Engineering 59 (2026) 337–351

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Safety for Intelligent and Connected Vehicles—Article

Embodied Interactive Intelligence Towards Autonomous Driving

Nan Ma [a,b,*], Jia Pan [c], Yongjin Liu [d], Yajue Yang [a,b], Yiheng Han [a,b], Jiacheng Guo [a,b], Zhixuan Wu [e], Zecheng Yang [f], Zhiwei Yang [g], Deyi Li [d,*]

[a] School of Information Science and Technology, Beijing University of Technology, Beijing 100124, China
[b] Beijing Key Laboratory of Embodied Interactive Intelligence, Beijing University of Technology, Beijing 100124, China
[c] Faculty of Engineering, The University of Hong Kong, Hong Kong 999077, China
[d] Department of of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[e] School of Computer Science, Beijing University of Posts and Technology, Beijing 100876, China
[f] Robotics College, Beijing Union University, Beijing 100101, China
[g] Dongfeng Usharing Technology Co., Ltd., Wuhan 430199, China

ARTICLE INFO

Article history:
Received 10 December 2024
Revised 30 August 2025
Accepted 23 September 2025
Available online 3 December 2025

Keywords:
Embodied interactive intelligence
Autonomous driving
Cognition behaviour
Continuous learning
Hypergraph learning

ABSTRACT

Autonomous driving depends on successful interactions among humans, vehicles, and roads. However, people often lack an understanding of autonomous vehicle (AV) behaviours and decisions. Moreover, AVs have difficulty aligning with human intentions in their interactions. To overcome the obstacles associated with the absence of interactive intelligence, especially in complex and uncertain environments, we introduce the concept of embodied interactive intelligence towards autonomous driving (EIIAD), which establishes representation and learning methods aligned with the physical world, enhancing human–machine integration. Building on this concept, we propose an end-to-end unified constrained vehicle environment interaction (UniCVE) model, which involves the construction of an end-to-end perception–cognition–behaviour closed-loop feedback paradigm and continuous learning through accumulated split driving scenarios. This model realizes interaction cognition through networks designed for pedestrians and vehicles, and it unifies the cognition as a value network of AVs to generate socially compatible behaviours. The UniCVE model is implemented on Dongfeng autonomous buses, which have successfully travelled 22 thousand kilometres and completed 45 thousand navigation tasks in Xiong'an New Area, China, demonstrating its general applicability in various driving scenarios. In addition, we highlight the high-level interactive intelligence of the UniCVE model in selected simulated complex interaction scenarios, demonstrating that it makes AVs more intelligent, more reliable, and more attuned to human relationships. Furthermore, the UniCVE model's capacity for self-learning and self-growth allows it to infinitely approximate true intelligence, even with limited experience.

1. Introduction

Autonomous driving (AD) is transforming human travel and lifestyles. The complexity of a realistic driving environment, characterized by the presence of multiple traffic participants, requires autonomous vehicles (AVs) to be equipped not only with specific driving competencies but also with a profound capacity for interactive cognition [1–3]. Real-world interactions, such as a pedestrian waving at a taxi or an oncoming driver signalling to yield, are ubiquitous. As AVs increase in prevalence, they will need to adapt to diverse cultures and regions, broadening the spectrum of interactive cognition needs.

The questions of whether AVs can genuinely comprehend human behaviours and whether people can understand a vehicle's decision-making strategy present ongoing interaction challenges [4–6]. If harmony between humans and AVs is not established, the reliability of autonomous driving may be doubted. Hence, for AVs to gain widespread acceptance, they must align with social norms and demonstrate the abilities to interact, learn, and build trust. We introduce the concept of embodied interactive intelligence towards autonomous driving (EIIAD), which encapsulates the crucial cognitive and behavioural intelligence necessary for AVs to integrate seamlessly into human society.

* Corresponding authors.
E-mail addresses: manan123@bjut.edu.cn (N. Ma), lidy@cae.cn (D. Li).

https://doi.org/10.1016/j.eng.2025.09.032

Embodied intelligence, defined as the purposeful exchange of energy and information with a physical environment [7–9], has been the focus of emerging academic research. Various frameworks have been proposed to enhance the ability of embodied intelligence agents to evolve and interact with the real world more effectively. Computational frameworks, such as deep evolutionary reinforcement learning (RL) for modeling morphological evolution [10] and mathematical physics models for soft robots [11], underscore the importance of environmental interactions, demonstrating how embodied intelligence evolves through adaptation to surroundings. Therefore, we propose embodied interactive cognition, which uses techniques such as cross-modality perception, machine learning, cognitive computing, and generative artificial intelligence to construct unified intelligent expressions and learning methodologies, thereby enhancing machine intelligence and promoting human–machine harmony [12,13].

As embodied intelligence agents within society, AVs must engage in bidirectional interactions with their environments, including humans, vehicles, and roads. Deficiencies in interaction capabilities, such as the inability to accurately discern pedestrian intentions, can result in significant traffic issues, as noted in the literature [14–16]. AD research is confronted with two primary challenges: attaining the level of a human driver's cognitive interaction, particularly in complex scenarios, and mastering the human driver's ability to adapt in real time during unforeseen emergencies. To address these challenges, AVs must exhibit advanced embodied interaction intelligence, which is characterized by cognitive and behavioural capabilities, such as decision-making, interaction cognition, and life-long learning [17–19].

Inspired by Li [20], we propose the concept of EIIAD. This concept can be considered a perception–cognition–behaviour closed-loop feedback paradigm that enables AVs to continuously construct and optimize mappings between physical and cognitive spaces through actively interacting with their environments and gathering feedback (Fig. 1(a)). The driving environment, also known as the physical space, typically includes a variety of dynamic and static traffic participants, all of which significantly influence the driving policies of AVs. Dynamic elements, such as pedestrians and other vehicles, move on the basis of their own intentions, leading to a constantly changing environment. Static elements such as traffic lights and road lanes enforce traffic regulations. The driving brain, which is the cognitive space within AVs, is used to process information from the physical environment to generate appropriate driving policies. This brain analyses the behavioural intentions of pedestrians and other vehicles and determines the level of attention each requires. For instance, while the AV will primarily focus on a pedestrian wearing a red top who is crossing the road, it will continue to observe a pedestrian wearing a green top standing on the side of the road. The blue vehicle turning left is marked for attention, while vehicles whose paths do not intersect with the AVs are deemed safe. For objects that require high levels of attention, the driving brain formulates corresponding driving policies. For example, virtual stop fences indicate that the AV will yield to these objects. Traffic rules are translated into executable mathematical formulas that guide the behaviours of AVs. On the basis of external costs and internal motivations, the driving brain makes the final decisions and executes driving actions to ensure socially compatible behaviours.

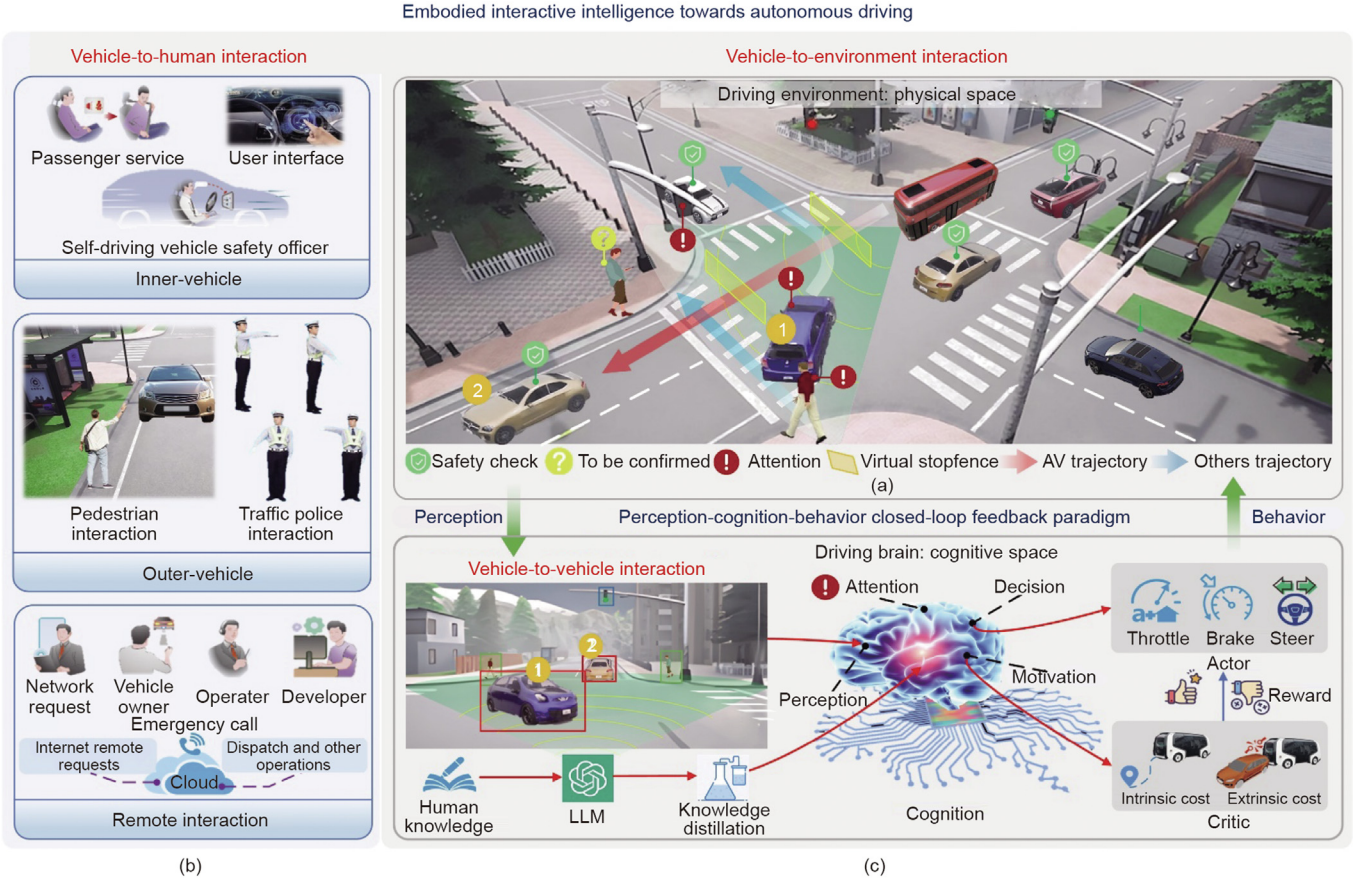


Fig. 1. EIIAD. (a) Vehicle-to-environment interactions and the proposed perception–cognition–behaviour closed-loop feedback paradigm. (b) Vehicle-to-human interactions. (c) Vehicle-to-vehicle interactions. LLM: large language model.

Notably, the interactions between AVs and pedestrians or other vehicles are distinct. Given the unpredictability and unique communication methods of pedestrians, AVs need to be able to recognize human actions and gestures. However, interactions between vehicle are focused more on deciphering driving patterns and behaviours. Thus, we devise separate models to address these different types of interactions within our AD algorithm.

Vehicle-to-human interactions are further divided into inner-vehicle, outer-vehicle, and remote cloud interactions (Fig. 1(b)). Inner-vehicle interactions involve real-time engagement with passengers, whereas remote cloud interactions facilitate online services and remote control. Outer-vehicle interactions focus on understanding and predicting the behaviours of pedestrians and instructions from traffic officers. Owing to complex traffic conditions, such as adverse weather conditions and varied human behaviours, AVs face challenges in recognizing human body language signals promptly and accurately. To address these issues, we propose a hypergraph neural network based on multiview spatiotemporal features (HGNN-MSTF). This model involves the construction of high-order semantic associations of human joints, enhancing the accuracy and robustness of action recognition and behaviour understanding.

Vehicle-to-vehicle interactions present significant challenges for EIIAD, which involves effective communication with other vehicles using nonverbal "vehicle language," such as headlight signals and speed (Fig. 1(c)). On the one hand, AVs need to predict other vehicles' intentions, such as merging or lane changing, to maintain safety. On the other hand, AVs must actively interact with the environment, initiating manoeuvres to enhance traffic efficiency. This process requires advanced cognitive processing, where AVs interpret behaviours, anticipate future trajectories, and devise optimal driving strategies to guarantee traffic safety and fluidity.

Mainstream AD approaches include imitation learning (IL) and RL. IL is simple and efficient, but it relies heavily on expert data, risking reliability in unfamiliar situations. RL, through trial and error, involves learning essential driving characteristics, offering better generalizability and robustness than IL; however, it requires extensive data collection. To address this issue, Hafner et al. [21] proposed Dreamer, a RL framework based on a world model that needs relatively few actual samples [21–23]. Inspired by Hafner, we propose intervehicle interactions based on a joint trajectory prediction world model (JTPWM) within the deep RL framework (JTPWM-DRL). This network builds a world model of AD by jointly predicting the driving trajectories of AVs and surrounding vehicles.

To manage complex human–vehicle–road interactions, we integrate the interaction cognition model of AVs and humans (HGNN-MSTF) and environmental constraints into JTPWM-DRL, achieving an end-to-end unified constrained vehicle environment interaction (UniCVE) model. Our model features the cognition and decision-making capabilities of large language model (LLM) assisted AD algorithms such as LMDrive [24–26], which not only preserves real-time computational performance but also enhances the cognitive decision-making capabilities of the system.

2. Methods and materials

2.1. Problem formulation

EIIAD enables AVs to become interactive, learning, and self-growing mobile intelligent agents. It is necessary to construct perception–cognition–behaviour closed-loop feedback, which can be systematically formulated with the partially observable Markov decision process (POMDP). We first define a state space $\mathscr{S}$ to encapsulate all possible states of traffic elements in the AD environment, such as the position and speed of vehicles and pedestrians and the road structure. The perception module uses various sensors (such as cameras and light laser detection and ranging (LiDAR)) to gather environmental information and recognize key interactive elements. However, accurate estimation of these states presents significant perception challenges because of sensor limitations, environmental uncertainties, and the complexity of interpreting human behavioural intentions from visual cues such as gestures and body language, under adverse conditions, such as occlusion and varying lighting. Owing to sensor noise and occlusions, the AV cannot obtain all real environmental states. Thus, we express the perception result as a conditional probability distribution based on the real state: $o_t \sim \mathscr{p}(o_t|s_t)$, where $\mathscr{p}$ represents the probability distribution function, $s \in \mathscr{S}$ denotes the state, o is the perception result, and the subscript t represents the moment (i.e., s_t and o_t express the state and perception results at time t, respectively).

The cognition module in this work forms an interaction model between the vehicle and traffic elements, focusing on understanding pedestrian behaviour, assessing surrounding vehicle intentions, predicting vehicle and pedestrian paths, and constraining driving behaviour on the basis of road and traffic signals. However, this process presents significant cognition challenges as the AV must predict future states of each traffic participant; hence, it faces uncertainty, requiring accurate modeling of individual behavioural patterns and their complex interdependent interactions within the dynamic traffic environment. Given the uncertainty in traffic participant behaviours due to environmental factors, we formulate the cognition model using conditional probability distributions of state prediction: $s_{t+1}^{k_e} \sim \mathscr{p}\left(s_{t+1}^{k_e}|s_t, o_t\right), k_e = 0, 1, 2, ..., K_e - 1$, where the superscript denotes the index of the K_e variable elements (such as pedestrians and vehicles) in the environment, with 0 specifically denoting the AV, and $s_{t+1}^{k_e}$ represents the state of the k_eth object at time $t+1$. This model predicts the intentions and future states of other elements on the basis of the current entire interaction environment state s_t, representing the vehicle's understanding of the entire dynamic driving environment over time.

On the basis of the interaction cognition model, the behaviour module learns a driving policy π, which represents the action probability distribution given a state and observation at time t, that is, $\pi(a_t|s_t, o_t)$:$a_t \sim \mathscr{p}_\pi(a_t|s_t, o_t)$, where a_t denotes the action to be executed at time t, $\mathscr{p}_\pi$ denotes stochastic decision-making distribution. Actions of the AV include throttle, braking, and steering. After the AV takes an action, the environmental state changes according to the transition model $\mathscr{T}$:$s_{t+1} \sim \mathscr{p}(s_{t+1}|s_t, o_t, a_t)$, which represents the probability of the state at time $t+1$ based on the state, observation, and action at time t. $\mathscr{T}$ is the underlying state-transition dynamics of the environment. The information from the dynamic environment is continuously collected by the perception module, forming a closed feedback loop. The objective is to maximize the expected return over a future time horizon H to find the optimal driving policy $\pi^*(a_t|s_t, o_t)$. The expected return is defined as $G_t = \sum_{k_t=0}^{H} \gamma^{k_t} r_{t+k_t}$, where k_t is the incremental time step, r_{t+k_t} is the reward at time $t+k_t$, the reward function $r_t \sim \mathscr{p}(r_t|s_t)$ is a probability distribution conditional on the state s_t, and $\gamma \in [0, 1]$ is the discount factor. We utilize the action value function $q_\pi(s_t, o_t, a_t)$ to estimate the expectation of G_t for the executed policy π: $q_\pi(s_t, o_t, a_t) = \mathbb{E}_\pi(G_t|s_t, o_t, a_t)$. $\mathbb{E}_\pi$ is the expectation taken with respect to the trajectory distribution induced by policy π. Leveraging the recursive Bellman equation, we ultimately obtain the optimal policy:

$$a_t^* = \pi^*(a_t|s_t, o_t) = \underset{a}{\operatorname{argmax}} \left(r_t + \gamma Q_{t+1}\right) \tag{1}$$

$$Q_{t+1} = \int \mathscr{p}(s_{t+1}|s_t, o_t, a_t) q_{t+1}^* \mathrm{d}s_{t+1} \tag{2}$$

$$q_{t+1}^{*}=\max_{a}(q_{\pi}(s_{t+1},o_{t+1},a_{t+1})) \tag{3}$$

where a_t^* represents the optimal action, q_{t+1}^* represents the optimal action value at the next time step, q_π represents the action-value function under policy π, and Q_{t+1} represents the expected state value at time $t+1$. Generating optimal actions that satisfy multiple competing objectives presents behaviour optimization challenges, where the AV must balance safety requirements, traffic rule compliance, efficiency goals, and social compatibility expectations while adapting to dynamic environmental constraints and human behavioural patterns.

To systematically address these interconnected challenges within our POMDP framework, we propose a UniCVE architecture comprising three specialized modules. The perception module employs the HGNN-MSTF to enhance human behaviour recognition through multiview analysis, the cognition module uses the JTPWM-DRL to predict future states of traffic participants through joint trajectory modelling, and the behaviour module implements unified multiobjective optimization to generate socially compatible driving policies. This tripartite architecture ensures synergistic integration where perception uncertainty is reduced, traffic participant behaviours are accurately predicted, and optimal actions are generated within the closed-loop feedback paradigm. Each module directly addresses its corresponding challenge while contributing to the overall embodied interactive intelligence framework.

2.2. Body language interaction model based on multiview spatiotemporal hypergraphs

Embodied interactive intelligence relies on robust perceptual abilities. However, real-world scenarios present challenges for accurate recognition, such as target occlusion and insufficient lighting. These perception challenges require accurate human behaviour recognition, which we formulate as an estimation of the conditional probability distribution $\mathcal{P}(b|s)$, where b represents the behavioural intention. In this work, we propose a multiview video data acquisition method, which has better potential consistency and complementarity than static images/single-view video data, resolving the problem of the lack of adequate action recognition data in complex scenes. We propose a HGNN-MSTF, which uses a hypergraph convolution module based on dynamic spatiotemporal attention mechanisms to capture the salient regions of the target object and its surrounding environment. This model extracts the features and models the correlations among the spatial structures and temporal data through deep convolutional networks (Fig. 2(a)).

Specifically, we propose two strategies for constructing a spatial hypergraph, one based on limb components and the other based on adjacent joints. The spatial hypergraph, which is based on a limb construction strategy, is denoted as follows:

$$\mathcal{G}^{L-\mathrm{spa}}=\left(\mathcal{V}^{L-\mathrm{spa}},\mathcal{E}^{L-\mathrm{spa}},\boldsymbol{W}^{L-\mathrm{spa}}\right) \tag{4}$$

where $\mathcal{V}^{L-\mathrm{spa}}=\left\{\mathcal{V}_1^{L-\mathrm{spa}},\mathcal{V}_2^{L-\mathrm{spa}},\ldots,\mathcal{V}_m^{L-\mathrm{spa}}\right\}$, $\mathcal{E}^{L-\mathrm{spa}}=\left\{e_1^{L-\mathrm{spa}},e_2^{L-\mathrm{spa}},\ldots,e_m^{L-\mathrm{spa}}\right\}$, and $\boldsymbol{W}^{L-\mathrm{spa}}$ represent the vertex sets, hyperedge sets, and weight matrix of the spatial hypergraph, respectively. Five parts comprise the hyperedges of the limb spatial hypergraph: the trunk, left hand, right hand, left leg, and right leg. The mth vertex set $\mathcal{V}_m^{L-\mathrm{spa}}$ in each hyperedge $e_m^{L-\mathrm{spa}}$ can be expressed as follows:

$$\mathcal{V}_m^{L-\mathrm{spa}}=\left\{v_{p,n}^{(i)}|\forall v_{p,n}^{(i)}\in\mathcal{V}^{L-\mathrm{spa}}\right\} \tag{5}$$

where $p=1,2,\ldots,P$; $n=1,2,\ldots,N$; $i=1,2,\ldots,I$; $m=1,2,\ldots,M$. P, I, and M denote the numbers of views, nodes, and hyperedges, respectively. $v_{p,n}^{(i)}$ is the ith node in the nth frame of the pth view.

Given that each video sequence contains N frames, we construct N spatial multiple hypergraphs. For each nth spatial hypergraph, the following equation can be considered:

$$\mathcal{G}_n^{L-\mathrm{spa}}=\left(\mathcal{V}_n^{L-\mathrm{spa}},\mathcal{E}_n^{L-\mathrm{spa}},\boldsymbol{W}_n^{L-\mathrm{spa}}\right) \tag{6}$$

where n ranges from 1 to N and is based on the limb construction strategy. The initial incidence matrix of each spatial hypergraph is defined as follows:

$$\boldsymbol{H}_n^{L-\mathrm{spa}}\left(v_{p,n}^{(i)},e_{m,n}^{L-\mathrm{spa}}\right)=\begin{cases}1 & v_{p,n}^{(i)}\in e_{m,n}^{L-\mathrm{spa}}\\ 0 & v_{p,n}^{(i)}\notin e_{m,n}^{L-\mathrm{spa}}\end{cases} \tag{7}$$

where $e_{m,n}^{L-\mathrm{spa}}$ represents the mth hyperedge in the nth spatial hypergraph. The diagonal matrices of the hyperedge degree and vertex degree are represented by $\boldsymbol{D}_{e_n}^{L-\mathrm{spa}}$ and $\boldsymbol{D}_{v_n}^{L-\mathrm{spa}}$, respectively. The Laplace matrix $\boldsymbol{G}_n^{L-\mathrm{spa}}$, which integrates high-order semantic information, is generated according to the incidence matrix $\boldsymbol{H}_n^{L-\mathrm{spa}}$, which is formulated as follows:

$$\boldsymbol{G}_n^{L-\mathrm{spa}}=\boldsymbol{D}_{v_n}^{L-\mathrm{spa}\,-1/2}\boldsymbol{H}_n^{L-\mathrm{spa}}\boldsymbol{W}_n^{L-\mathrm{spa}}\boldsymbol{D}_{e_n}^{L-\mathrm{spa}\,-1}\left(\boldsymbol{H}_n^{L-\mathrm{spa}}\right)^{\mathrm{T}}\boldsymbol{D}_{v_n}^{L-\mathrm{spa}\,-1/2} \tag{8}$$

We propose a different type of spatial hypergraph based on the adjacent joint construction strategy. Each nth spatial hypergraph can be denoted as follows:

$$\mathcal{G}_n^{J-\mathrm{spa}}=\left(\mathcal{V}_n^{J-\mathrm{spa}},\mathcal{E}_n^{J-\mathrm{spa}},\boldsymbol{W}_n^{J-\mathrm{spa}}\right) \tag{9}$$

where n ranges from 1 to N and is constructed on the basis of the adjacent joint construction strategy. The Laplace matrix $\boldsymbol{G}_n^{J-\mathrm{spa}}$ is generated according to the incidence matrix $\boldsymbol{H}_n^{J-\mathrm{spa}}$ and is formulated as follows:

$$\boldsymbol{G}_n^{J-\mathrm{spa}}=\boldsymbol{D}_{v_n}^{J-\mathrm{spa}\,-1/2}\boldsymbol{H}_n^{J-\mathrm{spa}}\boldsymbol{W}_n^{J-\mathrm{spa}}\boldsymbol{D}_{e_n}^{J-\mathrm{spa}\,-1}\left(\boldsymbol{H}_n^{J-\mathrm{spa}}\right)^{\mathrm{T}}\boldsymbol{D}_{v_n}^{J-\mathrm{spa}\,-1/2} \tag{10}$$

where $\boldsymbol{D}_{e_n}^{J-\mathrm{spa}}$ and $\boldsymbol{D}_{v_n}^{J-\mathrm{spa}}$ represent the diagonal matrices of hyperedge degree and vertex degree, respectively.

We construct a temporal hypergraph by grouping sequential frames from the same view. Given P views, we form P temporal hypergraphs. Each pth temporal hypergraph is denoted as follows:

$$\mathcal{G}_p^{\mathrm{tem}}=\left(\mathcal{V}_p^{\mathrm{tem}},\mathcal{E}_p^{\mathrm{tem}},\boldsymbol{W}_p^{\mathrm{tem}}\right) \tag{11}$$

where $\mathcal{V}_p^{\mathrm{tem}}$, $\mathcal{E}_p^{\mathrm{tem}}$, and $\boldsymbol{W}_p^{\mathrm{tem}}$ are temporal nodes sets, temporal hyperedges sets, and hyperedge weight matrix, respectively.

The initial incidence matrix of each temporal hypergraph is defined as follows:

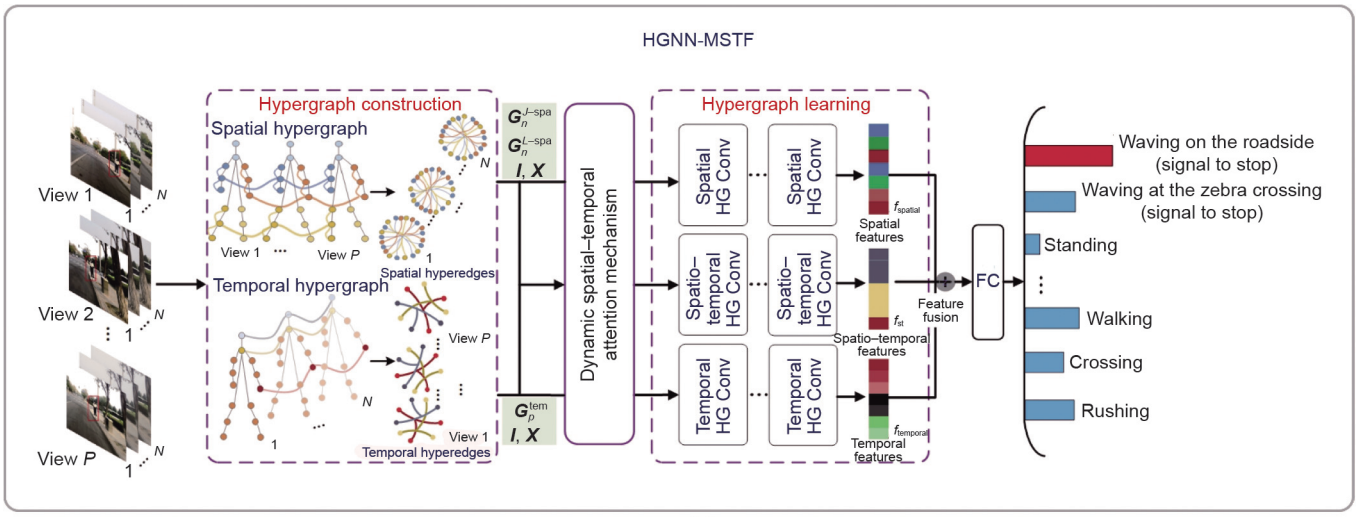
$$\boldsymbol{H}_p^{\mathrm{tem}}\left(v_{p,n}^{(i)},e_{m,p}^{\mathrm{tem}}\right)=\begin{cases}1 & v_{p,n}^{(i)}\in e_{m,p}^{\mathrm{tem}}\\ 0 & v_{p,n}^{(i)}\notin e_{m,p}^{\mathrm{tem}}\end{cases} \tag{12}$$

where $e_{m,p}^{\mathrm{tem}}$ represents the mth hyperedge in the pth temporal hypergraph. To capture the high-order semantic information between different sequences of human body joints, the Laplace matrix $\boldsymbol{G}_p^{\mathrm{tem}}$ is generated according to the incidence matrix $\boldsymbol{H}_p^{\mathrm{tem}}$ is formulated as follows:

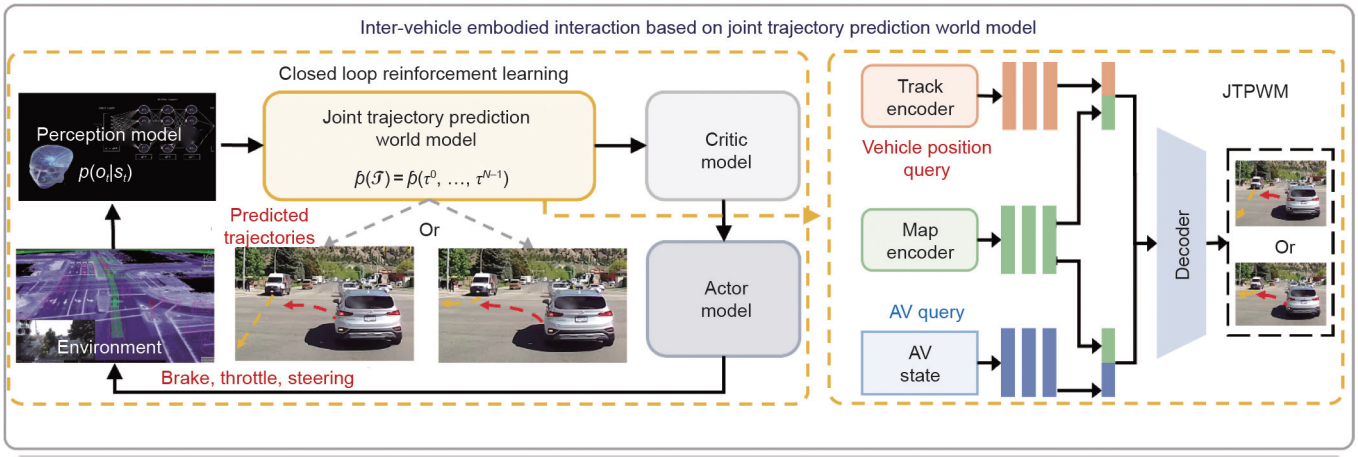
$$\boldsymbol{G}_p^{\mathrm{tem}}=\boldsymbol{D}_{v_p}^{-1/2}\boldsymbol{H}_p^{\mathrm{tem}}\boldsymbol{W}_p^{\mathrm{tem}}\boldsymbol{D}_{e_p}^{-1}\left(\boldsymbol{H}_p^{\mathrm{tem}}\right)^{\mathrm{T}}\boldsymbol{D}_{v_p}^{-1/2} \tag{13}$$

where $\boldsymbol{D}_{v_p}$ and $\boldsymbol{D}_{e_p}$ are the diagonal matrices of vertex and hyperedge degree, respectively.

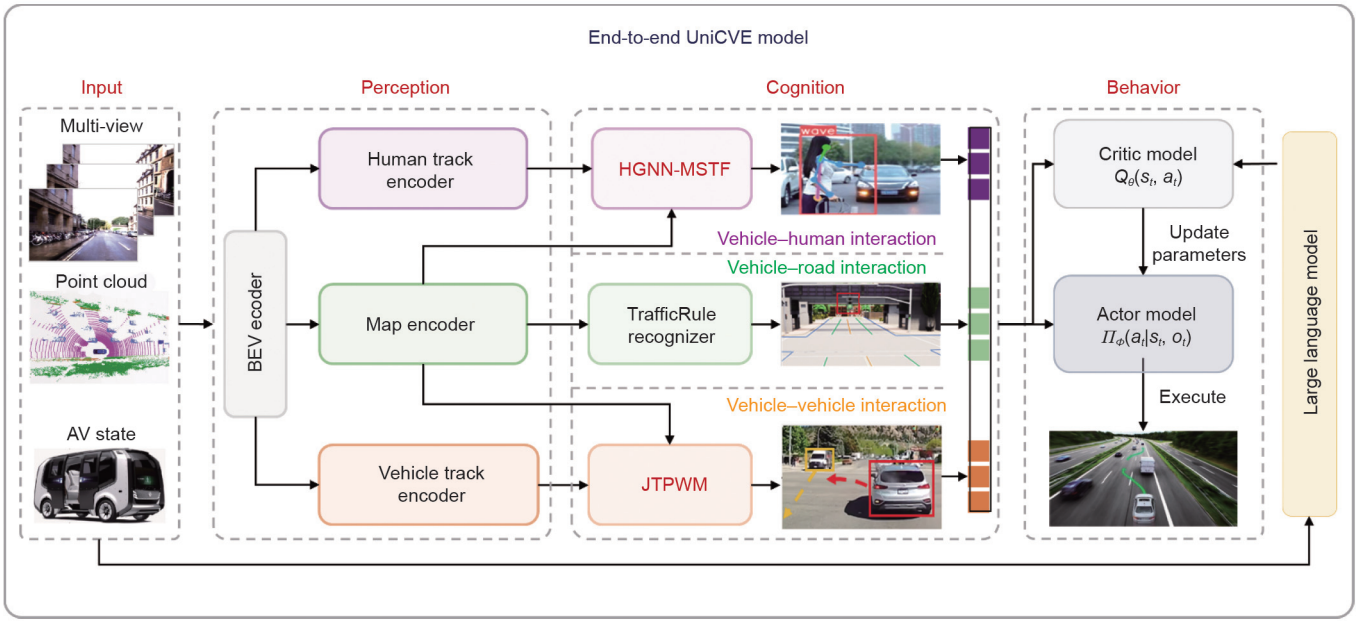
Expanding on the foundation of spatial and temporal hypergraphs, we incorporate a dynamic spatial attention mechanism to extract features from dynamically significant regions of interest. We then introduce spatial, temporal, and spatiotemporal hypergraph neural networks. These networks are designed to learn high-order semantic associations from temporal and spatial



(a)



(b)



(c)

Fig. 2. Models contained in EIIAD and their interconnected relationships. (a) Body language interaction model based on a spatiotemporal hypergraph (HGNN-MSTF) that processes multimodal sensory inputs to extract human behavioural features and interaction intentions. FC: fully connected. (b) Intervehicle embodied interaction based on a JTPWM-DRL that leverages environmental dynamics to predict future vehicle trajectories and assess potential collision risks. (c) End-to-end UniCVE model that serves as an integration framework that can systematically combine outputs from the HGNN-MSTF and JTPWM-DRL modules. Specifically, the UniCVE model incorporates human behaviour understanding from the perception module and trajectory predictions from the cognition module into a unified reward function structure within the RL actor–critic framework, enabling the generation of socially compliant and safety-aware driving policies through continuous multimodal information fusion and constraint optimization.

human actions, effectively addressing the issue of inaccurate recognition due to indeterminate information. The feature matrix of the vertices is denoted $\boldsymbol{X}$; we employ Laplacian matrices $\boldsymbol{G}_n^{L-\mathrm{spa}}$, $\boldsymbol{G}_n^{J-\mathrm{spa}}$, $\boldsymbol{G}_p^{\mathrm{tem}}$, and self-connection features $\boldsymbol{I}$ to capture high-order semantic information [40].

Upon training the aforementioned hypergraph neural networks, we derive the spatial hypergraph features f_{spatial}, temporal hypergraph features f_{temporal}, and spatiotemporal hypergraph features f_{st}. The features are then concatenated to yield the final features

$$f \in \mathbb{R}^{((P\times\boldsymbol{I})+(N\times\boldsymbol{I})+(N+P)\times\boldsymbol{I})\times C'} \tag{14}$$

where C' is the updated dimension.

We use the softmax function to compute the probability value of behaviour prediction, with the action category with the maximum probability value being the final prediction result. To optimize the proposed framework, we use cross-entropy loss for supervision during training. The cross-entropy loss is computed by adding a trainable linear layer to the final representation f. Mathematically, this parameter is expressed as follows:

$$L = -\sum_{r=1}^{Q} y_r \log \frac{\mathrm{e}^{f_r}}{\sum_{u=1}^{Q}\mathrm{e}^{f_u}} \tag{15}$$

where L is the loss function, y_r is the true feature label, and Q is the number of classes.

2.3. Intervehicle embodied interaction based on the JTPWM

AD is a typical long-horizon task. For instance, an autonomous bus usually spends several minutes to half an hour executing a sequence of actions to travel between two stations. The driving behaviour of the AV at each moment can influence its own future state and that of surrounding vehicles. Conversely, potential future interaction scenarios among vehicles can constrain the current driving decisions of AVs. Hence, AVs must rapidly and thoroughly evaluate the potential risks and benefits of different strategies over an extended future period to make informed decisions. The length of the prediction time window affects the assessment of risks and benefits. For example, in terms of collision risk, a shorter prediction window might cause a delay in the AV's emergency braking response. To capture these complex temporal dependencies, it is necessary to accurately model the system transition dynamics (i.e., $s_{t+1}\sim\mathscr{P}(s_{t+1}|s_t,o_t,a_t)$ defined in Section 2.1).

To equip AVs with far-sighted prediction capabilities in realistic complex interaction scenarios, we construct a world model based on joint trajectory prediction to characterize the cooperative interaction among AVs and surrounding vehicles. On this basis, we design an intervehicle embodied interaction network based on the JTPWM-DRL. Our novel actor–critic method leverages the predicted trajectory generated by the world model to update the critic network $Q_\theta(s_t,a_t)$ to approximate the action value function $q_\pi(s_t,o_t,a_t)$ and actor network $\Pi_\phi(a_t|s_t,o_t)$ to estimate the driving policy $\pi(a_t|s_t,o_t)$; θ and ϕ are learnable parameters (Fig. 2(b)).

The world model involves the construction of a compressed internal representation of the environment and the prediction of its evolution [8]. Our world model accurately represents vehicle interactions by predicting the joint trajectory probability distribution among multiple vehicles, thereby enhancing the embodied interactive intelligence of AVs. We argue that vehicle interactions are manifested mainly in interdependent driving trajectories. Unlike world models that generate images or videos, our model captures finer interaction features by concentrating on vehicle trajectory prediction and actively omitting elements that are irrelevant to interactions, such as distant buildings. Assuming the current time $t=0$, a predicted trajectory can be expressed as a series of states within a future period H: $\tau=\{s_1,s_2,...,s_H\}$. In a driving environment with R vehicles, a potential vehicle interaction scenario $\mathscr{T}$ is defined as the set of predicted trajectories of the vehicles: $\mathscr{T}=\{\tau^0,\tau^1,...,\tau^{R-1}\}$. Owing to interaction uncertainties, there may be multiple future interaction scenarios, each with a probability of $\mathscr{P}(\mathscr{T})=\mathscr{P}(\tau^0,\tau^1,...,\tau^{R-1})$. To improve the learning efficiency, JTPWM-DRL estimates the long-term return with the predicted scenarios instead of learning the results by trial and error. The estimated return is computed as follows:

$$Q_{\mathrm{w}}(s_t,a_t)=\sum_{i=1}^{\mathscr{M}}\mathscr{P}(\mathscr{T}^i)G(\mathscr{T}^i) \tag{16}$$

where $\mathscr{M}$ denotes the number of potential scenarios and the subscript w represents that the value is estimated with the world model. $G(\cdot)$ denotes the cumulative discounted return obtained along a sampled trajectory. We update the θ of $Q_\theta(s_t,a_t)$ by minimizing the critic loss function $\mathscr{L}_c$, which represents the difference between the estimated action state value and the expected return of possible future scenarios, as follows:

$$\mathscr{L}_c=\min_\theta \mathbb{E}\left[\left(Q_\theta(s_t,a_t)-Q_{\mathrm{w}}(s_t,a_t)\right)^2\right] \tag{17}$$

where $\mathbb{E}(\cdot)$ stands for expected value.

We then update ϕ of Π_ϕ by maximizing the soft action value loss function $\mathscr{L}_A$ as follows:

$$\mathscr{L}_A=\max_\phi \mathbb{E}\left[Q_\theta(s_t,a_t)-\log\left(\Pi_\phi(a_t|s_t,o_t)\right)\right] \tag{18}$$

Specifically, we design a JTPWM through a transformer network based on the cross-attention mechanism (Fig. 2(b)). The perception model employs the TrackEncoder and MapEncoder to encode the states of the surrounding vehicles and map environment, respectively. To establish the interaction relationship between the AV and social vehicles, we use the encoded states of the AV and social vehicles as queries, namely, the AV Query and Vehicle Position Query. The encoded map environment is used as both the key and the value. Given a dataset $\mathscr{D}$, we train the world model by minimizing the negative logarithm of the joint trajectory likelihood: $\min_\kappa \sum_{\mathscr{T},\mathscr{D}} -\log\mathscr{P}\{\mathscr{T}|\mathscr{D},\kappa\}$, where κ is the parameter of the world model.

2.4. End-to-end UniCVE model

The realistic interactive environment is composed of AVs, pedestrians, surrounding vehicles, and road structures. To handle such complex situations, we develop the end-to-end UniCVE model by formulating interaction intentions and rules as unified rewards and soft constraints in the RL actor–critic framework. Specifically, by expanding upon JTPWM-DRL, we build a cross-modal structured hypergraph neural network based on HGNN-MSTF and recognize traffic rules using map segmentation features. We further enhance the cognitive and decision-making capabilities of the model by inputting human knowledge from a LLM into our real-time AD interactive intelligence model through tailored reward functions. By incorporating multiple perception and cognition results through a unified reward function framework, we solve for the optimal policy $\pi^*(a_t|s_t,o_t)$ that generates driving behaviours that simultaneously satisfy safety requirements, traffic regulations, and social compatibility expectations. This approach ensures that our model delivers instant responsiveness while remaining adept at understanding the complex interactions within the world (Fig. 2(c)).

We unify multiview and cross-modal perception information in the Bird's eye view (BEV) space, with the vehicle centre serving as the origin of the coordinates. This space provides a comprehensive

representation of the surroundings, including static elements such as lane structures and dynamic elements such as other vehicles and pedestrians. We utilize TrackEncoders, including the VehicleTrackEncoder and HumanTrackEncoder, for object detection by extracting features such as the position and speed of traffic participants. Concurrently, MapEncoder extracts map features to segment driving environment areas, including lane lines and zebra crossings. Note that in our network, the TrackEncoder and MapEncoder can be implemented with any object detection and map segmentation model; therefore, we utilize the outputs of the corresponding modules in the Learning from All Vehicles (LAV) [38] as the encoding features.

We jointly train the BEV features through an object detection task and a map segmentation task. Specifically, we utilize point painting to fuse cross-modal data of multiview images and point clouds and extract features through PointPillars [39]. This results in the basic cross-modal BEV features $F \in \mathbb{R}^{C\times H\times W}$ that describe the environmental information around the AV, where C denotes the dimension of the BEV features, and H and W represent the height and width of the BEV grids, respectively. We reshape F to $F' = \{F_1, ..., F_{\mathrm{HW}}\} \in \mathbb{R}^{C\times H\times W}$ for easier indexing. On the basis of F', we dynamically select Y important grid regions $g_1, ..., g_Y$, potentially occupied by important objects such as pedestrians, buildings, and vehicles. We design the important region proposal function $G(\cdot)$ by selecting the elements of F with the top-N L2-Norm values:

$$g_1, ..., g_Y = G(F') = \mathrm{Top}k(\|F_1\|_2, ..., \|F_{\mathrm{HW}}\|_2)$$

where Top$k()$ denotes the operator that selects the indices of the k largest L2-Norm values from its input.

To capture the contextual information between different regions, we further construct a hypergraph $\mathscr{G}_{\mathrm{IR}} = \{\mathscr{V}_{\mathrm{IR}}, \mathscr{E}_{\mathrm{IR}}\}$, where $\mathscr{V}_{\mathrm{IR}} = \{v_\alpha | \alpha = 1, ..., Y\}$ and v_α represents the αth region g_α. The vertex features $F_{\mathrm{IR}} = \left\{F_{g_1}, ..., F_{g_\gamma}\right\}$ are BEV features in the corresponding regions. To construct the hyperedge set $\mathscr{E}_{\mathrm{IR}}$, we connect vertices with closed semantic relations by using K-nearest neighbor (KNN) on the feature space:

$$\mathscr{E}_{\mathrm{IR}} = \left\{v_\alpha, \forall v_\beta \in \mathrm{KNN}(v_\alpha) | v_\alpha, v_\beta \in \mathscr{V}_{\mathrm{IR}}\right\} \tag{19}$$

The association matrix $\boldsymbol{H}_{\mathrm{IR}}$ of the hypergraph is as follows:

$$\boldsymbol{H}_{\mathrm{IR}}(\forall v_\alpha, \forall e_\beta) = \begin{cases} 1 & v_\alpha \in e_\beta \\ 0 & v_\alpha \notin e_\beta \end{cases}, e_\beta \in \mathscr{E}_{\mathrm{IR}} \tag{20}$$

where e_β is the βth hyperedge. On the basis of $\mathscr{G}_{\mathrm{IR}}$, the AV learns the high-order relationships between different regions and extracts semantic features with multiple convolutional layers as follows:

$$F_{\mathrm{IR}}^{(l+1)} = F_{\mathrm{IR}}^{(l)} + \sigma(A) \tag{21}$$

$$A = D_{\mathrm{IR},v}^{-1/2,(l)} H_{\mathrm{IR}}^{(l)} W^{(l)} D_{\mathrm{IR},e}^{-1,(l)} H_{\mathrm{IR}}^{T,(l)} D_{\mathrm{IR},v}^{-1/2,(l)} F_{\mathrm{IR}}^{(l)} \theta^{(l)} \tag{22}$$

where $F_{\mathrm{IR}}^{(l)}$ is the vertex feature of the lth layer of the hypergraph, $\sigma(\cdot)$ is the activation function, and $D_{\mathrm{IR},v}^{(l)}$ and $D_{\mathrm{IR},e}^{(l)}$ are the diagonal weights of the vertex degree and hyperedge degree in the lth layer of the hypergraph, respectively. $H_{\mathrm{IR}}^{(l)}$ is the association matrix of the lth layer of the hypergraph, and $W^{(l)}$ and $\theta^{(l)}$ are the learnable hyperedge weight and vertex weight in the lth layer of the hypergraph, respectively. Through this method, we continuously select significant regions and propagate the hypergraph to obtain context-based BEV features.

With context-based BEV features, we construct interaction cognition models of AVs with pedestrians and surrounding vehicles and obtain a semantic understanding of road traffic rules. The HGNN-MSTF recognizes the intentions of pedestrians, such as crossing the road or making a roadside request to board the car, by considering both their positions and body language. For the interaction among AVs and surrounding vehicles, we utilize the proposed JTPWM to predict trajectories. Our implementation is based on the LAV's trajectory prediction module [38]. In addition, the traffic rule recognition operator uses traffic semantic features generated by the map encoder to recognize traffic rules, such as stopping at red lights and not occupying bus lanes from 9 am to 5 pm.

Similar to the approach in JTPWM-DRL, where the predicted trajectories of vehicles are utilized to estimate the action state values of the AV, we regard the pedestrian intention and traffic rules as constraints by designing specific reward functions. For example, if there is a red light l metres ahead of an AV, it is penalized if it moves forwards more than l metres. Assuming that the position of the AV at this time is ($x = 0$ m, $y = 0$ m), the reward function specific for traffic rules $r_t(s)$ can be expressed as follows:

$$r_t(s) = -1000, \text{if } \sqrt[2]{{x_s}^2 + {y_s}^2} \geq l \tag{23}$$

where x_s and y_s represent the coordinates of the AV. In this manner, the AV can calculate the optimal strategy through understanding the comprehensive interaction relationship with the overall environment.

In addition, we input the driving strategy of the LLM-based AD model LMDrive [24] into our AD model, addressing the challenge of deploying LLM on real-time platforms. LMDrive receives multiview and multimodal perception signal inputs and yields a series of control instruction outputs under the guidance of text navigation or text reminders. These control instructions ultimately form a planning trajectory $\widehat{\tau}$. We use this trajectory as a reference trajectory and penalizing trajectories that are largely different from it. This difference is formulated with a reward function specific to the LLM $r_l(s, \tau) = -w\|\tau - \widehat{\tau}\|^2$, where w represents the weight parameter. In general, the reward function of the critic network consists of four parts: $r = r_\mathrm{b} + r_\mathrm{h} + r_\mathrm{t} + r_\mathrm{l}$, where r_b, r_h, r_t, and r_l are related to the basic reward items of the driving tasks, pedestrian intentions, traffic rules, and the results of the LLM, respectively. This design enables our AD system to better understand and adapt to complex driving environments.

In summary, the end-to-end perception–cognition–behaviour network realizes the core idea of embodied interactive intelligence, that is, the unity of knowledge and action, which perceives and influences the surrounding physical or social environment through embodied behaviour interactions. Embodied interactive intelligence will eventually be ubiquitous and omnipotent, similar to the self-driving vehicles and humanoid robots we see everywhere today.

3. Results and discussion

3.1. Experimental setup

To fully evaluate EIIAD, we designed an experiment in which an autonomous bus had to navigate a long distance to safely transport passengers. During operation, the bus boarded hailing passengers and bypassed stops if not needed. En route to the next stop, the bus encountered a series of interaction scenarios, such as changing lanes, unprotected left turns, narrow road encounters, and pedestrians signalling for precedence at crosswalks during traffic light changes. As shown in Fig. 3, we conducted experiments on a Dongfeng autonomous minibus equipped with 4 NVIDIA Orin-X chips (167 TOPS each, NVIDIA, USA), 6 LiDAR sensors (Hesai, China), 7 cameras, and 2 millimetre-wave radars, achieving an end-to-end inference latency of approximately 60–100 ms depending on the driving scenario. Our AD system achieves approximately

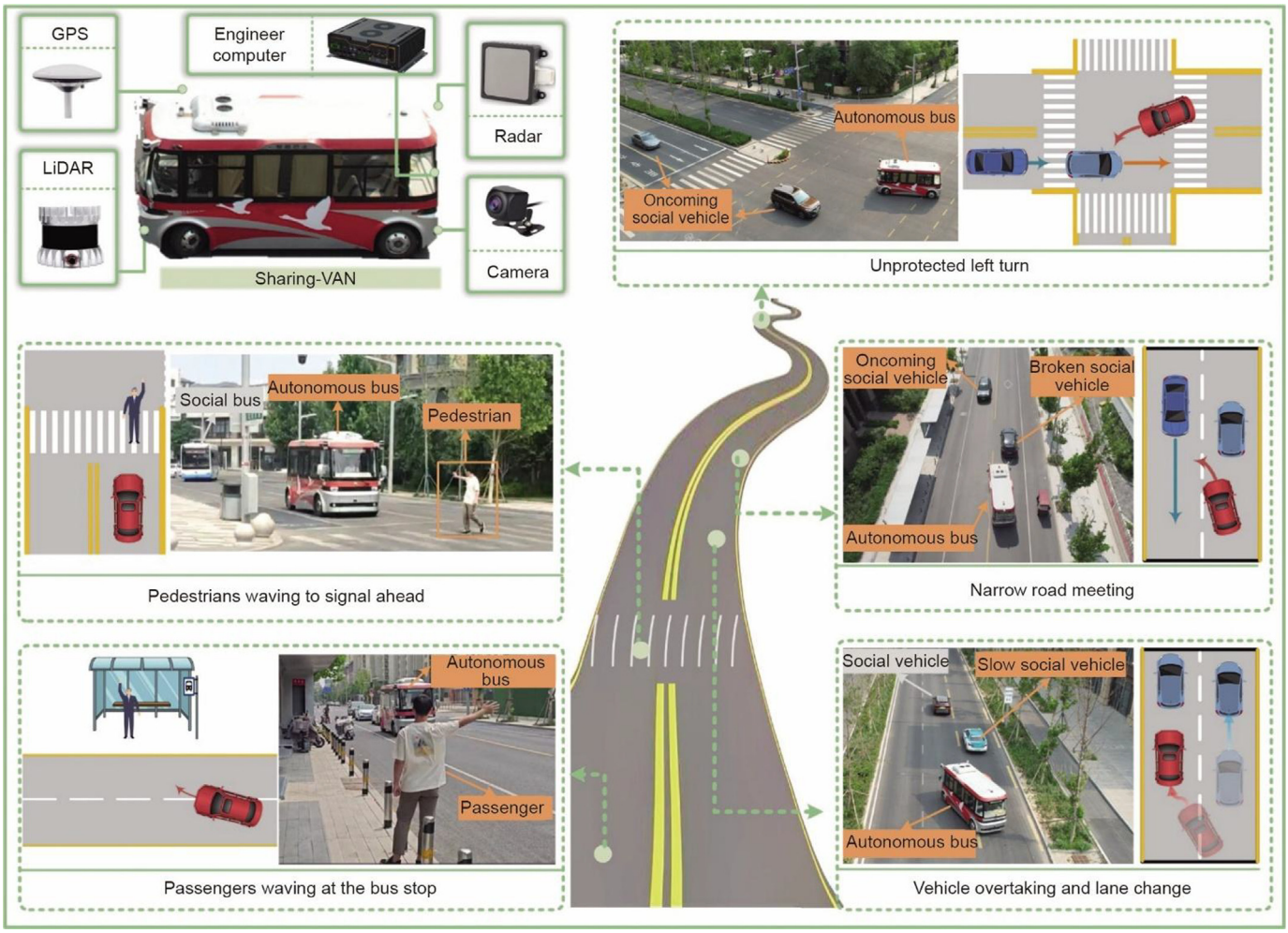


Fig. 3. Dongfeng AD bus navigation task. GPS: Global Positioning System.

26 000 km per disengagement, demonstrating competitive performance that positions us favourably among industry leaders according to the 2024 California DMV disengagement reports.

For comparison with the SOTA models, we conducted comparative experiments on the CARLA simulation platform (Computer Vision Center, Spain). The evaluation criteria included two task-level metrics: the success rate of picking up passengers and the rate of safe arrivals. Additionally, a scene-level metric was included: the collision rate between vehicles and pedestrians. Finally, we incorporated scores from a user study that corresponded to the perceived level of intelligence.

The following sections present our experimental results in three parts: ① vehicle-to-human interaction intelligence shown through recognizing and understanding human actions in various contexts; ② vehicle-to-vehicle interaction intelligence exhibited in three complex interaction scenarios, for example, narrow road encounters, overtaking and lane changes, and unprotected left turns against oncoming traffic; and ③ vehicle-to-environment interaction intelligence demonstrated by the overall performance of the autonomous bus in the navigation task.

3.2. Embodied interactive intelligence: vehicle-to-human

AVs often encounter complex road conditions in open scenarios. For effective human vehicle interactions, it is imperative to promptly and accurately discern the intentions of pedestrians on the basis of diverse contexts. This recognition and differentiation of pedestrian behaviour is critical for ensuring the safe and efficient operation of a vehicle. The proposed action recognition algorithm—a HGNN-MSTF—captures the spatial and temporal features of key joints on the human skeleton and combines contextual semantic information; this algorithm is an extension of our previous work [27]. We first validate our method's performance on a self-collected cross-modal human behaviour dataset for AD (CMHBD-AD) [28], verifying our algorithm's ability to understand pedestrian intentions in various AD scenarios; to better contextualize the uniqueness and relevance of our proposed CMHBD-AD dataset, we created a comparative summary with the widely used NTU-RGB+D dataset, as shown in Table 1 and Fig. 4. Afterwards, we demonstrated the wide applicability of our algorithm on the public NTU-RGB+D dataset to record daily human actions [29]. Overall, our HGNN-MSTF method outperforms current methods on the NTU-RGB+D dataset and achieves the best results on the CMHBD-AD dataset. Experiments were conducted on only one 4090Ti GPU (NVIDIA, USA) using the stochastic gradient descent (SGD) optimization algorithm with a momentum of 0.9, a weight decay of 0.0004, a training iteration of 100, a learning rate of 0.05, and a sliding window that extracted 40 frames every 5 s for training.

3.2.1. Vehicle-to-human interaction cognition based on the context of scenarios

In real-world scenarios, pedestrians may exhibit the same actions (such as waving) in different environments (such as by

Table 1
Comparison between the CMHBD-AD and NTU-RGB+D datasets.

Aspect	CMHBD-AD	NTU-RGB+D
Application scenario	AD (real-world pedestrian behaviour understanding)	Indoor daily human action recognition
Modalities	Red–green–blue (RGB) video and LiDAR point clouds	RGB video, depth maps, skeleton data
Camera views	Three views (left, centre, and right)	Three fixed viewpoints (cameras 1, 2, and 3)
Number of behaviour classes	8 classes (AD-specific pedestrian behaviours)	60 classes (general daily human actions)
Subjects	Real pedestrians in outdoor driving environments	40 volunteers performing predefined actions indoors
Total dataset size	~60 GB (50 GB video + 10 GB point cloud)	56 880 video samples
Image samples	150 000+ raw images; 50 000+ annotated images	~56 000 labelled video clips
Point cloud samples	~4000 raw; ~2000 annotated	Not available
Scene diversity	Campuses, intersections, parking lots, day/night, and varying lighting	Indoor environments only
Notable strengths	Multiview, multimodal, robust under occlusion/low-light, and emergent behaviour modeling	Rich behaviour types; skeletal and depth information for action modeling

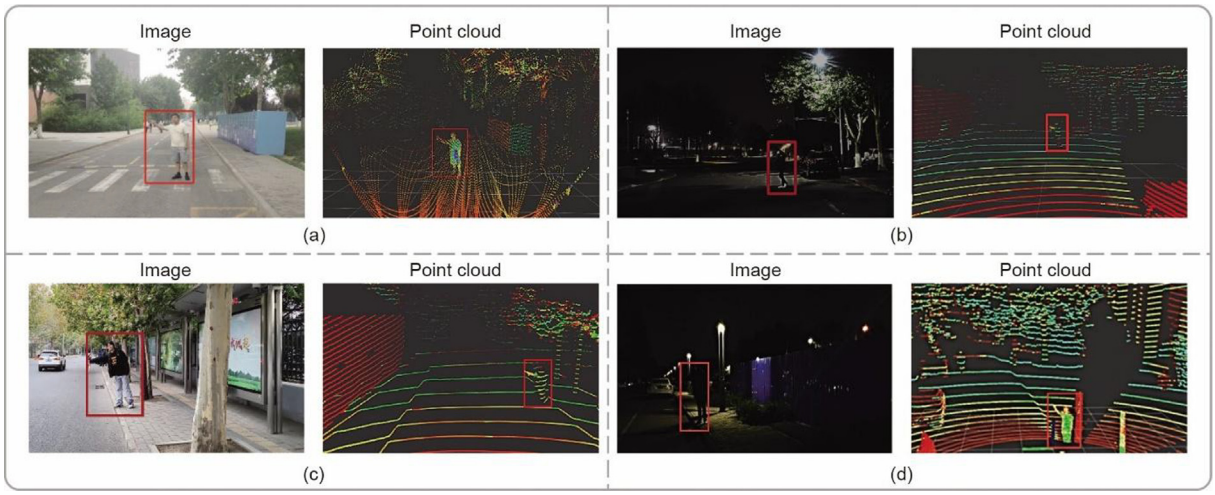


Fig. 4. Visualization of the cross-modal human behaviour dataset for AD and snapshots from simulation experiments of various interaction scenarios and corresponding statistical results. (a, b) pedestrian waves to vehicles from the road during the (a) day and (b) night. (c, d) pedestrian waves to vehicles by the roadside, again in both (c) daytime and (d) nighttime settings.

the roadside or on a zebra crossing), expressing different behavioural intentions (Fig. 5). We model the probability distributions of actions and human positions through the action recognition network and the position detection network, respectively. Combined with road areas from the map segmentation network, pedestrian intentions are expressed through joint probability distributions in specific areas. For example, in the process shown in Figs. 5(a-i) and (a-ii), when a pedestrian waves at the roadside, our system demonstrates embodied intelligence through the following cognition behaviour cycle. Specifically, the human behaviour cognition module takes the environmental state as the input and produces the probability distributions for the recognition results as the output; the formulation is $\mathbb{P}(\text{recognition}|s) = f_\theta(s)$, where θ represents the parameters of the network and s represents the environmental states. The probability of a pedestrian waving is calculated as $\mathbb{P}(b = \text{wave}) = 0.92$, where b represents the recognition result of pedestrian actions, while the probability of a pedestrian standing at the roadside is determined as $\mathbb{P}(\text{pos} \subseteq R_{\text{curb}}) = 0.89$, where pos represents the position of the pedestrian in the environment and R_{curb} represents the roadside area. Combining the results of action recognition and the environment, the AV analyses the joint probability of a pedestrian waving at the roadside as $\mathbb{P}(b = \text{wave}, \text{pos} \subseteq R_{\text{curb}}) = 0.82$, indicating that the pedestrian intends to take a ride. In response to this interpretation, our system generates an appropriate behavioural reaction with $\mathbb{P}(a_{\text{lon}} = -0.9\,\text{m}\cdot\text{s}^{-2}) = 0.87$ and $\mathbb{P}(a_{\text{lat}} = 15.2^\circ) = 0.78$ (a_{lon} is longitudinal acceleration, and a_{lat} is lateral acceletation, enabling the vehicle to simultaneously apply a moderate deceleration of 0.9 $\text{m}\cdot\text{s}^{-2}$ while executing a 15.2° steering angle to change lanes and reach the pickup location.

Conversely, when the same waving gesture occurs in a different context, such as at a crosswalk where $\mathbb{P}(b = \text{wave}) = 0.97$ and $\mathbb{P}(s \subseteq R_{\text{crosswalk}}) = 0.94$, the joint probability $\mathbb{P}(b = \text{wave}, s \subseteq R_{\text{crosswalk}}) = 0.91$ indicates a yielding request rather than a pickup intention. $R_{\text{crosswalk}}$ represents the zebra crossing area. Our embodied system responds accordingly to $\mathbb{P}(a_{\text{lon}} = -1.5\,\text{m}\cdot\text{s}^{-2}) = 0.95$, causing the AV to stop 2.0 m before the crosswalk and maintain position until the pedestrian crosses safely. This context-sensitive cognition and behaviour capability demonstrates the embodied nature of our system, where identical actions trigger different behavioural responses based on the environmental context.

Similarly, in the process shown in Figs. 5(a-iii) and (a-iv), when a pedestrian crosses the road on a zebra crossing, the traffic light controlling the vehicle turns from red to green, the pedestrian waves to signal the AV to stop and yield to the pedestrian, and the AV analyses the probability of a pedestrian waving on the zebra crossing as $\mathbb{P}(b = \text{wave}, s \subseteq R_{\text{crosswalk}})$. These findings show that environmental information is crucial for understanding human behavioural intentions. To achieve a more accurate understanding of behaviour, we have verified this process when fully navigating the task of picking up passengers with our autonomous bus.

We compared our proposed method on the CMHBD-AD dataset with SOTA methods, including the graph convolutional neural network (GCN)-based method MST–GCN [30] and the hypergraph GCN (HGNN)-based methods HyperGNN [31], 2s-AGCN [32],

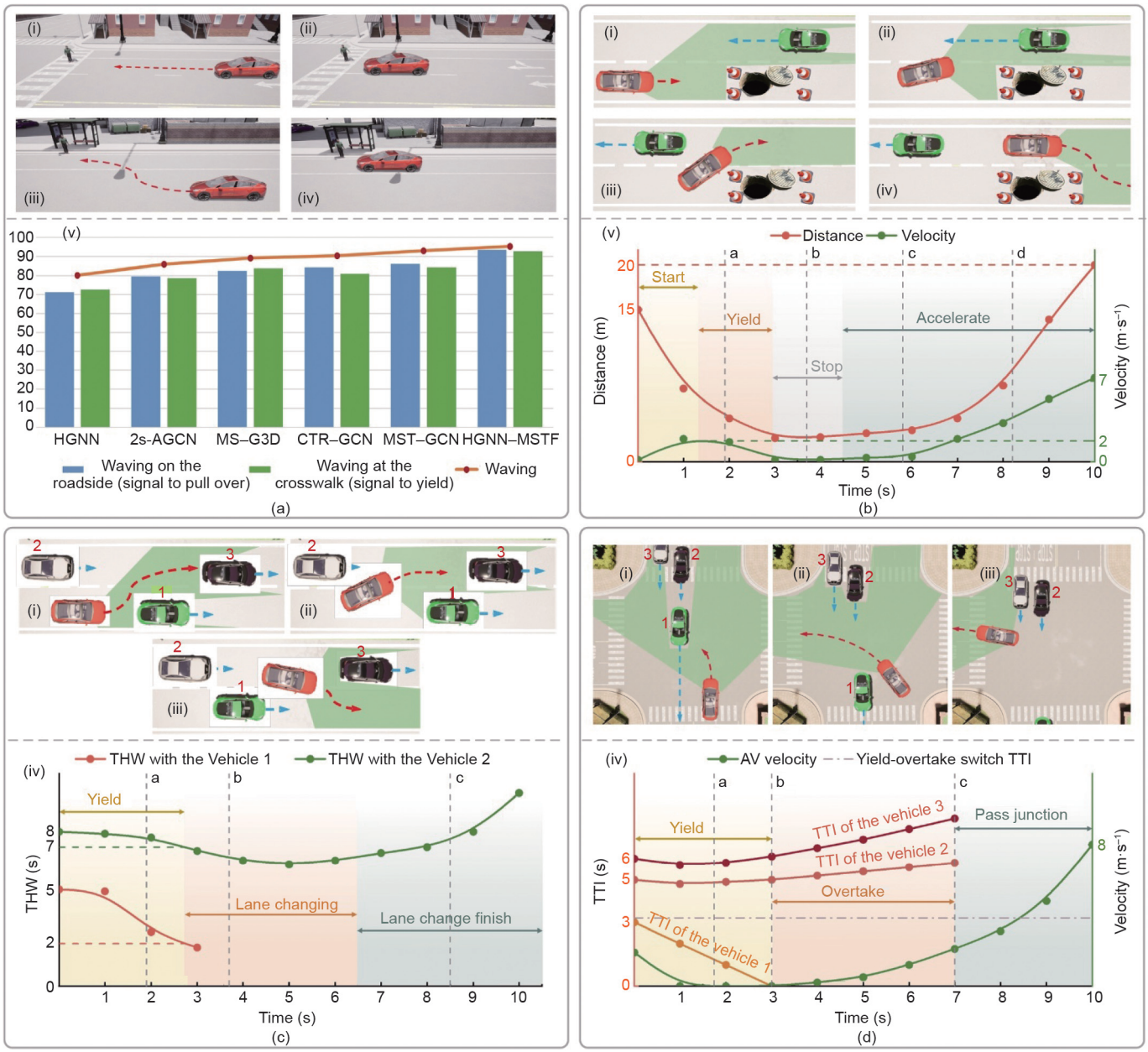


Fig. 5. Complex vehicle environment interaction scenarios. (a) Pedestrian waving under different scenarios. (a-i)–(a-iv) Sequential timesteps of AV interaction based on context-dependent behavior recognition. (a-v) Action recognition performance comparison on CMHBD-AD dataset. (b) Narrow road encounter. (b-i)–(b-iv) Sequential cooperative yielding interaction. (b-v) Distance-velocity profiles with interaction phase annotations (timestamps a–d correspond to (b-i)–(b-iv)). (c) Overtaking and lane change. (c-i)–(c-iii) Sequential overtaking maneuver with interactive decision-making. (c-iv) Time head way (THW) profiles with Vehicle 1 (red) and Vehicle 2 (green) during lane change (timestamps a–c correspond to (c-i)–(c-iii)). (d) Unprotected left turn against oncoming traffic. (d-i)–(d-iii) Sequential interactive decision-making based on right-of-way calculation. (d-iv) Time to intersection (TTI) profiles for oncoming vehicles, AV velocity (purple), and yield-overtake switch TTI (gray dashed line) (timestamps a–c correspond to (c-i)–(c-iii)).

MS–G3D [33], CTR–GCN [34], and HGNN [35] (Fig. 5(a-v)). The results indicate that our approach surpasses other algorithms in tasks that only require recognition of actions, and it demonstrates superior efficacy in identifying interaction intentions across various contexts. This finding underscores our method's enhanced ability to discern interaction intentions that are contingent upon the context.

3.2.2. Human action recognition based on high-order semantic relationships

In addition to excelling in CMHBD-AD, our HGNN-MSTF model matches the performance of state-of-the-art models but with relatively few parameters. Specifically, this model constructs a spatiotemporal hypergraph with a dynamic attention mechanism to capture high-order semantic feature relationships. We used the NTU-RGB+D dataset, which encompasses a wide range of daily human actions, to validate the robustness of the HGNN-MSTF. This dataset includes 60 types of behaviours performed by 40 volunteers from three camera views. We conducted two types of evaluations: ① cross-performer evaluation (X-sub), where training data come from 20 subjects and test data come from another 20, and ② cross-view evaluation (X-view), where training data come from camera views 2 and 3 and test data come from camera view 1.

Our experimental results corroborate our previous work [27], in which the HGNN-MSTF achieved accuracies of 91.2% on X-sub and 96.5% on X-view, demonstrating its strong cognition capabilities.

This model matches the performance of the GCN-based models while reducing the number of parameters by 5.4% and 81.6% compared with the MST–GCN and Dynamic GCN models, respectively, thus improving the computational efficiency. Compared with the STH–DRL, the HGNN-MSTF improves the X-sub evaluation benchmark by 0.4% [36]; it improves the X-view evaluation benchmark by 0.7% relative to the HyperGNN. In addition, we conducted ablation experiments of the HGNN-MSTF to demonstrate that jointly extracting spatial and temporal features benefits the action recognition task, achieving better performance than models with either spatial or temporal hypergraphs, as shown in Fig. 6(a).

3.3. Embodied interactive intelligence: Vehicle-to-vehicle

We propose a JTPWM-DRL to facilitate the interaction of AVs with surrounding vehicles in accordance with social norms. In this section, JTPWM-DRL is compared with SOTA algorithms across three complex interaction scenarios chosen from the autonomous bus navigation route. We used two typical indices to monitor the interaction states between vehicles. The first index is the time to intersection (TTI), which refers to the time from a vehicle to a path intersection point (Fig. 6(b)). The second index is the time headway (THW), which refers to the time gap between two vehicles travelling in the same direction (Fig. 6(b-i)). Both indices indicate how a vehicle perceives and responds to the dynamics of its surrounding traffic conditions, for example, the interaction states. Detailed scenario descriptions are provided in the following sections.

3.3.1. Narrow road encounter

Encountering involves two or more vehicles travelling in opposite directions interacting and cooperating to pass through a particular point. Vehicles must adhere to relevant regulations and yield promptly and reasonably. The interaction between AVs and oncoming vehicles when obstacles are encountered ahead on a narrow road is shown in Fig. 5(b). The AV needs to take a detour to avoid obstacles, but owing to its line of sight being blocked, it does not notice a green vehicle driving on the opposite side (Fig. 5(b-i)). After discovering the green vehicle, the AV stops to make space, allowing the other vehicle to go first (Fig. 5(b-ii)); then, the AV takes a detour to bypass obstacles and drive back to its original lane (Figs. 5(b-iii) and (b-iv)). This scenario demonstrates the excellent adaptability and the coordinated interaction capabilities of AD when unforeseen events are encountered. The relationships between the distance and speed of AVs and vehicles driving in the opposite lane in the narrow-road scenario and the interaction process are shown in Fig. 5(b-v). The red curve represents the changes in the distance between the AV and the vehicle driving in the opposite lane. The green curve illustrates the velocity profile of the AV. a, b, c, and d correspond to the timestamps referenced in the subplots of Figs. 5(b-i)–(b-iv). Start represents the starting phase of the AV, Yield represents the deceleration phase of the AV to yield to oncoming vehicles, Stop represents the parking phase of the AV, and Acceleration represents the acceleration phase of the AV through narrow roads.

3.3.2. Overtaking and lane change

During operation, AVs encounter a variety of scenarios, such as obstacles in the path and slow-moving vehicles ahead. These situations often necessitate manoeuvres such as overtaking and changing lanes. Fig. 5(c) illustrates that when encountering a slow car ahead, the AV successfully overtakes and merges into the faster lane by predicting the trajectory of adjacent white vehicles and calculating the right of way (Figs. 5(c-i) and (c-ii)). Then, following the same process, the AV bypasses the slow black car ahead and returns to its original lane (Fig. 5(c-iii)). The characteristic of this scenario is that AVs need to continuously calculate the right of way and adopt efficient embodied strategies to perform reasonable overtaking and lane change manoeuvres. The THW and lane change decisions of the AV, the leading vehicle, and the following vehicle during overtaking and merging are shown in Fig. 5(c–iv), illustrating the interaction process. The red curve represents the THW profile of the AV with respect to the leading vehicle. The green curve portrays the THW profile of the AV with the following vehicle in the adjacent lane. a, b, and c correspond to the timestamps referenced in the subplots of Figs. 5(c-i)–(c-iii).

3.3.3. Unprotected left turn against oncoming traffic flow

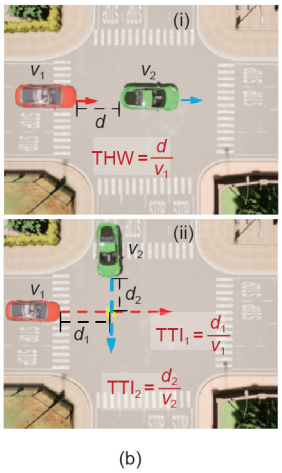
The embodied interaction intelligence of vehicles is manifested in complex situations such as unprotected left turns. At an unprotected intersection, a traffic light controls the behaviour of both left-turning (AVs) and oncoming vehicles simultaneously. AVs

Method	Accuracy	Precision	Recall	F1
HGNN-SPA	88.99%	90.18%	89.74%	89.36%
HGNN-TEM	92.19%	87.47%	87.18%	86.99%
HGNN-MSTF	93.47%	93.47%	92.31%	91.86%

(a)

Method	Basic interaction scenarios		Complex interaction scenarios	
	RC	DS	RC	DS
InterFuser (20)	88.90	51.60	87.0	53.0
LAV (21)	91.55	45.20	82.0	49.0
JTPWM-DRL	92.42	46.39	90.1	63.9

(c)



(b)

Fig. 6. (a) Ablation results of the HGNN-MSTF model. The HGNN-SPA model has only spatial hypergraphs, whereas the HGNN-TEM model has only temporal hypergraphs. The HGNN-MSTF model is a complete model with both spatial and temporal hypergraphs. (b) Illustration of time to intersection and time headway. (b-i) A vehicle to a path intersection point; (b-ii) two vehicles travelling in the same direction. v_1 and v_2 represent velocities of cars respectively. d represents the distance between two cars as shown in the figure. (c) Vehicle-to-vehicle interaction experiment. The basic interaction scenarios include following a car in a straight line, navigating lane changes, and making right turns. The complex interaction scenarios include narrow road meetings, overtaking and merging, and unprotected left turns facing oncoming straight traffic. The InterFuser is a cross-modal end-end AD navigation algorithm [37], and the LAV is an SOTA of the IL AD algorithm [38]. RC: the route completion; DS: driving score.

need to find the right of way for left turns amidst busy oncoming traffic while ensuring that there are no collisions with other vehicles. Our AV efficiently and safely completes a left turn by right-of-way calculation and strategy optimization amidst oncoming traffic, as shown in Fig. 5(d). Upon entering the intersection, the AV detects two oncoming vehicles and predicts their imminent crossing. Thus, the vehicle slows to yield (Fig. 5(d-i)). During the yielding process, another oncoming car appears. The AV anticipates a safe crossing before the car (Fig. 5(d-ii)). Noticing the car's deceleration and recognizing that the AV has the right of way, the AV continues to accelerate (Fig. 5(d-iii)). Finally, all the vehicles pass through the intersection safely and efficiently. In this scenario, with JTPWM-DRL, the AV can actively shuttle among traffic flows while ensuring safety. The interaction decision made by the AV based on the TTI of oncoming vehicles during an unprotected left turn is shown in Fig. 5(d-iv). The purple curve shows the AV's velocity profile. a, b, and c correspond to the timestamps referenced in the subplots of Figs. 5(d-i)–(d-iii). The grey dashed line represents the yield-overtake switch TTI. When the TTI is less than the switch TTI, the AV yields, and when the TTI is greater than the switch TTI, the AV overtakes.

3.3.4. Vehicle-to-vehicle interaction experiment results

We evaluated AD models on basic interaction scenarios in Town05 provided by the CARLA simulation platform to study vehicle-to-vehicle interactions. Additionally, we established the three aforementioned complex interaction scenarios for further examination. The experimental results are shown in Fig. 6(c). In the experiment, we use two metrics from the CARLA leaderboard: ① the route completion (RC) rate, which is the percentage of the actual driving distance of the AV to the prescribed route length (i.e., if the AV deviates from the prescribed route or does not yield to emergency vehicles as required during driving, then this part of the driving distance is not included in the calculation of the RC), and ② the driving score (DS), which is the product of the RC rate and the penalty coefficient for various violations. Violations include collisions with other traffic participants and the running of red lights, and each violation corresponds to a different penalty coefficient. For example, given a driving route of length 1000 m, if the actual path driven by the AV is only 900 m of the prescribed driving route, then RC = 90. During the driving process, if the AV has a collision with a pedestrian (penalty coefficient of 0.5) and runs a red light (penalty coefficient of 0.7), then the final DS is $DS = 90 \times 0.5 \times 0.7 = 31.5$.

The experimental results demonstrate that our method achieves the highest RC for basic interaction scenarios, surpassing InterFuser by 4% and LAV by 0.95%. Furthermore, our method stands out in complex interaction scenario experiments, outperforming LAV and InterFuser by 9.88% and 3.56% in terms of RC, respectively. In terms of DS, our method scores 30.4% and 20.1% higher than LAV and InterFuser, respectively.

3.4. Embodied interactive intelligence: Vehicle-to-environment

We designed a navigation task for autonomous buses to pick up passengers, aiming to evaluate the interaction capabilities of AD models in complex environments involving humans, vehicles, and roads. Along the route, the AV experienced five sequential interaction scenarios: ① unprotected left turn against oncoming traffic flow; ② picking up passengers waving at the bus stop; ③ narrow road encounter; ④ yielding pedestrians waving at the crosswalk; ⑤ overtaking and lane changing. In our experiments, we examined a scenario in which multiple buses stop at a single station. To enhance operational efficiency, the bus only came to a halt when passengers wave to signal their intention to board. The evaluation metrics included the RC of the total route, the average DS of the scenarios, the number of collisions, and whether the AV could successfully pick passengers if they waved at the bus stop. We proposed the end-to-end UniCVE model and compared it with the SOTA models. As illustrated in Fig. 7(a), the UniCVE model achieved an RC of 100 and the highest DS of 83, which exceeded those of the LAV and InterFuser models by 60.8% and 90.9%, respectively. Furthermore, our approach successfully avoided collisions. The low variance in our results (variance = 1.32 for DS, variance = 0 for RC) compared with the baseline methods demonstrated superior stability across different experimental conditions.

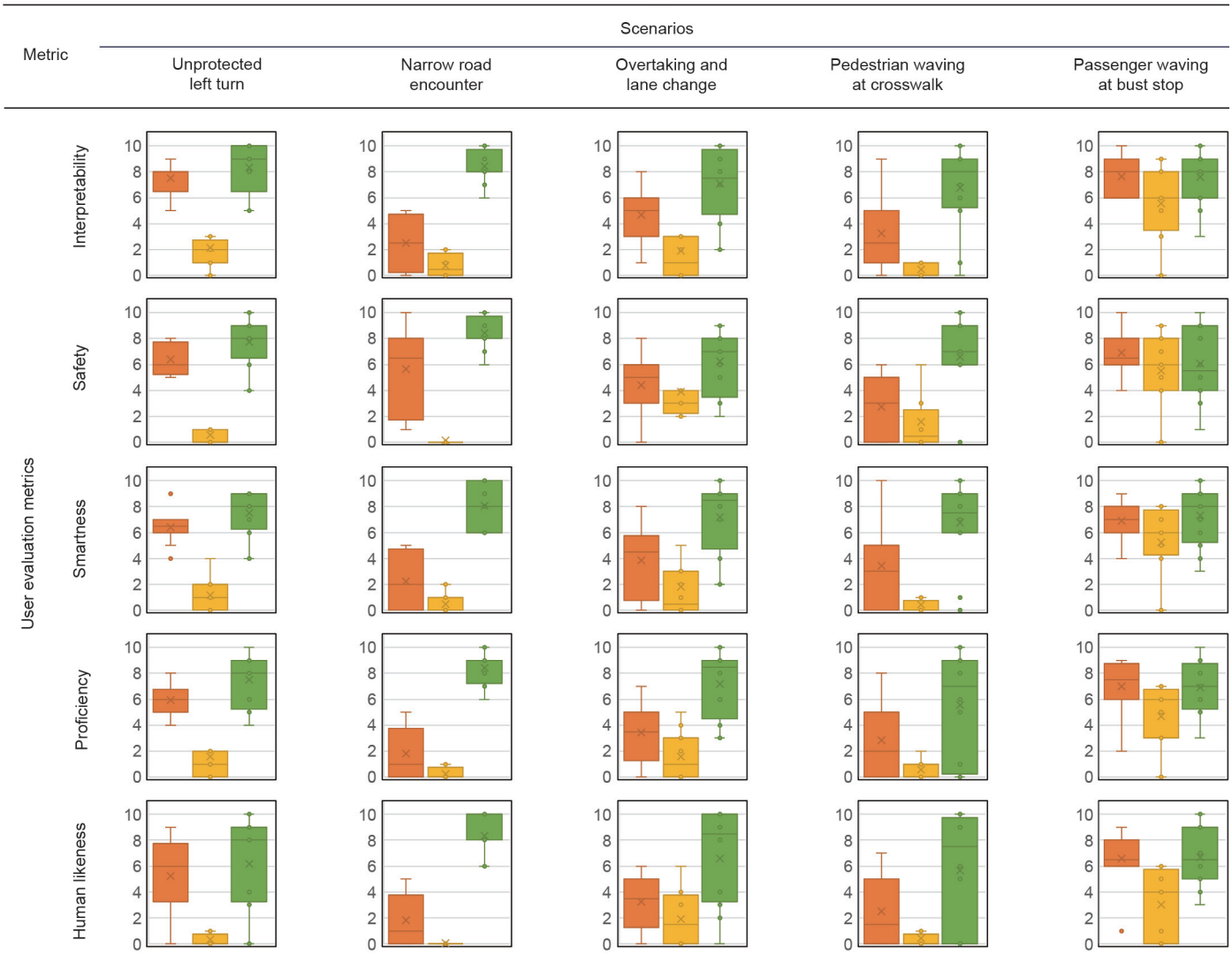
We conducted a user study employing a questionnaire to evaluate the interaction intelligence of various algorithms in complex scenarios, as depicted in Fig. 7(b). For this study, participants were shown videos of all the simulated interaction scenarios encountered along the navigation route. In these videos, the AV was equipped with different AD models, including LAV, InterFuser, and our UniCVE. Participants were not informed about the underlying AD models used in the simulations. They were asked to rate their level of agreement with a series of predefined statements on a scale from 0 to 10. The scoring was as follows: 0–2 indicated "strong disagreement," 3–4 indicated "disagreement," 5–6 indicated "neutrality," 7–8 indicated "agreement," and 9–10 indicated "strong agreement." The statements were designed to gauge various aspects of embodied interactive intelligence, namely, ① interpretability: I can always understand the car's decision; ② safety: I would feel safe riding in this car; ③ smartness: I think the car is smart; ④ proficiency: I think the car behaves as a human expert; and ⑤ human likeness: I would make the same decision with the car. Data were collected from a sample of 40 participants, comprising both undergraduate and postgraduate students.

We categorized the user study participants into two groups on the basis of their driving background: experienced drivers and nondrivers. The demographic analysis revealed nuanced differences in system evaluation patterns between these two groups. Notably, UniCVE maintained exceptional and consistent performance across both demographic segments, achieving nearly identical ratings (7.35 for nondrivers vs 7.46 for experienced drivers, representing only a 1.5% variance). This minimal performance difference suggests that UniCVE's design principles effectively transcend user experience backgrounds. The metric-specific analysis indicated that students without driving experience rated UniCVE's human-like behaviour 4.8% higher (7.0 vs 6.68) and interpretability marginally better (7.75 vs 7.67), whereas experienced student drivers showed slightly greater appreciation for safety features (7.02 vs 6.7). These variations reflected the different cognitive frameworks through which each group evaluated AD systems: inexperienced drivers appeared more receptive to AI-driven behavioural patterns, whereas experienced drivers demonstrated heightened sensitivity to safety-critical aspects on the basis of their practical driving knowledge. The overall consistency in the performance of the UniCVE model across these demographic segments validated the its robust design and broad applicability within academic populations. In future work, we plan to incorporate these demographic insights to develop adaptive user interfaces and expand our evaluation to broader age ranges and professional backgrounds to further validate the system universality.

Safety is the primary metric for evaluating interactive intelligence and the usability of AD algorithms. As illustrated in Fig. 6(b), UniCVE outperforms the other algorithms in terms of safety across all the scenarios, with average safety scores of 7.02, compared with 5.23 for LAV and 2.35 for InterFuser. This difference represents 34.2% and 198.7% improvements over LAV and InterFuser, respectively. Interpretability, the degree to which humans can understand AV behaviours, is vital for cooperation and traffic efficiency. On average, UniCVE, LAV, and InterFuser

Method	Evaluation metrics for embodied interaction intelligence						
	RC		DS		Times of collision	Pick up passengers	User study score
	Mean	Var	Mean	Var			
InterFuser (20)	73.35	3.56	43.47	2.74	1	N	1.85
LAV (22)	92.37	5.31	51.60	1.56	1	N	4.60
UniCVE	100.00	0	83.00	1.32	0	Y	7.17

(a)



(b)

Method	RC		DS		Times of collision
	Mean	Var	Mean	Var	
UniCVE with CTX-HG $K=5$	100.00	0	83.00	1.32	1
UniCVE with CTX-HG $K=10$	90.11	0.34	51.00	0.97	0
UniCVE with CTX-HG $K=15$	93.57	0.27	61.00	1.21	0
UniCVE with CTX-HG $K=20$	100.00	0	78.00	1.43	1
UniCVE w/o CTX-HG	98.21	0.41	64.00	1.57	2

(c)

Fig. 7. Experimental results of the autonomous bus transporting passenger navigation task. (a) Interaction experiments of pedestrians, vehicles, and roads in intricate environments. (b) Comparative analysis of AD models via user questionnaire evaluation. The plots illustrate the standardized mean differences in scores for three distinct AD models, each evaluated across a comprehensive set of metrics and interaction scenarios. The red bars denote the data corresponding to the LAV model, the yellow bars represent the InterFuser model, and the green bars depict the data for our proposed UniCVE model. (c) Ablation results of the scenario context hypergraph. CTX-HG represents the context hypergraph network. K is the K value in the KNN used to construct hyperedges in the hypergraph network. UniCVE w/o CTX-HG denotes the UniCVE model without the context hypergraph module. Var: variance.

achieve interpretability scores of 7.67, 5.12, and 2.18, respectively, which demonstrates that the UniCVE model is more comprehensible and compatible with human users. Notably, the UniCVE model significantly outperforms the other models in the narrow-road encounter scenario, not only with higher average scores but also with a smaller standard deviation. This scenario requires the AV to actively use the opposite lane to bypass a static obstacle, which is a comprehensive reflection of smartness, proficiency, and human likeness.

Furthermore, we conducted ablation experiments on overtaking and lane change scenarios to demonstrate that compared with bare BEV features (UniCVE w/o CTX-HG), extracting scenario contextual features with hypergraphs improves the DS, as shown in Fig. 7(c). In addition, the results show that the K value of the KNN used to construct hyperedges strongly affects the performance of the model. The optimal K value is 10 because a value of 5 is too small to obtain contextual relationships and a value of 20 leads to the selection of too many nodes for an edge without any specific relationship.

In addition to the five scenarios, there are many other scenarios. With its self-learning and self-growth capabilities, the UniCVE algorithm can use big data to approximate infinity, thereby achieving embodied intelligence.

3.5. Discussion

Embodied interaction intelligence is crucial for human machine collaboration, as it empowers AVs to interact, learn, and earn trust amongst humans. Our end-to-end AD model, UniCVE, offers distinct advantages in terms of perception and behavioural intelligence: ① it enhances perception accuracy through multiview cross-modal information fusion; ② it designs interaction cognition models that are deliberately tailored to different interaction objects, thereby capturing high-order semantic features of interaction and enhancing the rationality of AV behaviours in interaction scenarios; ③ it allows a LLM to incorporate human driving language guidance, enhancing scene semantic comprehension and decision-making in complex driving environments. We delve into these aspects in the subsequent discussion.

To address diverse perception sources, heterogeneous data, and complex driving environments, we design a cross-modal perception system and propose a hypergraph-based human action recognition model that leverages multiview spatiotemporal features. Our system outperforms pure visual systems by effectively using complementary 2D image and 3D point cloud data, significantly improving human action recognition accuracy and robustness under varying lighting and visibility conditions. Furthermore, our model captures high-order semantic associations through hyperedges among joints and limbs, enabling precise behaviour understanding. This model employs a dynamic spatiotemporal attention-based hypergraph convolution model to minimize information redundancy and training complexity, thereby saving computational resources and time. In future research, we need to address the adaptive construction of high-quality hypergraph structures in the absence of sufficient prior knowledge.

Considering complex and variable driving environments, RL methods that concurrently learn world and strategy models can struggle to converge. To mitigate this issue, we pretrain the world model on a large dataset via supervised learning, establishing vehicle interaction knowledge for AVs and reducing ineffective RL sampling. Our proposed JTPWM-DRL mitigates collision risk by estimating the joint probability distribution of AV and surrounding vehicle trajectories and inferring potential interaction scenarios. However, supervised learning models rely heavily on the quality and quantity of labelled data, making the acquisition of ample accurate data for AV trajectory prediction challenging. Addressing data's long-tail distribution through effective model training can enhance the generalizability.

LLMs have showcased their robust understanding and reasoning capabilities across various applications, effectively encapsulating human knowledge and wisdom. To leverage the performance of the LLMs while ensuring real-time responses in AD systems, we use a knowledge distillation method. This method transfers knowledge from the teacher LLM to a smaller but faster student model, enhancing AV driving capabilities in unknown obstacle scenarios. Future research involves designing efficient student models to improve knowledge integration and deploying large models in AD systems without compromising inference speed or LLM reasoning capabilities.

4. Conclusions and implications

Overall, our work introduces an end-to-end perception–cognition–behaviour closed-loop feedback framework for AD. This framework integrates the understanding of interactions with different traffic participants in various driving environments as unified rewards and soft constraints to enable AVs to drive in a socially compatible and human-predictable manner. By utilizing the RL paradigm, our model is capable of continual learning and self-growth. Our user study demonstrates that this model outperforms other state-of-the-art algorithms in terms of interaction intelligence, showing greater proficiency and human-like behaviour in complex simulated environments. Notably, we have successfully deployed the model on real buses, achieving 22 000 km of travel and completing 45 000 navigation tasks in the Xiong'an New Area, China. This deployment underscores our contribution to advancing embodied interactive intelligence in AD. However, our qualitative analysis reveals certain limitations in edge case scenarios. Specifically, at intersections with significant visual occlusions, pedestrians may suddenly emerge from behind parked vehicles or architectural barriers, creating challenging reactive scenarios. In these cases, our model initially exhibits delayed braking responses because of insufficient early visual cues and limited prediction time windows. Interestingly, we have observed that through repeated exposure to similar occlusion patterns at specific intersection locations, our model demonstrates memory-based adaptive learning—the development of enhanced anticipatory behaviours that enable more proactive safety responses in familiar high-risk areas. This finding suggests that experience accumulation at challenging locations contributes to improved risk assessment and behavioural planning over time.

In future work, we will focus on addressing these limitations by enhancing occlusion-aware perception, incorporating uncertainty-aware trajectory prediction, and strengthening memory-based modules to better anticipate latent risks across diverse intersection geometries. Moreover, we plan to explore real-world deployment and multiagent cooperative scenarios to further test the robustness and scalability of our framework.

CRediT authorship contribution statement

Nan Ma: Writing – original draft, Methodology, Conceptualization. **Jia Pan:** Writing – original draft, Supervision. **Yongjin Liu:** Writing – original draft, Supervision. **Yajue Yang:** Writing – original draft, Supervision, Investigation. **Yiheng Han:** Writing – original draft, Supervision. **Jiacheng Guo:** Writing – original draft, Validation. **Zhixuan Wu:** Validation. **Zecheng Yang:** Visualization, Validation. **Zhiwei Yang:** Resources, Data curation. **Deyi Li:** Writing – original draft, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62371013), the National Key Research and Development Program of China (2023YFF0615800), the National Natural Science Foundation of China-Research Grants Council (NSFC-RGC) Joint Research Scheme (62461160309), and the Beijing Natural Science Foundation (L247007).

References

[1] Crosato L, Tian K, Shum HPH, Ho ESL, Wang Y, Wei C. Social interaction-aware dynamical models and decision-making for autonomous vehicles. Adv Intell Syst 2024;6(3):2300575.
[2] Ettinger S, Cheng S, Caine B, Liu C, Zhao H, Pradhan S, et al. Large scale interactive motion forecasting for autonomous driving: the Waymo open motion dataset. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. Piscataway: IEEE; 2021. p. 9690–9.
[3] Huang Z, Liu H, Lv C. Gameformer: game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France. Piscataway: IEEE; 2023. p. 3880–90.
[4] Wang WS, Wang L, Zhang C, Liu C, Sun L. Social interactions for autonomous driving: a review and perspectives. Found Trends Robotics 2022;10:198–376.
[5] Luo W, Park C, Cornman A, Sapp B, Anguelov D. JFP: joint future prediction with interactive multi-agent modeling for autonomous driving. In: Proceedings of the 6th Conference on Robot Learning; 2022 Dec 14–18; Auckland, New Zealand. New York City: PLMR; 2022. p. 1457–67.
[6] Xia C, Xing M, He S. Interactive planning for autonomous driving in intersection scenarios without traffic signs. IEEE Trans Intell Transp Syst 2022;23(12):24818–28.
[7] Floreano D, Mondada F, Perez-Uribe A, Roggen D. Evolution of embodied intelligence. In: Iida F, Pfeifer R, Steels L, Kuniyoshi Y, editors. Embodied artificial intelligence. Berlin: Springer; 2004. p. 291–311.
[8] Roy N, Posner I, Barfoot T, Beaudoin P, Bengio Y, Bohg J, et al. From machine learning to robotics: challenges and opportunities for embodied intelligence. 2021. arXiv:2110.15245.
[9] Howard D, Eiben AE, Kennedy DF, Mouret JB, Valencia P, Winkler D. Evolving embodied intelligence from materials to machines. Nat Mach Intell 2019;1(1):12–9.
[10] Gupta A, Savarese S, Ganguli S, Li F. Embodied intelligence via learning and evolution. Nat commun 2021;12:5721.
[11] Mengaldo G, Renda F, Brunton SL, Bächer M, Calisti M, Duriez C, et al. A concise guide to modelling the physics of embodied intelligence in soft robotics. Nat Rev Phys 2022;4(9):595–610.
[12] Cross ES, Ramsey R. Mind meets machine: towards a cognitive science of human–machine interactions. Trends Cogn Sci 2021;25(3):200–12.
[13] Hoc JM. Towards a cognitive approach to human–machine cooperation in dynamic situations. Int J Hum Comput Stud 2001;54(4):509–40.
[14] Chougule A, Chamola V, Sam A, Yu FR, Sikdar B. A Comprehensive review on limitations of autonomous driving and its impact on accidents and collisions. IEEE Open J Veh Technol 2023;5:142–61.
[15] Ryan C, Murphy F, Mullins M. End-to-end autonomous driving risk analysis: a behavioural anomaly detection approach. IEEE Trans Intell Transp Syst 2021;22(3):1650–62.
[16] Chia WMD, Keoh SL, Goh C, Johnson C. Risk assessment methodologies for autonomous driving: a survey. IEEE Trans Intell Transp Syst 2022;23(10):16923–39.
[17] Cao Z, Jiang K, Zhou W, Xu S, Peng H, Yang D. Continuous improvement of self-driving cars using dynamic confidence-aware reinforcement learning. Nat Mach Intell 2023;5(2):145–58.
[18] Zhuang H, Fang D, Tong K, Liu Y, Zeng Z, Zhou X, et al. Online analytic exemplar-free continual learning with large models for imbalanced autonomous driving task. IEEE Trans Vehicular Technol 2024;74(2):1949–58.
[19] Niu H, Xu Y, Jiang X, Hu J. Continual driving policy optimization with closed-loop individualized curricula. In: Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA); 2024 May 13–17; Yokohama, Japan. Piscataway: IEEE; 2024. p. 6850–7.
[20] Li D. Cognitive physics—the enlightenment by Schrödinger, turing, and wiener and beyond. Intell Comput 2023;2:0009.
[21] Hafner D, Lillicrap T, Ba J, Norouzi M. Dream to control: learning behaviors by latent imagination. 2019. arXiv:1912.01603.
[22] Zhu W, Hayashibe M. Autonomous navigation system in pedestrian scenarios using a dreamer-based motion planner. IEEE Robot Autom Lett 2023;8(6):3836–43.
[23] Gao Y, Zhang Q, Ding DW, Zhao D. Dream to drive with predictive individual world model. IEEE Trans Intell Veh 2024;9(12):8224–38.
[24] Shao H, Hu Y, Wang L, Song G, Waslander SL, Liu Y, et al. Lmdrive: closed-loop end-to-end driving with large language models. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. Piscataway: IEEE; 2024. p. 15120–30.
[25] Xu Z, Zhang Y, Xie E, Zhao Z, Guo Y, Wong KYK, et al. Drivegpt4: interpretable end-to-end autonomous driving via large language model. IEEE Robot Autom Lett 2024;9(10):8186–93.
[26] Fu D, Li X, Wen L, Dou M, Cai P, Shi B, et al. Drive like a human: rethinking autonomous driving with large language models. In: Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW); 2024 Jan 1–6; Waikoloa, HI, USA. Piscataway: IEEE; 2024. p. 910–9.
[27] Ma N, Wu Z, Feng Y, Wang C, Gao Y. Multi-view time-series hypergraph neural network for action recognition. IEEE Trans Image Process 2024;33:3301–13.
[28] Ma N. Cross-modal human behavior dataset for autonomous driving (CMHBD-AD) [Internet]. Beijing: e Intelligent Interaction Team. 2024 [cited 2025 Sep 19]. Available from: http://www.mananlab.tech/Cross-Modal-Human-Behavior. Chinese.
[29] Shahroudy A, Liu J, Ng T, Wang G. NTU RGB+D: a large scale dataset for 3D human activity analysis. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. Piscataway: IEEE; 2016. p. 1010–9.
[30] Chen Z, Li S, Yang B, Li Q, Liu H. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. Proc Conf AAAI Artif Intell 2021;35(2):1113–22.
[31] Chen Y, Li Y, Zhang C, Zhou H, Luo Y, Hu C. Informed patch enhanced HyperGCN for skeleton-based action recognition. Inf Process Manage 2022;59(4):102950.
[32] Shi L, Zhang YF, Cheng J, Lu H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. Piscataway: IEEE; 2019. p. 12018–27.
[33] Liu ZY, Zhang HW, Chen ZH, Wang Z, Ouyang W. Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. Piscataway: IEEE; 2020. p. 140–9.
[34] Chen Y, Zhang Z, Yuan C, Li B, Deng Y, Hu W. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. Piscataway: IEEE; 2021. p. 13339–48.
[35] Feng YF, You H, Zhang Z, Ji R, Gao Y. Hypergraph neural networks. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence; 2019 Jan 27–Feb 1; Honolulu, HI, USA. Palo Alto: AAAI Press; 2019. p. 3558–65.
[36] Nikpour B, Armanfard N. Spatio–temporal hard attention learning for skeleton-based activity recognition. Pattern Recognit 2023;139:109428.
[37] Shao H, Wang L, Chen R, Li H, Liu Y. Safety-enhanced autonomous driving using interpretable sensor fusion transformer. In: Proceedings of the 6th Conference on Robot Learning; 2022 Dec 14–18; Auckland, New Zealand. New York: PLMR; 2022. p. 726–37.
[38] Chen D, Krähenbühl P. Learning from all vehicles. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. Piscataway: IEEE; 2022. p. 17201–10.
[39] Lang AH, Vora S, Caesar H, Zhou L, Yang J, Beijbom O. Pointpillars: fast encoders for object detection from point clouds. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. Piscataway: IEEE; 2019. p. 12689–97.
[40] Wu Z, Ma N, Wang C, Xu C, Xu G, Li M. Spatial–temporal hypergraph based on dual-stage attention network for multi-view data lightweight action recognition. Pattern Recognit 2024;151:110427.