–5373. https://doi.org/10.1109/ACCESS.2016.2603219

Karnchanachari N, Geromichalos D, Tan KS, et al (2024) Towards learning-based planning: The nuPlan benchmark for real-world autonomous driving. In: 2024 IEEE International Conference on Robotics and Automation (ICRA), pp 629–636, https://doi.org/10.1109/icra57147.2024.10610077

Karuppasamy M, Gangisetty S, Rai SN, et al (2025) Towards safer and understandable driver intention prediction. arXiv preprint arXiv:251009200 URL https://arxiv.org/abs/2510.09200

Khindkar V, Balasubramanian V, Arora C, et al (2024) Can reasons help improve pedestrian intent estimation? a Cross-Modal approach. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp 11515–11522, https://doi.org/10.1109/iros58592.2024.10802097

Kim J, Rohrbach A, Darrell T, et al (2018) Textual explanations for Self-Driving vehicles. In: Ferrari V, Hebert M, Sminchisescu C, et al (eds) Computer Vision –

ECCV 2018, Lecture Notes in Computer Science, vol 11206. Springer, Cham, pp 577–593, https://doi.org/10.1007/978-3-030-01216-8_35

Kotseruba I, Rasouli A, Tsotsos JK (2021) Benchmark for evaluating pedestrian action prediction. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 1258–1268, https://doi.org/10.1109/WACV48630.2021.00130

Kushwaha D, Biron Z (2026) A review on safe reinforcement learning using lyapunov and barrier functions. Artificial Intelligence Review https://doi.org/10.1007/s10462-026-11611-9

Łach Ł (2026) Multi-sensor fusion and deep learning for road scene understanding: a comprehensive survey. Artificial Intelligence Review 59:147. https://doi.org/10.1007/s10462-026-11542-5, URL https://doi.org/10.1007/s10462-026-11542-5

Lee D, Eom C, Kwon M (2024) AD4RL: Autonomous driving benchmarks for offline reinforcement learning with value-based dataset. In: 2024 IEEE International Conference on Robotics and Automation (ICRA), pp 8239–8245, https://doi.org/10.1109/icra57147.2024.10610308

Lei Z, Ren S, Hu Y, et al (2022) Latency-Aware collaborative perception. In: Avidan S, Brostow GJ, Cissé M, et al (eds) Computer Vision – ECCV 2022, Lecture Notes in Computer Science, vol 13692. Springer, Cham, pp 316–332, https://doi.org/10.1007/978-3-031-19824-3_19

Li C, Chan SH, Chen YT (2023) DROID: Driver-Centric risk object IDentification. IEEE Transactions on Pattern Analysis and Machine Intelligence 45(11):13683–13698. https://doi.org/10.1109/TPAMI.2023.3294305

Li D, Zhang Q, Zhao D (2026a) Ranging from prediction to planning via machine learning approaches for autonomous driving: A survey. Artificial Intelligence Review 59:200. https://doi.org/10.1007/s10462-026-11604-8

Li M, Wang B, Wang S, et al (2025a) Serial distributed reinforcement learning for enhanced multi-objective platoon control in curved road coordinates. Expert Systems with Applications 269:126493. https://doi.org/10.1016/j.eswa.2025.126493

Li Q, Wang Y, Wang Y, et al (2022a) HDMapNet: An online HD map construction and evaluation framework. In: 2022 International Conference on Robotics and Automation (ICRA), pp 4628–4634, https://doi.org/10.1109/icra46639.2022.9812383

Li T, Jia P, Wang B, et al (2024) LaneSegNet: Map Learning with Lane Segment Perception for Autonomous Driving. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=LsURkIPYR5, accessed 29 June 2026

Li T, Chen L, Wang H, et al (2026b) Graph-Based Topology Reasoning for Driving Scenes. Science China Information Sciences 69:152103. https://doi.org/10.1007/s11432-025-4815-9

Li Y, Ren S, Wu P, et al (2021) Learning distilled collaboration graph for multi-agent perception. In: Advances in Neural Information Processing Systems, vol 34. Curran Associates, Inc., pp 29541–29552, URL https://proceedings.neurips.cc/paper/2021/hash/f702defbc67edb455949f46babab0c18-Abstract.html, accessed 29 June 2026

Li Y, Ren S, Wu P, et al (2022b) V2X-Sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. IEEE Robotics and Automation Letters 7(4):10914–10921. https://doi.org/10.1109/LRA.2022.3192802

Li Y, Fan C, Ge C, et al (2025b) WOMD-Reasoning: A Large-Scale Dataset for Interaction Reasoning in Driving. In: Proceedings of the 42nd International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 267. PMLR, pp 34288–34311, URL https://proceedings.mlr.press/v267/li25l.html, accessed 29 June 2026

Liao B, Chen S, Yin H, et al (2025) Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 12037–12047, https://doi.org/10.1109/CVPR52734.2025.01124

Lin J, Yu W, Zhang N, et al (2017) A survey on internet of things: Architecture, enabling technologies, security and privacy, and applications. IEEE Internet of Things Journal 4(5):1125–1142. https://doi.org/10.1109/JIOT.2017.2683200

Lisondra M, Benhabib B, Nejat G (2026) Embodied AI with foundation models for mobile service robots: A systematic review. Robotics 15(3):55. https://doi.org/10.3390/robotics15030055

Liu B, Adeli E, Cao Z, et al (2020a) Spatiotemporal relationship reasoning for pedestrian intent prediction. IEEE Robotics and Automation Letters 5(2):3485–3492. https://doi.org/10.1109/LRA.2020.2976305

Liu G, Hu Y, Xu C, et al (2025a) Toward collaborative autonomous driving: Simulation platform and End-to-End system. IEEE Transactions on Pattern Analysis and Machine Intelligence 47(8):6566–6584. https://doi.org/10.1109/tpami.2025.3560327

Liu L, Chen C, Pei Q, et al (2021) Vehicular edge computing and networking: A survey. Mobile Networks and Applications 26:1145–1168. https://doi.org/10.1007/s11036-020-01624-1

Liu Q, Liu Y, Liu C, et al (2018) Hierarchical lateral control scheme for autonomous vehicle with uneven time delays induced by vision sensors. Sensors 18(8):2544. https://doi.org/10.3390/s18082544

Liu S, Chen W, Li W, et al (2026) Bridgedrive: Diffusion bridge policy for closed-loop trajectory planning in autonomous driving. In: The Fourteenth International Conference on Learning Representations, URL https://arxiv.org/abs/2509.23589

Liu Y, Huang Q, Li R, et al (2024) Select2Col: Leveraging spatial-temporal importance of semantic information for efficient collaborative perception. IEEE Transactions on Vehicular Technology 73(9):12556–12569. https://doi.org/10.1109/TVT.2024.3390414

Liu Y, Cao X, Chen T, et al (2025b) From screens to scenes: A survey of embodied AI in healthcare. Information Fusion 119:103033. https://doi.org/10.1016/j.inffus.2025.103033

Liu Y, Chen W, Bai Y, et al (2025c) Aligning cyber space with physical world: A comprehensive survey on embodied AI. IEEE/ASME Transactions on Mechatronics 30(6):7253–7274. https://doi.org/10.1109/tmech.2025.3574943

Liu YC, Tian J, Glaser N, et al (2020b) When2com: Multi-agent perception via communication graph grouping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4106–4115, https://doi.org/10.1109/CVPR42600.2020.00416

Liu YC, Tian J, Ma CY, et al (2020c) Who2com: Collaborative perception via learnable handshake communication. In: Proceedings of the IEEE International Conference on Robotics and Automation, pp 6876–6883, https://doi.org/10.1109/ICRA40945.2020.9197364

Lu Y, Li Q, Liu B, et al (2023) Robust collaborative 3D object detection in presence of pose errors. In: Proceedings of the IEEE International Conference on Robotics and Automation, pp 4812–4818, https://doi.org/10.1109/ICRA48891.2023.10160546

Luo C, Liu Y, Sun Z (2026) Ensuring safety in cooperative CAV lane change: A multi-agent reinforcement learning strategy with control barrier function. In: Intelligent Transportation Engineering: Proceedings of the 10th International Conference (ICITE 2025), Advances in Transdisciplinary Engineering, https://doi.org/10.3233/ATDE251474

Ma Y, Wei T, Zhong N, et al (2025) LeapVAD: A leap in autonomous driving via cognitive perception and Dual-Process thinking. IEEE Transactions on Neural Networks and Learning Systems 37(4):1963–1977. https://doi.org/10.1109/tnnls.2025.3626711

Ma Y, Song Z, Zhuang Y, et al (2026) A survey on vision-language-action models for embodied AI. IEEE Transactions on Neural Networks and Learning Systems https://doi.org/10.1109/TNNLS.2025.3650584

Malik S, Khan MA, El-Sayed H (2021) Collaborative autonomous driving–a survey of solution approaches and future challenges. Sensors 21(11):3783. https://doi.org/10.3390/s21113783

Malla S, Choi C, Dwivedi I, et al (2023) DRAMA: Joint risk localization and captioning in driving. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp 1043–1052, https://doi.org/10.1109/wacv56688.2023.00110

Man Y, Gui LY, Wang YX (2023) BEV-Guided Multi-Modality fusion for driving perception. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 21960–21969, https://doi.org/10.1109/cvpr52729.2023.02103

Manzinger S, Althoff M (2017) Negotiation of drivable areas of cooperative vehicles for conflict resolution. In: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), pp 1–8, https://doi.org/10.1109/itsc.2017.8317765

Manzinger S, Althoff M (2018) Tactical decision making for cooperative vehicles using reachable sets. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), pp 444–451, https://doi.org/10.1109/itsc.2018.8569560

Mao R, Guo J, Jia Y, et al (2023) DOLPHINS: Dataset for collaborative perception enabled harmonious and interconnected self-driving. In: Wang L, Gall J, Chin TJ, et al (eds) Computer Vision – ACCV 2022, Lecture Notes in Computer Science, vol 13845. Springer, Cham, pp 495–511, https://doi.org/10.1007/978-3-031-26348-4_29

Ming Z, Stephany Berrio J, Shan M, et al (2025) OccFusion: Multi-Sensor fusion framework for 3D semantic occupancy prediction. IEEE Transactions on Intelligent Vehicles 10(5):3421–3433. https://doi.org/10.1109/tiv.2024.3453293

Mirchevska B, Hugle M, Kalweit G, et al (2021) Amortized q-learning with model-based action proposals for autonomous driving on highways. In: 2021 IEEE International Conference on Robotics and Automation (ICRA), pp 1028–1035, https://doi.org/10.1109/icra48506.2021.9560777

Mizutani M, Tsukada M, Esaki H (2021) AutoMCM: Maneuver coordination service with abstracted functions for autonomous driving. In: 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), pp 1069–1076, https://doi.org/10.1109/itsc48978.2021.9564556

Molina-Masegosa R, Avedisov SS, Sepulcre M, et al (2023) V2X Communications for Maneuver Coordination in Connected Automated Driving: Message Generation Rules. IEEE Vehicular Technology Magazine 18(3):91–100. https://doi.org/10.1109/MVT.2023.3284562

Mukhopadhyay S, Roy-Chowdhury A, Qiu H (2026) CooperTrim: Adaptive data selection for uncertainty-aware cooperative perception. In: International Conference on

Learning Representations, URL https://openreview.net/forum?id=8NgKNuHRiH

Newell A, Simon HA (1976) Computer science as empirical inquiry: Symbols and search. Communications of the ACM 19(3):113–126. https://doi.org/10.1145/360018.360022

Nichting M, Hess D, Schindler J, et al (2020) Space time reservation procedure (STRP) for V2X-Based maneuver coordination of cooperative automated vehicles in diverse conflict scenarios. In: 2020 IEEE Intelligent Vehicles Symposium (IV), pp 502–509, https://doi.org/10.1109/iv47402.2020.9304769

Nie M, Peng R, Wang C, et al (2025) Reason2Drive: Towards interpretable and Chain-Based reasoning for autonomous driving. In: Leonardis A, Ricci E, Roth S, et al (eds) Computer Vision – ECCV 2024, Lecture Notes in Computer Science, vol 15084. Springer, Cham, pp 292–308, https://doi.org/10.1007/978-3-031-73347-5_17

Nie T, Tang Y, He J, et al (2026) ADV-0: Closed-loop min-max adversarial training for long-tail robustness in autonomous driving. arXiv preprint arXiv:260315221 https://doi.org/10.48550/arXiv.2603.15221, URL https://arxiv.org/abs/2603.15221

Ning J, Behl M (2024) DKMGP: A Gaussian Process Approach to Multi-Task and Multi-Step Vehicle Dynamics Modeling in Autonomous Racing. arXiv preprint arXiv:241113755 URL https://arxiv.org/abs/2411.13755

Paden B, Čáp M, Yong SZ, et al (2016) A survey of motion planning and control techniques for self-driving urban vehicles. IEEE Transactions on Intelligent Vehicles 1(1):33–55. https://doi.org/10.1109/TIV.2016.2578706

Paek DH, Kong SH (2025) Availability-aware sensor fusion via unified canonical space. In: Advances in Neural Information Processing Systems, vol 38. Curran Associates, Inc., URL https://openreview.net/forum?id=foy0BWzOFL, accessed 29 June 2026

Pfeifer R, Iida F (2004) Embodied artificial intelligence: Trends and challenges. In: Iida F, Pfeifer R, Steels L, et al (eds) Embodied Artificial Intelligence, Lecture Notes in Computer Science, vol 3139. Springer, Berlin, Heidelberg, pp 1–26, https://doi.org/10.1007/978-3-540-27833-7_1

Rasouli A, Kotseruba I, Tsotsos JK (2017) Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), pp 206–213, https://doi.org/10.1109/ICCVW.2017.33

Rasouli A, Kotseruba I, Kunic T, et al (2019) PIE: A Large-Scale dataset and models for pedestrian intention estimation and trajectory prediction. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp 6261–6270, https://doi.org/10.1109/iccv.2019.00636

Ristani E, Solera F, Zou R, et al (2016) Performance measures and a data set for multi-target, multi-camera tracking. In: Hua G, Jégou H (eds) Computer Vision – ECCV 2016 Workshops, Lecture Notes in Computer Science, vol 9914. Springer, Cham, pp 17–35, https://doi.org/10.1007/978-3-319-48881-3_2

Rowe L, Ethier M, Dykhne EH, et al (2023) FJMP: Factorized joint Multi-Agent motion prediction over learned directed acyclic interaction graphs. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 13745–13755, https://doi.org/10.1109/cvpr52729.2023.01321

Ruan H, Yu H, Yang W, et al (2024) Learning cooperative trajectory representations for motion forecasting. In: Advances in Neural Information Processing Systems 37, pp 13430–13457, https://doi.org/10.52202/079017-0429

Salzmann T, Ivanovic B, Chakravarty P, et al (2020) Trajectron++: Dynamically-Feasible trajectory forecasting with heterogeneous data. In: Vedaldi A, Bischof H, Brox T, et al (eds) Computer Vision – ECCV 2020, Lecture Notes in Computer Science, vol 12363. Springer, Cham, pp 683–700, https://doi.org/10.1007/978-3-030-58523-5_40

Sauer A, Savinov N, Geiger A (2018) Conditional affordance learning for driving in urban environments. In: Proceedings of the 2nd Conference on Robot Learning, Proceedings of Machine Learning Research, vol 87. PMLR, pp 237–252, URL https://proceedings.mlr.press/v87/sauer18a.html, accessed 29 June 2026

Schwarting W, Pierson A, Alonso-Mora J, et al (2019) Social behavior for autonomous vehicles. Proceedings of the National Academy of Sciences of the United States of America 116(50):24972–24978. https://doi.org/10.1073/pnas.1820676116

Shah G, Tian D, Moradi-Pari E, et al (2024) AROW: V2X-Based automated Right-of-Way algorithm for cooperative intersection management. IEEE Transactions on Intelligent Transportation Systems 25(9):10983–10999. https://doi.org/10.1109/tits.2024.3384982

Shi S, Jiang L, Dai D, et al (2022) Motion transformer with global intention localization and local movement refinement. In: Advances in Neural Information Processing Systems 35, pp 6531–6543, https://doi.org/10.52202/068431-0473

Shi W, Dong S, Zhang C, et al (2025) Learning-based real-time model predictive tracking control for autonomous vehicles with path-pattern adaptability. Control Engineering Practice 164:106480. https://doi.org/10.1016/j.conengprac.2025.106480

Sima C, Renz K, Chitta K, et al (2025) DriveLM: Driving with graph visual question answering. In: Leonardis A, Ricci E, Roth S, et al (eds) Computer Vision – ECCV 2024, Lecture Notes in Computer Science, vol 15110. Springer, Cham, pp 256–274, https://doi.org/10.1007/978-3-031-72943-0_15

Singh G, Akrigg S, Maio MD, et al (2023) ROAD: The road event awareness dataset for autonomous driving. IEEE Transactions on Pattern Analysis and Machine Intelligence 45(1):1036–1054. https://doi.org/10.1109/tpami.2022.3150906

So O, Serlin Z, Mann M, et al (2024) How to train your neural control barrier function: Learning safety filters for complex input-constrained systems. In: 2024 IEEE International Conference on Robotics and Automation (ICRA), pp 11532–11539, https://doi.org/10.1109/ICRA57147.2024.10610418

Song R, Liang C, Cao H, et al (2024) Collaborative semantic occupancy prediction with hybrid feature fusion in connected automated vehicles. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 17996–18006, https://doi.org/10.1109/cvpr52733.2024.01704

Song Z, Wen F, Zhang H, et al (2023) A cooperative perception system robust to localization errors. In: Proceedings of the IEEE Intelligent Vehicles Symposium, pp 1–6, https://doi.org/10.1109/IV55152.2023.10186727

Suman V, Pham P, Bera A (2023) RAIST: Learning risk aware traffic interactions via Spatio-Temporal graph convolutional networks. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp 5545–5550, https://doi.org/10.1109/iros55552.2023.10341578

Sun F, Chen R, Ji T, et al (2024) A comprehensive survey on embodied intelligence: Advancements, challenges, and future perspectives. CAAI Artificial Intelligence Research 3(1):9150042. https://doi.org/10.26599/AIR.2024.9150042

Sun Q, Huang X, Gu J, et al (2022) M2I: From factored marginal trajectory prediction to interactive prediction. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 6533–6542, https://doi.org/10.1109/cvpr52688.2022.00643

Tang P, Wang Z, Wang G, et al (2024) SparseOcc: Rethinking sparse latent representation for vision-based semantic occupancy prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 15035–15044, https://doi.org/10.1109/CVPR52733.2024.01424

Vadivelu N, Ren M, Tu J, et al (2021) Learning to communicate and correct pose errors. In: Proceedings of the 2020 Conference on Robot Learning, Proceedings of Machine Learning Research, vol 155. PMLR, pp 1195–1210, URL https://proceedings.mlr.press/v155/vadivelu21a.html, accessed 29 June 2026

Wan J, Liu H, Xu M, et al (2024) Lane-changing tracking control of automated vehicle platoon based on MA-DDPG and adaptive MPC. IEEE Access 12:58078–58096. https://doi.org/10.1109/ACCESS.2024.3381629

Wang B, Han Y, Wang S, et al (2022) A review of intelligent connected vehicle cooperative driving development. Mathematics 10(19):3635. https://doi.org/10.3390/math10193635

Wang C, Zhou S, Wang L, et al (2025a) Autonomous driving via knowledge-enhanced safe reinforcement learning. IEEE Transactions on Intelligent Vehicles 10(1):359–372. https://doi.org/10.1109/TIV.2024.3412916

Wang H, Li T, Li Y, et al (2023) OpenLane-V2: A topology reasoning benchmark for unified 3D HD mapping. In: Advances in Neural Information Processing Systems 36, pp 18873–18884, https://doi.org/10.52202/075280-0827

Wang J, Nordström T (2025) Latency robust cooperative perception using asynchronous feature fusion. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp 4862–4871, https://doi.org/10.1109/WACV61041.2025.00476

Wang J, Zhang Q, Zhao D, et al (2019) Lane change decision-making through deep reinforcement learning with rule-based constraints. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp 1–6, https://doi.org/10.1109/ijcnn.2019.8852110

Wang J, Malawade AV, Zhou J, et al (2024a) RS2G: Data-Driven Scene-Graph extraction and embedding for robust autonomous perception and scenario understanding. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp 7478–7487, https://doi.org/10.1109/wacv57701.2024.00732

Wang J, Wang H, Song J, et al (2025b) Knowledge-guided self-learning control strategy for mixed vehicle platoons with delays. Nature Communications 16:7705. https://doi.org/10.1038/s41467-025-62597-x

Wang J, Jiang Z, Sun W, et al (2026) SparseCoop: Cooperative perception with Kinematic-Grounded queries. Proceedings of the AAAI Conference on Artificial Intelligence 40(12):9876–9884. https://doi.org/10.1609/aaai.v40i12.37952

Wang P, Liu D, Chen J, et al (2021) Decision making for autonomous driving via augmented adversarial inverse reinforcement learning. In: 2021 IEEE International Conference on Robotics and Automation (ICRA), pp 1036–1042, https://doi.org/10.1109/icra48506.2021.9560907

Wang S, Wang Z, Jiang R, et al (2024b) A multi-agent reinforcement learning-based longitudinal and lateral control of CAVs to improve traffic efficiency in a mandatory lane change scenario. Transportation Research Part C: Emerging Technologies 158:104445. https://doi.org/10.1016/j.trc.2023.104445

Wang T, Kim S, Ji W, et al (2024c) DeepAccident: A motion and accident prediction benchmark for V2X autonomous driving. Proceedings of the AAAI Conference on

Artificial Intelligence 38(6):5599–5606. https://doi.org/10.1609/aaai.v38i6.28370

Wang TH, Manivasagam S, Liang M, et al (2020) V2VNet: Vehicle-to-Vehicle communication for joint perception and prediction. In: Vedaldi A, Bischof H, Brox T, et al (eds) Computer Vision – ECCV 2020, Lecture Notes in Computer Science, vol 12347. Springer, Cham, pp 605–621, https://doi.org/10.1007/978-3-030-58536-5_36

Wang W, Wang J, Deng Y, et al (2025c) BEVSync: Asynchronous data alignment for camera-based vehicle-infrastructure cooperative perception under uncertain delays. Proceedings of the AAAI Conference on Artificial Intelligence 39(14):14699–14707. https://doi.org/10.1609/aaai.v39i14.33611

Wang Y, He J, Fan L, et al (2024d) Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 14749–14759, https://doi.org/10.1109/cvpr52733.2024.01397

Wang Y, Huang X, Sun X, et al (2025d) Uniocc: a unified benchmark for occupancy forecasting and prediction in autonomous driving. In: 2025 IEEE/CVF International Conference on Computer Vision (ICCV), pp 25560–25570, https://doi.org/10.1109/iccv51701.2025.02371

Wei S, Wei Y, Hu Y, et al (2023) Asynchrony-Robust collaborative perception via bird's eye view flow. In: Advances in Neural Information Processing Systems 36, pp 28462–28477, https://doi.org/10.52202/075280-1236

Wei Z, Zheng L, Liu J, et al (2026) MS-Occ: Multi-Stage LiDAR-Camera fusion for 3D semantic occupancy prediction. IEEE Robotics and Automation Letters 11(1):370–377. https://doi.org/10.1109/lra.2025.3632759

Wen L, Fu D, Li X, et al (2024) DiLu: A Knowledge-Driven Approach to Autonomous Driving with Large Language Models. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=OqTMUPuLuC, accessed 29 June 2026

Wilson B, Qi W, Agarwal T, et al (2021) Argoverse 2: Next generation datasets for self-driving perception and forecasting. In: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (Round 2), URL https://openreview.net/forum?id=vKQGe36av4k, accessed 29 June 2026

Wilson M (2002) Six views of embodied cognition. Psychonomic Bulletin & Review 9(4):625–636. https://doi.org/10.3758/BF03196322

Wong LHK, Kang X, Bai K, et al (2025) A survey of robotic navigation and manipulation with physics simulators in the era of embodied AI. arXiv preprint arXiv:2505.01458, URL https://arxiv.org/abs/2505.01458, accessed 1 July 2026

Xia Q, Chen P, Xu G, et al (2025) Adaptive path-tracking controller embedded with reinforcement learning and preview model for autonomous driving. IEEE Transactions on Vehicular Technology 74(3):3736–3750. https://doi.org/10.1109/TVT.2024.3502640

Xiang H, Zheng Z, Xia X, et al (2025) V2X-Real: A Large-Scale dataset for Vehicle-to-Everything cooperative perception. In: Leonardis A, Ricci E, Roth S, et al (eds) Computer Vision – ECCV 2024, Lecture Notes in Computer Science, vol 15110. Springer, Cham, pp 455–470, https://doi.org/10.1007/978-3-031-72943-0_26

Xie Y, Xu C, Rakotosaona MJ, et al (2023) SparseFusion: Fusing multi-modal sparse representations for multi-sensor 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 17591–17602, https://doi.org/10.1109/ICCV51070.2023.01613

Xu K, Zhao H, Hu R, et al (2025) Embodied Intelligence for Flexible Manufacturing: A Survey. Robot 47(4):581–624. https://doi.org/10.13973/j.cnki.robot.250208

Xu R, Xiang H, Tu Z, et al (2022) V2X-ViT: Vehicle-to-Everything cooperative perception with vision transformer. In: Avidan S, Brostow GJ, Cissé M, et al (eds) Computer Vision – ECCV 2022, Lecture Notes in Computer Science, vol 13699. Springer, Cham, pp 107–124, https://doi.org/10.1007/978-3-031-19842-7_7

Xu R, Tu Z, Xiang H, et al (2023a) CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers. In: Proceedings of the 6th Conference on Robot Learning, Proceedings of Machine Learning Research, vol 205. PMLR, pp 989–1000, URL https://proceedings.mlr.press/v205/xu23a.html, accessed 29 June 2026

Xu R, Xia X, Li J, et al (2023b) V2V4Real: A Real-World Large-Scale dataset for Vehicle-to-Vehicle cooperative perception. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 13712–13722, https://doi.org/10.1109/cvpr52729.2023.01318

Xu X, Liu K, Dai P, et al (2023c) Joint task offloading and resource optimization in NOMA-based vehicular edge computing: A game-theoretic DRL approach. Journal of Systems Architecture 134:102780. https://doi.org/10.1016/j.sysarc.2022.102780

Xu Y, Yang X, Gong L, et al (2020) Explainable Object-Induced action decision for autonomous vehicles. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 9520–9529, https://doi.org/10.1109/cvpr42600.2020.00954

Xu Y, Shi Y, Chen S, et al (2026) DCT-MARL: Dynamic communication topology adaptation for cooperative vehicle platooning under non-ideal V2V communications. Computer Networks 286:112439. https://doi.org/10.1016/j.comnet.2026.112439

Xu Z, Wu K, Wen J, et al (2024) A survey on robotics with foundation models: Toward embodied AI. arXiv preprint arXiv:2402.02385, URL https://arxiv.org/abs/2402.02385, accessed 1 July 2026

Yan J, Liu Y, Sun J, et al (2023a) Cross modal transformer: Towards fast and robust 3D object detection. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp 18222–18232, https://doi.org/10.1109/iccv51070.2023.01675

Yan T, Wu D, Han W, et al (2025) DrivingSphere: Building a high-fidelity 4D world for closed-loop simulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 27531–27541, URL https://openaccess.thecvf.com/content/CVPR2025/html/Yan_DrivingSphere_Building_a_High-fidelity_4D_World_for_Closed-loop_Simulation_CVPR_2025_paper.html

Yan Z, Li P, Fu Z, et al (2023b) INT2: Interactive trajectory prediction at intersections. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp 8502–8513, https://doi.org/10.1109/iccv51070.2023.00784

Yang D, Yang K, Wang Y, et al (2023) How2comm: Communication-Efficient and Collaboration-Pragmatic Multi-Agent perception. In: Advances in Neural Information Processing Systems 36, pp 25151–25164, https://doi.org/10.52202/075280-1093

Yang L, Zhang X, Li J, et al (2025a) V2X-Radar: A Multi-Modal Dataset with 4D Radar for Cooperative Perception. In: Advances in Neural Information Processing Systems, Datasets and Benchmarks Track, vol 38. Curran Associates, Inc., URL https://openreview.net/forum?id=sTKsFIVqik, accessed 29 June 2026

Yang SG, Kim J, Lim SC (2026) High-accuracy path tracking for autonomous vehicle navigation: A hierarchical deep reinforcement learning approach. IEEE Transactions on Vehicular Technology 75(2):3342–3347. https://doi.org/10.1109/TVT.2025.3600091

Yang X, Xiong L, Leng B, et al (2020) Design, validation and comparison of path following controllers for autonomous vehicles. Sensors 20(21):6052. https://doi.org/10.3390/s20216052

Yang Y, Mei J, Ma Y, et al (2025b) Driving in the occupancy world: Vision-Centric 4D occupancy forecasting and planning via world models for autonomous driving. Proceedings of the AAAI Conference on Artificial Intelligence 39(9):9327–9335. https://doi.org/10.1609/aaai.v39i9.33010

Yao X, Du Z, Sun Z, et al (2024a) Cooperative lane-changing in mixed traffic: A deep reinforcement learning approach. Transportmetrica A: Transport Science 22(1):1. https://doi.org/10.1080/23249935.2024.2343048

Yao Y, Yan S, Goehring D, et al (2024b) Improving out-of-distribution generalization of trajectory prediction for autonomous driving via polynomial representations.

arXiv preprint arXiv:240713431 URL https://arxiv.org/abs/2407.13431

Yin J, So O, Yu EY, et al (2025) Safe beyond the horizon: Efficient sampling-based MPC with neural control barrier functions. In: Proceedings of Robotics: Science and Systems, Los Angeles, CA, USA, https://doi.org/10.15607/RSS.2025.XXI.112

Yu H, Luo Y, Shu M, et al (2022a) DAIR-V2X: A Large-Scale dataset for Vehicle-Infrastructure cooperative 3D object detection. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 21329–21338, https://doi.org/10.1109/cvpr52688.2022.02067

Yu H, Yang W, Ruan H, et al (2023) V2X-Seq: A Large-Scale sequential dataset for Vehicle-Infrastructure cooperative perception and forecasting. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 5486–5495, https://doi.org/10.1109/cvpr52729.2023.00531

Yu K, Fu M, Zhang T, et al (2025a) Enhancing safety in autonomous racing with constrained reinforcement learning. IEEE Robotics and Automation Letters 10(6):6448–6455. https://doi.org/10.1109/LRA.2025.3566591

Yu SY, Malawade AV, Muthirayan D, et al (2022b) Scene-Graph augmented Data-Driven risk assessment of autonomous vehicle decisions. IEEE Transactions on Intelligent Transportation Systems 23(7):7941–7951. https://doi.org/10.1109/tits.2021.3074854

Yu Y, Yan Y, Jin Y (2025b) Structural knowledge: from brain to artificial intelligence. Artificial Intelligence Review 58:264. https://doi.org/10.1007/s10462-025-11270-2

Zhan W, Sun L, Wang D, et al (2019) INTERACTION Dataset: An International, Adversarial and Cooperative Motion Dataset in Interactive Driving Scenarios with Semantic Maps. INTERACTION Dataset, URL https://interaction-dataset.com/, accessed 29 June 2026

Zhang B, Song N, Jin X, et al (2025a) Bridging past and future: End-to-End autonomous driving with historical prediction and planning. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 6854–6863, https://doi.org/10.1109/cvpr52734.2025.00643

Zhang D, Yuan Q, Meng L, et al (2026) Reinforcement learning for single-agent to multi-agent systems: from basic theory to industrial application progress, a survey. Artificial Intelligence Review 59:46. https://doi.org/10.1007/s10462-025-11439-9

Zhang L, Ding W, Chen J, et al (2020) Efficient uncertainty-aware decision-making for automated driving using guided branching. In: 2020 IEEE International Conference on Robotics and Automation (ICRA), pp 3291–3297, https://doi.org/10.1109/icra40945.2020.9197302

Zhang X, Zhou Z, Wang Z, et al (2025b) Co-MTP: A cooperative trajectory prediction framework with Multi-Temporal fusion for autonomous driving. In: 2025 IEEE International Conference on Robotics and Automation (ICRA), pp 801–807, https://doi.org/10.1109/icra55743.2025.11127303

Zhang Y, Tian J, Xiong Q (2025c) A review of embodied intelligence systems: A three-layer framework integrating multimodal perception, world modeling, and structured strategies. Frontiers in Robotics and AI 12:1668910. https://doi.org/10.3389/frobt.2025.1668910

Zhang Y, Zheng Z, Liu J, et al (2025d) MIC-BEV: Infrastructure-based multi-camera bird's-eye-view perception transformer for 3D object detection. In: ICCV 2025 DriveX Workshop, URL https://drivex-workshop.github.io/iccv2025/static/papers/15_MIC_BEV_Infrastructure_Base.pdf, accessed 29 June 2026

Zhang Z, Zhang Y, Ding X, et al (2025e) Online vectorized hd map construction using geometry. In: Computer Vision – ECCV 2024, Lecture Notes in Computer Science, vol 15107. Springer, pp 73–90, https://doi.org/10.1007/978-3-031-72967-6_5

Zhao R, Wang K, Li Y, et al (2025) Centralized cooperative control for autonomous vehicles at unsignalized all-directional intersections: A multi-agent projection-based constrained policy optimization approach. Expert Systems with Applications 267:126153. https://doi.org/10.1016/j.eswa.2024.126153

Zhao S, Zhang J, Masoud N, et al (2024) High-speed cornering control and real-vehicle deployment for autonomous electric vehicles. arXiv preprint arXiv:241111762 URL https://arxiv.org/abs/2411.11762

Zheng W, Chen W, Huang Y, et al (2025a) OccWorld: Learning a 3D occupancy world model for autonomous driving. In: Leonardis A, Ricci E, Roth S, et al (eds) Computer Vision – ECCV 2024, Lecture Notes in Computer Science, vol 15071. Springer, Cham, pp 55–72, https://doi.org/10.1007/978-3-031-72624-8_4

Zheng Y, Liang R, Zheng K, et al (2025b) Diffusion-Based Planning for Autonomous Driving with Flexible Guidance. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=wM2sfVgMDH, accessed 29 June 2026

Zheng Y, Yao L, Su Y, et al (2025c) A survey of embodied learning for object-centric robotic manipulation. Machine Intelligence Research 22:588–626. https://doi.org/10.1007/s11633-025-1542-8

Zhou H, Xu W, Chen J, et al (2020) Evolutionary V2X technologies toward the internet of vehicles: Challenges and opportunities. Proceedings of the IEEE 108(2):308–323. https://doi.org/10.1109/JPROC.2019.2961937

Zhou J, Yan L, Liang J, et al (2026) Enforcing cooperative safety for reinforcement learning-based mixed-autonomy platoon control. IEEE Transactions on Intelligent Transportation Systems 27(1):1592–1605. https://doi.org/10.1109/TITS.2025.3627592

Zhu M, Zhu Y, Zhu Y, et al (2026) M3CAD: Towards Generic Cooperative Autonomous Driving Benchmark. In: 2026 IEEE International Conference on Robotics and Automation (ICRA), accepted for publication, URL https://arxiv.org/abs/2505.06746, accessed 29 June 2026

Zia MU, Xiang W, Huang T, et al (2026) Unifying ground and air: a comprehensive review of deep learning-enabled CAVs and UAVs. Artificial Intelligence Review 59:19. https://doi.org/10.1007/s10462-025-11425-1, URL https://doi.org/10.1007/s10462-025-11425-1

Zimmer W, Wardana GA, Sritharan S, et al (2024) TUMTraf V2X cooperative perception dataset. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 22668–22677, https://doi.org/10.1109/cvpr52733.2024.02139