

Towards the Intelligence Evaluation of High-level Autonomous Vehicles: A Subjective-Objective Mapping Method via LLM

Jiarui Zhang¹, Jierui Chen¹, Shiyu Fang¹, Chao Huang², Peng Hang^{1,✉}, Jian Sun¹

Cite this article: <https://doi.org/10.26599/COMMTR.2026.9640041>

ABSTRACT: With the advent of the commercialization phase for autonomous vehicles (AVs), the evaluation of their intelligence has become essential for regulators. However, existing evaluation methods are still largely traditional and experience based. Therefore, this paper proposes a Subjective-Objective Mapping Evaluation Method to evaluate the intelligence of high-level Autonomous Driving Systems (ADS). First, a five-dimensional evaluation metrics system is developed to represent the overall performance of AV during testing or actual driving. Next, the performance of AVs in real-world driving dataset is evaluated based on the Large Language Model (LLM). This approach can significantly enhance evaluation efficiency, achieving a passing rate of 86.75% in the Turing Test. Finally, a deep neural network with attention mechanism is trained using quantified metrics and LLM-based evaluation label to serve as the evaluation model. Ablation experiments are conducted on both the LLM-based evaluation strategy and the evaluation model, demonstrating the necessity of each module. During the application phase, the evaluation model's effectiveness and reliability in scenario-based testing were validated using data from the OnSite Autonomous Driving Challenge Competition results, and comparative analysis was conducted against traditional evaluation methods. The results show that our model's evaluations align closely with those of human experts and outperform traditional evaluation methods.

KEYWORDS: Autonomous Vehicles, Intelligence Evaluation, Large Language Models, Interaction Behavior, Scenario-based Testing.

1 Introduction

AVs have reached the stage of commercial deployment, yet despite multiple ADS iterations and broad scenario testing, its performance on open roads often remains below expected levels. This imbalance largely stems from the gap between the high evaluation scores achieved during testing and the disappointing ride experience. In a real-world ride survey involving Level 4 autonomous vehicles, some participants noted subpar performance in several scenarios. Key issues included indecisive driving, abrupt braking, unnatural trajectory and speed control at intersections like roundabouts, and some vehicles even exhibited erratic lane changes, compromising both passenger experience and traffic efficiency (Johansson et al., 2023). Such observations collectively indicate that current evaluation practices in both academia and industry still exhibit significant limitations. Considering the ongoing transition of ADS from L2 systems to higher level L3+ systems, there is an urgent need to develop a systematic intelligence evaluation method specifically targeting these high-level ADS.

Existing research on evaluating ADS intelligence primarily relies on the integration of quantitative metrics, which mainly involves metric weighting and aggregation. Metrics weighting can be determined either through expert judgment or derived from data. Methods based on expert judgment, such as the Analytic Hierarchy Process (AHP), align evaluations with human insight (Darko et al., 2019), but differences in expertise may introduce variability and bias. Data-driven methods infer weights from large datasets,

reducing reliance on expert input, yet they depend on mathematical models and data validity, which can limit universal applicability. Metrics aggregation mainly includes linear aggregation (Sun et al., 2021), multi-attribute decision-making (Li et al., 2025), and fuzzy evaluation methods (Li and Kwan, 2003), with linear aggregation being most common. However, ADS intelligence involves complex nonlinear relationships and metric interdependencies, making linear and multi-attribute approaches less effective in capturing overall performance. Fuzzy evaluation can address metric uncertainty but still relies on expert-defined weights and membership settings, which may compromise robustness and consistency. However, most approaches focus on metrics within isolated modules and fail to capture the true performance of algorithms in complex environments. This limitation is particularly pronounced for L3+ ADS and contributes to the gap between test results and real-world user experience.

Upon deeper analysis, we uncover the two major challenges in intelligence evaluation of ADS: **1) High dimensionality and strong module coupling:** Evaluating low-level ADS typically involves assessing isolated sub-functions such as perception, prediction, planning, and control separately. In contrast, high-level ADS are often highly integrated systems — for example, end-to-end architectures — making it difficult to isolate and evaluate the performance of any single function. This necessitates a holistic assessment of overall system intelligence. Moreover, evaluating from a system-level perspective dramatically increases the number

¹ The Key Laboratory of Road and Traffic Engineering of the State, Ministry of Education, Tongji University, Shanghai 201804, China. ² School of Electrical and Mechanical Engineering, Faculty of Sciences, Engineering and Technology, Adelaide University, Adelaide 5000, Australia.

✉ Corresponding author. E-mail: hangpeng@tongji.edu.cn

Received: March 5, 2026; Revised: May 26, 2026; Accepted: July 29, 2026

© The Author(s) 2026. This is an open access article under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0, <http://creativecommons.org/licenses/by/4.0/>).

of metrics, further complicating the evaluation process. 2) **Difficulty in knowledge-driven subjective evaluation:** Current evaluations of low-level ADS rely almost exclusively on quantitative metrics, which primarily measure performance in well-defined or structured scenarios. High-level ADS, however, must handle unknown and uncertain environments, which place greater demands on aspects that are inherently difficult to quantify, such as human-likeness, generalization ability, and ethical reasoning. Consequently, evaluation of high-level ADS requires the integration of subjective, knowledge-driven assessments alongside traditional quantitative metrics.

Recent advances in neural networks have made it feasible to handle extremely high-dimensional input spaces. Moreover, as network size continues to grow, large models have demonstrated human-level thinking patterns, capable of intuitive reasoning. This progress provides a novel approach to addressing the challenges discussed above.

Motivated by this, we propose a Subjective-Objective Mapping Evaluation (SOME) framework for the intelligence of L3+ ADS. This framework combines quantitative data with qualitative evaluations generated by large models, effectively integrating objective measurements with subjective reasoning. The main contributions of this work are summarized as follows:

1) A Subjective-Objective Mapping Evaluation Framework is proposed for the intelligence of ADS, in which LLM-based reasoning is first used to transform interaction data into expert-style subjective judgments that are then aligned with objective metrics through a neural network model. This method leverages the powerful reasoning capabilities of LLMs and the learning generalization of deep neural networks, effectively reducing evaluation result bias caused by expert subjectivity differences and inaccurate metrics weighting, while significantly improving evaluation efficiency and enabling automated evaluation.

2) A multi-dimensional evaluation system for ADS intelligence is constructed, which covers safety, comfort, efficiency, social interaction, and impact on traffic system, providing a holistic depiction of ADS performance. Based on these metrics, a neural network model is then built by taking the normalized metric set as inputs and learning to map them to LLM-derived intelligence scores, enabling the system to capture nonlinear relationships across dimensions.

3) An LLM-based ADS evaluation dataset has been constructed. It includes AV interaction data, their natural-language descriptions, as well as evaluation results and quantitative scores. The dataset is created by converting raw trajectories into structured language via the Data-Language Mapping Module and obtaining expert-like evaluations from LLMs, supporting both LLM fine-tuning and evaluation validation.

The rest of the paper is organized as follows. Section 2 presents a literature review of existing studies on ADS evaluation dimensions and paradigms. Section 3 provides a detailed description of the construction of the evaluation system and evaluation model. In Section 4, we experimentally validate the reliability of the LLM-based evaluation results and the effectiveness of the proposed evaluation model. Finally, conclusions are made in Section 5.

2 Literature Review

Accurately evaluating the performance of AVs is a crucial aspect in the scenario-based testing. This section will summarize the progress and shortcomings in evaluation dimensions and evaluation methods in recent years and explain the motivation behind this research.

2.1. Evaluation dimensions

ADAS functions are designed to ensure the safety of vehicle operation, aiming to reduce safety risks caused by human drivers' delayed reactions, fatigue, or other factors. Therefore, traditional evaluations mainly focus on safety, such as the evaluation of AEB (Hu et al., 2022), Adaptive Cruise Control (Lee et al., 2021), and Lane Keeping Assist (Wei et al., 2023). With the development of ADS, the demand for AVs has further increased. Firstly, as a vehicle, comfort directly affects the passenger experience. Winkel et al. (2023) conducted a study on longitudinal comfort, verifying the relationship between acceleration, jerk, and passenger discomfort. A comfort analysis method based on road information was proposed by Chen et al. (2025) to evaluate the comfort of vehicles from a macro perspective (traffic flow and road conditions).

In addition, improving vehicle driving efficiency helps increase road throughput (Montanaro et al., 2019), and efficiency is also one of the key indicators of high intelligence in AVs. Apollo Go AVs in Wuhan, China experienced "deadlock" at intersections due to insufficient intelligence, while Waymo's Robotaxi was stuck in traffic for over 15 minutes on San Francisco streets, both causing severe impacts on the urban traffic system. Existing studies primarily assess vehicle efficiency in terms of speed and energy consumption (Zhu et al., 2020). Chen et al. (2017) investigated the impact of vehicle automation on traffic system performance by modeling roadway capacity in mixed traffic flows, highlighting how different levels of automation can alter traffic flow characteristics and overall road capacity. Additionally, since most existing AVs are new energy vehicles, their energy consumption during operation is also a key focus of public attention (Sun et al., 2025). Finally, AVs' social interaction capability, are important dimensions for evaluating their intelligence (Zhao et al., 2023), as the ability of understanding and reacting to the behaviors of other road users can significantly enhance AVs' rapid deployment. To accurately evaluate the social interaction capability of AVs in complex traffic environments, a multi-agent simulation model named Social-MAIL was proposed by Wang et al. (2025). Zhao et al. (2024) introduced the Interaction Preference Value to quantify human driving preferences, incorporating it into AV planning for more human-like interactions. Similarly, the Interaction Orientation (IO) metric was developed by Liu et al. (2024) to assess the degree of human-likeness in AVs' interactions.

However, most existing studies evaluate the performance of ADS from a specific dimension. While this approach accurately captures the system's performance in these dimensions, it fails to capture the overall intelligence of the system. This limitation makes it difficult to conduct holistic assessments and achieve global optimization of the system.

2.2. Evaluation methods

The evaluation methods must be designed after the determination of evaluation dimensions and metrics. Qualitative evaluation relies

on the experience and knowledge of experts, assessing the subject through methods such as inductive analysis based on their observed performance (Grosse and Alahi, 2024). Woo and Ahn, (2024) employed qualitative evaluation methods to assess the interaction level of ADS, using subjective judgments from participants to identify key driving characteristics of autonomous vehicles.

However, qualitative evaluation is highly subjective and lacks data support, which makes the evaluation results less persuasive. It is generally used only when the evaluation metrics are difficult to quantify. Quantitative evaluation involves key steps such as metrics definition, weighting and aggregation. Studies (Jahangirova et al., 2021; Guo et al., 2019) have proposed evaluation frameworks for quantifying the driving quality and overall performance of autonomous vehicles from multiple dimensions.

In terms of metric aggregation methods, Li et al. (2025) proposed a multi-dimensional evaluation framework for ADS scenarios-based testing on the TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) method. This method uses an objective approach for metric aggregation, reducing the impact of subjective factors. However, in practical applications, the ideal solution may be difficult to define or even nonexistent, and changes in metric weights can lead to significant fluctuations in the evaluation results. Wang et al. (2024) determined both the objective and subjective weights of the metrics using CRITIC and AHP, respectively, and applied an improved Gray Relational Analysis to aggregate metrics. This method combines objective and subjective weighting approaches to ensure the comprehensiveness and fairness of the evaluation, but the determination of the ratio coefficient between subjective and objective weighting remains a new challenge.

Therefore, traditional aggregation models still have drawbacks. Du et al. (2017) proposed an autonomous parking system evaluation model based on neural networks, where the metric results from the parking process are input into the neural network for system evaluation. The neural network can learn the complex relationships between the metrics by training effectively. However, this method relies on human experts' subjective evaluations, and different experts' knowledge, experience, and preferences can influence the evaluation results. Additionally, the method requires large amounts of sample data and relying on a single human expert to evaluate hundreds or even thousands of tests individually can lead to a decrease in efficiency.

2.3. Applications for LLM in AVs

Currently, several studies have integrated LLMs into autonomous driving systems (Qu et al., 2023; Zhang et al., 2024). For instance, Fang et al. (2025) proposed an LLM-based Actor-Reasoner module to enhance the interaction capabilities between autonomous vehicles and human-driven vehicles. GPT-Driver redefined motion planning tasks as a language modeling problem and leveraged the reasoning abilities of the GPT model to generate precise driving trajectories (Mao et al., 2023).

LLMs have also found widespread application in the field of autonomous driving testing and evaluation (Liu et al., 2023). LLMs

can generate massive test scenarios by simulating typical urban or highway traffic environments, providing autonomous driving systems with a safe and efficient testing platform and use cases (Zhao et al., 2026). Zhao et al. (2025) proposed a novel world model, DriveDreamer-2, based on LLMs, which can be used for rapid generations of long-tail scenarios. Some scholars have also applied artificial intelligence to evaluations in the fields of autonomous driving and transportation. You et al. (2025) proposed an LLM-based driving performance evaluation framework, which provides a comprehensive assessment of vehicle safety, comfort, and intelligence by building a real-world driving corpus knowledge base. However, LLMs still have limitations in understanding and applying domain-specific knowledge, and their inherent randomness can introduce potential bias in the results (Nie et al., 2025).

3 Methodology

This section presents the proposed LLM-based Subjective-Objective Mapping Evaluation framework for automated ADS intelligence evaluation. As shown in Fig. 1, the framework consists of three modules: dataset selection and processing, evaluation model construction, and model validation and application. The first module extracts high-quality interaction segments as evaluation and training samples; the second trains an attention-based deep neural network using calculated metric values as inputs and LLM-generated intelligence scores as supervisory labels; and the third validates and applies the proposed model.

3.1. Definition of ADS intelligence

In this study, the comprehensive intelligence of autonomous driving system is defined as "the capability of an ADS to perceive, understand, decide, and act appropriately in complex traffic interactions, while maintaining safe, comfortable, efficient, socially compatible, and surrounding-traffic friendly driving performance". Different from the intelligence of a single functional module, comprehensive intelligence emphasizes the system-level behavioral performance of ADS in interaction with surrounding road users and traffic environments.

From this perspective, autonomous driving intelligence is reflected in the coordinated performance of multiple behavioral dimensions. Safety serves as the fundamental baseline, ensuring that ADS can control risks and avoid conflicts. Efficiency and impact on surrounding traffic reflect whether the ADS can complete driving tasks effectively while minimizing disturbances to nearby traffic participants. Comfort and social interaction further characterize whether the ADS can generate human-acceptable motion and make reasonable decisions in interactions with other road users. Therefore, the proposed five-dimensional framework serves as an observable representation of comprehensive autonomous driving intelligence, where each dimension is regarded as a complementary manifestation of system-level intelligence rather than an isolated performance metric.

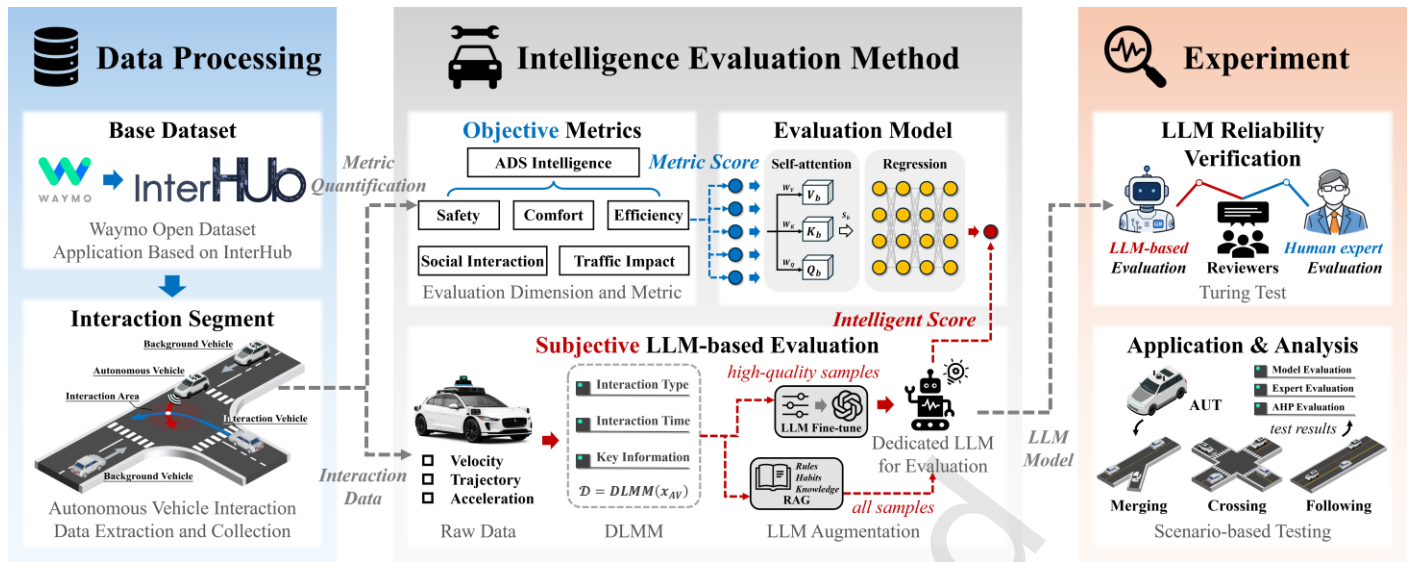


Fig. 1. Overview of the proposed evaluation model for ADS intelligence.

3.2. Dataset and processing

The Waymo Open Dataset (Sun et al., 2020) provides large-scale real-world driving scenarios and map information for motion prediction tasks and is used as the basis for training the evaluation model in this study. Since the dataset contains extensive human-driving trajectories and non-interactive segments, we used the InterHub toolkit (Jiang et al., 2025a) to extract high-interaction AV segments for accurate evaluation and model training.

However, the Waymo Open Dataset mainly contains normal and high-quality driving behaviors, which makes the evaluation scores concentrate in the high-score range and may limit the model's ability to learn low-intelligence behaviors and typical failure modes. To address this issue, this study augments the original dataset by generating low-score interaction segments based on the Waymo Open Dataset. Specifically, feasible interaction modes are first enumerated through conflict-point-based local negotiation analysis, and kinematically feasible anchors are then generated using graph-based spatiotemporal optimization. These anchors guide a controllable diffusion model to synthesize complete multi-agent trajectories, followed by rationality-based pruning to remove physically or interactively implausible samples (Jiang et al., 2025b). Detailed distributional consistency validation for the generated data is provided in Appendix A.

The results show that the generated samples remain within the physically reasonable range of real-world trajectories. This indicates that the augmented data can improve the coverage of typical failure modes without deviating from naturalistic driving behavior. Finally, 5200 valid interaction segments were extracted and constructed, including **Following**, **Crossing**, and **Merging** interactions, as shown in Fig. 2.

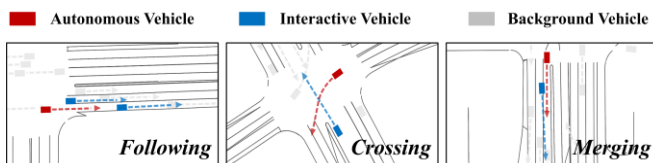


Fig. 2. Three interaction types in Waymo Open Dataset.

3.3. Evaluation dimensions and metrics

This section constructs a five-dimensional evaluation system, including safety, comfort, efficiency, social interaction, and surrounding traffic impact. The following will discuss the metrics for each dimension and quantification methods.

3.3.1 Safety

In this study, TTC is selected as one of the safety metrics, and the quantification method is described as follows:

$$TTC = \begin{cases} \frac{|\overline{DTC}|}{|\Delta \vec{v}|}, & \overline{DTC} \cdot \Delta \vec{v} < 0 \\ \infty, & \overline{DTC} \cdot \Delta \vec{v} \geq 0 \end{cases} \quad (1)$$

where $\overline{DTC}$ is the distance from the front to the rear vehicle, and $\Delta \vec{v}$ is the difference between two vehicles. Fig. 3 (a) shows how they are calculated. Draw a line parallel to $\Delta \vec{v}$, and if there exists at least one line segment that intersects both the front and rear edges of the leading and following vehicles, the shortest of these segments is taken as $\overline{DTC}$, with the direction pointing from the following to the leading vehicles.

In addition, Post Encroachment Time (PET) is chosen as another safety evaluation metric, as shown in Fig. 3 (b). it quantifies the driving safety of vehicles at intersection scenarios, which is defined as follows:

$$PET = \frac{d_B}{v_B} - \frac{d_A}{v_A} \quad (2)$$

where d_A is the displacement of vehicle A from the current time until its rear exits the conflict zone, and d_B is the displacement of vehicle B from the current time until its front enters the conflict zone. v_A and v_B are the instantaneous speeds of vehicle A and vehicle B at the current time.

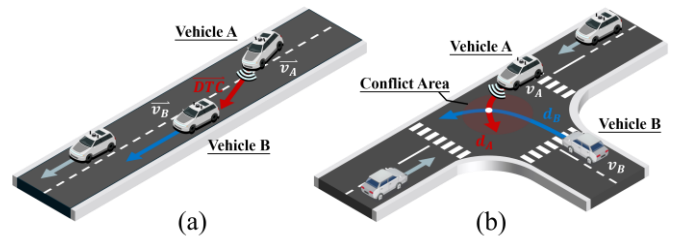


Fig. 3. (a) TTC in Merging scenario; (b) PET in Crossing scenario.

3.3.2 Comfort

Some studies (Winkel et al., 2023; Du et al., 2018; Chen et al., 2025) have examined vehicle comfort from various perspectives and statistically determined the threshold values of vehicle state parameters under different comfort levels. Based on these, this study selects longitudinal acceleration a_x , lateral acceleration a_y , Jerk, and yaw rate ω as comfort evaluation metrics, and these metrics are divided into multiple levels as shown in Table 1.

Table 1. Passenger comfort evaluation feedback interval

Metric	Comfortable	Acceptable	Uncomfortable
a_x	0-0.89	0.89-2.12	>2.12
a_y	0-0.40	0.40-1.80	>1.80
Jerk	0-0.60	0.60-1.00	>1.00
ω	0-0.10	0.10-0.37	>0.37

3.3.3 Efficiency

Task time and average delay are selected as efficiency evaluation metrics. Eq. (3) shows the calculation method of task time metric TT.

$$TT = \min\left(\frac{T_{\text{ego}} \cdot v_{\text{max}}}{d_{\text{min}}}, 1\right) \quad (3)$$

where T_{ego} is the testing time of ego vehicle, d_{min} is the length of the shortest compliant path to complete the task, and v_{max} is the maximum speed limit of the road.

Additionally, efficiency is represented by the average speed of traffic flow and then define the avg delay metric aD as Eq. (4).

$$aD = \min\left[\max\left(\frac{\bar{v}_{\text{ego}} - v_{\text{min}}}{v_{\text{max}} - v_{\text{min}}}, 0\right), 1\right] \quad (4)$$

where $\bar{v}_{\text{ego}}$ is the average speed of the ego vehicle, and v_{min} and v_{max} represent the minimum and maximum average speeds of the surrounding traffic vehicles, respectively

3.3.4 Social interaction

AV should make reasonable interaction decisions based on right-of-way and the states of other vehicles. This study extent metric IO from research (Liu et al., 2024) to general scenarios, evaluating the social interaction rationality of AVs, as follows:

$$IO = 1 - \text{ITSI} \cdot S_{\text{norm}} \quad (5)$$

$$\text{ITSI} = \text{softmax}[\Delta\text{TTCP}_{\text{norm}}, a_{\text{norm}}] \left[\frac{\Delta\text{TTCP}_{\text{norm}}}{a_{\text{norm}}} \right] \quad (6)$$

$$S_{\text{norm}} = \min\left[\max\left(\frac{\Delta S - S_{\text{min}}}{S_{\text{max}} - S_{\text{min}}}, 0\right), 1\right] \quad (7)$$

where ITSI is the Interaction Transient State Index, $\Delta\text{TTCP}_{\text{norm}}$ is the normalized time difference for the two vehicles to pass through the conflict zone, a_{norm} is the normalized cooperative acceleration of the two vehicles, S_{norm} is the normalized vehicle motion trajectory feature, ΔS is the longitudinal displacement generated by the ego vehicle at the current moment, and S_{min} and S_{max} are the minimum and maximum longitudinal displacements achievable under the collision-avoidance boundary acceleration. When IO approaches 0, it indicates that AV is adopting a more aggressive strategy.

3.3.5 Impact on surrounding traffic

In local interaction scenarios, unreasonable ADS behaviors may affect the normal movement of nearby traffic participants. This dimension evaluates the extent to which the ADS affects the driving

efficiency of surrounding vehicles during short-term interactions. This metric is quantified by the speed variation of affected background vehicles before and after AV interaction:

$$\text{Impact} = 1 - \frac{1}{n} \sum_{i=1}^n \max\left(\frac{v_{i,1} - v_{i,2}}{v_{i,1}}, 0\right) \quad (8)$$

where n is the number of Background Vehicles (BVs), and $v_{i,1}$ and $v_{i,2}$ represent the speeds of the i -th affected BV before and after the AV enters the traffic system, respectively.

Finally, the quantified results of all metrics need to be converted into normalized metric scores, ensuring that higher scores consistently indicate better performance. Therefore, appropriate transformations are applied to each metric. The detailed processing method for each metric is discussed in Appendix B.

3.4. LLM-based evaluation for ADS

LLMs are employed as expert-like evaluators to assess ADS intelligence in AV interaction processes, reducing human labor and improving evaluation consistency. This subsection introduces the LLM-based evaluation process used for the interaction segments in the dataset.

3.4.1 Data-Language Mapping Module

Interaction segments contain extensive trajectory and map data, while LLMs are not inherently sensitive to raw numerical inputs. Therefore, a Data-Language Mapping Module (DLMM) is developed to convert interaction data into structured natural-language descriptions. Specifically, DLMM adopts a coarse-to-fine processing strategy.

At the coarse-grained level, it parses the ego vehicle trajectory, surrounding-agent trajectories, lane topology, and interaction time window, and extracts key events according to the scenario type, such as car-following and lane-change initiation, conflict-zone entry and exit, and interaction completion. At the fine-grained level, it calculates trajectory-level and interaction-level features at these key moments, including position, speed, acceleration, yaw rate, relative speed, among others. Abnormal motions, such as sudden braking, using the anomaly function. Finally, the extracted events and features are filled into predefined rule-based templates to generate descriptions.

The coarse-grained information includes the scene type $\mathcal{S}$, interaction type $\mathcal{I}$, and key interaction time domains $\mathcal{T}$, which are extracted from the interaction process using LLM. The fine-grained information consists of vehicle states and environmental attributes. The state of each vehicle is:

$$\mathbf{x}(t) = [\mathbf{X}(t), \mathbf{s}(t)] \quad (9)$$

Where $\mathbf{X}(t) = [x(t), y(t)]$ represent the vehicle's position, and $\mathbf{s}(t) = [v(t), a(t), \theta(t)]$ represent velocity, acceleration and heading, respectively.

To simplify the data, non-interactive trajectory points are removed. For the three interaction scenarios **Following**, **Merging**, and **Crossing**, the effective interaction ranges are defined as follows:

$$\mathbf{R}_F = \{t | l_i(t) = l_j(t), |\Delta\theta_{ij}(t)| < \theta_F, d_{ij}(t) < d_F\} \quad (10)$$

$$\mathbf{R}_M = \{t | l_i(t) \rightarrow l_j(t), |\Delta\theta_{ij}(t)| > \theta_M, d_{ij}(t) < d_M\} \quad (11)$$

$$\mathbf{R}_C = \{t | \mathcal{T}_i \cap \mathcal{T}_j \neq \emptyset, |\Delta\theta_{ij}(t)| > \theta_C, d_{ij}(t) < d_C\} \quad (12)$$

where $l_i(t)$ denotes the lane where the vehicle i is located at time t , $\Delta\theta_{ij}(t)$ denotes the heading-angle difference between vehicles i and j at time t , and $\mathcal{T}_i$ denotes the trajectory sequence from current time t_0 to the end time t_{end} of vehicle i . θ_F , θ_M , and θ_C are the

heading-angle difference thresholds, while d_F , d_M , and d_C are the relative-distance thresholds for the three scenarios, respectively.

Additionally, for abnormal states encountered by AVs during interactions—such as sudden braking or sharp turns—an anomaly function $\mathcal{A}(t)$ is employed for monitoring.

$$\mathcal{A}(t) = \left\{ \left| \frac{s_i(t) - s_i(t-1)}{\Delta t} \right| > r_i, i = 1, 2, 3 \right\} \quad (13)$$

where $s_i(t)$ is the state component of AV, Δt is the time interval, and r_i is the anomaly threshold. To ensure the accuracy of LLM-based evaluation, we also structure and embed vehicle state data within the key interaction time domains T and anomaly $\mathcal{A}$ into the LLM.

Based on the above processing, the interaction trajectory data can be converted into prompts that are acceptable to the LLM. For a single interaction, Algorithm 1 demonstrates the operational process of DLMM.

Algorithm 1: Data–Language Mapping Module

Input: Interaction type I , vehicle states x , key time T , abnormal time set $\mathcal{A}$

Output: Structured natural-language description $\mathcal{D}$

```

1 Initialize interaction time  $t \leftarrow 0$ , and  $\mathcal{D} \leftarrow \emptyset$ ;
2 Set effective interaction range  $R \leftarrow R_I$ ;
3  $t_1, t_2, t_3 \leftarrow R[1], T[1], R[-1]$ ;
4 Append basic interaction information to  $\mathcal{D}$ , including  $S, I, R$ ,
  and involving agents;
5 Extract initial states of AV and HVs:  $x_{AV}(t_1)$  and  $x_{HV}(t_1)$ ;
6 Append initial-state description  $x_{AV}(t_1)$  and  $x_{HV}(t_1)$  to  $\mathcal{D}$ ;
7 Append AV status variation description  $s_{AV}(t_1 \rightarrow t_2)$  to  $\mathcal{D}$ ;
8 for each key time  $t \in T$  do
9   Calculate vehicle states  $s$  and interaction features at  $t$ ;
10  Append key behavior of interacting vehicles description
     $s_{AV}(t)$  and  $s_{HV}(t)$  to  $\mathcal{D}$ ;
11 end for
12 Append key interaction details description of  $x_{AV}(T)$  and
     $x_{HV}(T)$  to  $\mathcal{D}$ ;
13 Append AV status variation description  $s_{AV}(t_2 \rightarrow t_3)$  to  $\mathcal{D}$ ;
14 for each interaction time  $t \in R$  do
15   if ( $t \in \mathcal{A}$ ) then
16     Detect abnormal motion states  $\mathcal{A}(t)$ ;
17     Append AV abnormal status description  $\mathcal{A}(t)$  to  $\mathcal{D}$ ;
18   end if
19 end for
20 Organize  $\mathcal{D}$  using the predefined rule-based template
  corresponding to  $I$ ;
21 return  $\mathcal{D}$ .
```

3.4.2 Knowledge augmentation for LLMs evaluation

Although LLMs exhibit strong reasoning capabilities, their performance in ADS evaluation may be constrained by insufficient domain-specific knowledge and limited task-oriented adaptation. Prior research has shown that integrating LLMs with structured domain knowledge and retrieval-augmented mechanisms can improve the reliability of knowledge-intensive decision support (Liu et al., 2025). Motivated by this insight, this study adopts a knowledge-augmented (KA) strategy, where relevant AV knowledge, including traffic rules, driving behavior guidelines, and autonomous-driving knowledge, is retrieved and combined with DLMM-generated descriptions to guide domain-informed LLM evaluation.

Furthermore, the LLM is fine-tuned using high-quality interaction samples. These samples are converted into natural-language descriptions by DLMM and evaluated by human experts

through qualitative reasoning and fuzzy assessments rather than direct scoring. The expert reasoning process is then organized into a Chain-of-Thought format for fine-tuning, enabling the LLM to better align its evaluation logic with human expert judgment.

GPT-4o is selected as the base LLM for this study and accessed through the OpenAI API, owing to its strong language understanding and reasoning capabilities in complex evaluation tasks. The effectiveness of the proposed LLM processing modules and other LLMs is further analyzed in Section 4.1.2 to verify the robustness of the LLM-based evaluation strategy.

3.5. Evaluation model construction

In this study, a neural network model incorporating a self-attention mechanism is proposed. The model consists of a self-attention layer and a regression head. The former uses self-attention mechanism to weigh and integrate each input metric, calculating the interrelationships between the metrics, and the latter utilizes this comprehensive representation for evaluation prediction. The model architecture is shown in Fig. 4.

In this task, batch processing is used to improve the model's generalizability of the interrelationships mining between metrics. The batch size is defined as B . Larger B helps to reveal the implicit relationships between metrics more effectively, while an excessively large batch size will also increase the consumption. Assuming the total number of samples is N , the number of batches is N/B . Samples in i -th batch can be represented as:

$$\mathcal{D}_i = \{(\mathbf{x}_i^{(b)}, \mathbf{y}_i^{(b)})\}_{b=1}^B, \mathbf{x}_i^{(b)} \in \mathbb{R}^{10 \times 1}, \mathbf{y}_i^{(b)} \in \mathbb{R} \quad (14)$$

where $\mathbf{x}_i^{(b)}$ is the quantified result of the 10 metrics, and $\mathbf{y}_i^{(b)}$ is the intelligence evaluation label obtained from LLM. For each input $\mathbf{x}_i^{(b)}$, it is mapped to a d -dimension embedding matrix to facilitate subsequent processing by the self-attention layer:

$$\mathbf{t}_i^{(b)} = \mathbf{x}_i^{(b)} \mathbf{W}_d + \mathbf{b}_d \quad (15)$$

where $\mathbf{W}_d \in \mathbb{R}^{1 \times d}$ is the weight matrix, and $\mathbf{b}_d \in \mathbb{R}^d$ is the bias vector, which is broadcast to each sample. In the self-attention layer, the embedding matrix $\mathbf{t}_i^{(b)}$ is mapped to query, key, and value spaces:

$$\mathbf{Q}_i^{(b)} = \mathbf{t}_i^{(b)} \mathbf{W}_Q, \mathbf{K}_i^{(b)} = \mathbf{t}_i^{(b)} \mathbf{W}_K, \mathbf{V}_i^{(b)} = \mathbf{t}_i^{(b)} \mathbf{W}_V \quad (16)$$

where $\mathbf{W}_Q \in \mathbb{R}^{d \times d_K}$, $\mathbf{W}_K \in \mathbb{R}^{d \times d_K}$, and $\mathbf{W}_V \in \mathbb{R}^{d \times d_V}$ are learnable parameter matrices, d_K and d_V are the dimensions of the key and value spaces, respectively. Based on the similarity between each pair of queries and keys, the attention weight matrix in i -th batch is obtained:

$$\mathbf{A}_i = \frac{1}{B} \sum_{b=1}^B \mathbf{A}_i^{(b)} \quad (17)$$

$$\mathbf{A}_i^{(b)} = \text{softmax} \left(\frac{\mathbf{Q}_i^{(b)} (\mathbf{K}_i^{(b)})^T}{\sqrt{d_K}} \right) \quad (18)$$

where $\mathbf{A}_i^{(b)}$ is the attention weight matrix computed for the b -th sample, and $\mathbf{A}_i$ represents the attention levels between the metrics when computing the overall score. $\mathbf{A}_i$ is an intermediate output of the self-attention layer, while the final output is the attention score, which is obtained by multiplying the attention weight matrix with the value space matrix. Additionally, the attention score is linearly transformed to map it back to a scalar, facilitating subsequent processing by the regression head.

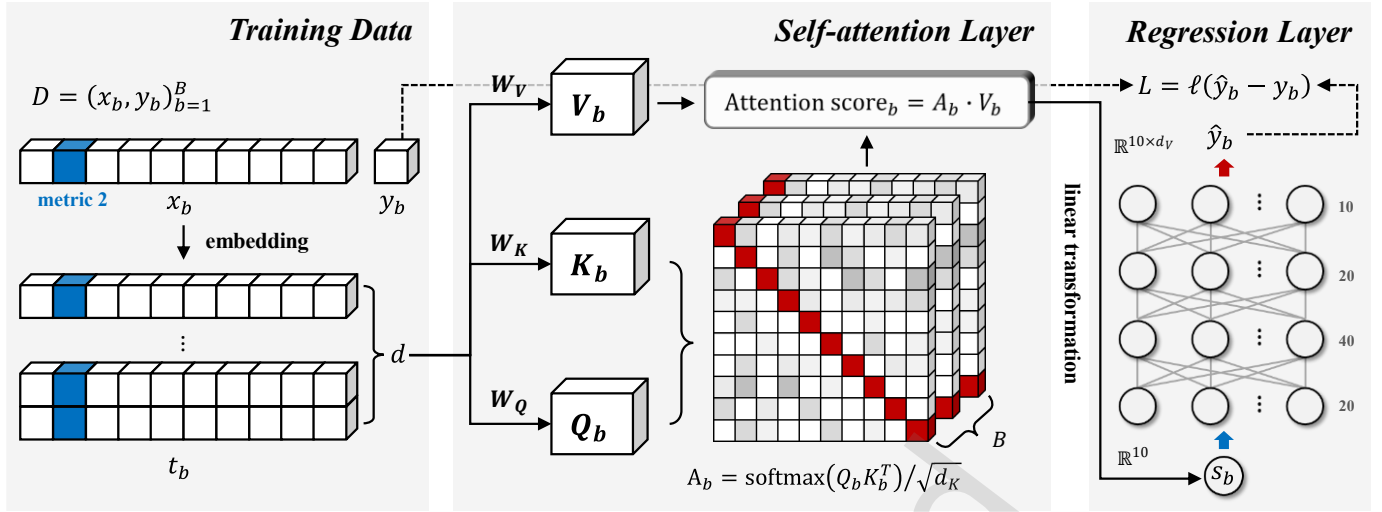


Fig. 4. Architecture of the proposed attention-based evaluation model.

$$\mathbf{s}_i^{(b)} = (\mathbf{A}_i^{(b)} \mathbf{V}_i^{(b)}) \mathbf{W}_s + \mathbf{b}_s \quad (19)$$

where $\mathbf{s}_i^{(b)}$ is the output of the self-attention layer, $\mathbf{W}_s \in \mathbb{R}^{d_v \times 1}$ is the weight matrix, and $\mathbf{b}_s \in \mathbb{R}^{d_v}$ is the bias vector. $\mathbf{s}_i^{(b)}$ is then fed into the regression head layer to achieve a nonlinear aggregation of all metrics. The regression head consists of a Multi-Layer Perceptron (MLP), where the input layer is composed of 10 neurons, each representing the result of each metric after weighted fusion through the self-attention mechanism. The output layer contains a single neuron, which represents the prediction of the final overall evaluation result. Although increasing the number of hidden layers can help the network learn more complex features, too many hidden layers could lead to overfitting (Uzair and Jamil, 2020). Ultimately, 4 hidden layers were selected. The number of neurons in each layer is set using an empirical formula:

$$N_h = \frac{N}{\alpha(N_i + N_o)} \quad (20)$$

where N_h is the number of neurons in each hidden layer, N_i and N_o are the number of neurons in the input and output layers, and α is a constant which is set to 10. The parameters in two layers are trained based on the evaluation label $y_i^{(b)}$, and the training objective is:

$$\min_{\theta} L(\theta) = \min_{\theta} \frac{1}{B} \sum_{b=1}^B (\hat{y}_i^{(b)}(\theta) - y_i^{(b)})^2 \quad (21)$$

$$\hat{y}_i^{(b)}(\theta) = \Phi(\mathbf{s}_i^{(b)}) \quad (22)$$

where θ is the model parameters, which are optimized by minimizing $L(\theta)$ and computing gradients using backpropagation, and Φ represents the nonlinear transformation performed by the regression head on the input $\mathbf{s}_i^{(b)}$. Based on the optimized model parameters, the final attention matrix between the metrics is computed:

$$\mathbf{A} = \frac{B}{N} \sum_{i=1}^N \mathbf{A}_i \quad (23)$$

This matrix represents the degree to which each metric attends to other metrics during the aggregation process and can be seen as a representation of the interrelationships between the metrics. The model's training performance results are further analyzed in Section 4.

4 Reliability and Effectiveness Validation

This section validates the reliability and effectiveness of the proposed evaluation framework. LLM-based evaluation results are first examined, followed by an analysis of the proposed evaluation model's performance and interpretability.

4.1. Reliability analysis of LLM-based evaluation

The accuracy of LLM-based evaluation is critical to the effectiveness of the evaluation model. Therefore, this section validates the reliability, stability, and expert consistency of LLM evaluation results through a Turing Test and ablation experiments.

4.1.1 Turing Test for expert-like evaluation

Ten PhD students specializing in the field of autonomous driving were invited as third-party evaluators. They were asked to assess the source of the evaluation results from 66 AV interaction evaluation samples, including whether the evaluation results came from human experts, from the LLM, or were indeterminable.

The Turing Test results are shown in Fig. 5, where $P(N | M)$ and $P(H | M)$ represent the ratio where the evaluator cannot determine the source and where the results are incorrectly judged to be from the LLM, respectively.

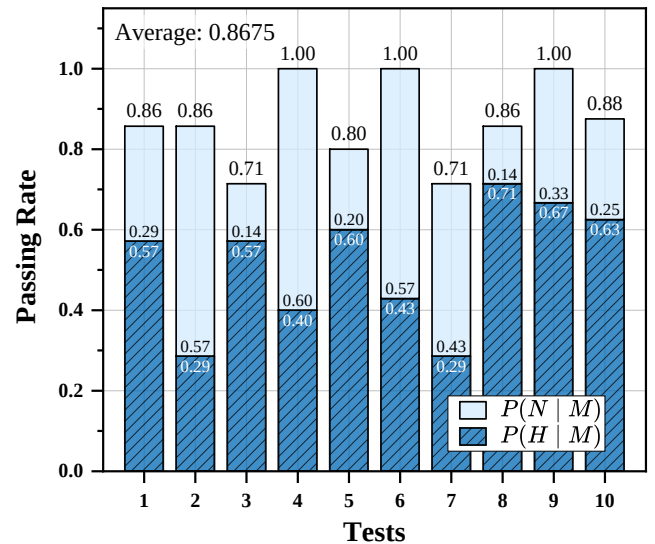


Fig. 5. Turing Test passing rates of LLM-based evaluation results.

Among them, the rate at which the evaluation results from the LLM were judged to be from human experts was 0.51, and the rate of unable to judge was 0.35, with an average passing rate of 86.75%. Additionally, based on Eq. (24), the 95% confidence interval for the test is (78.57%, 94.93%). This means that even though there may be some fluctuation in the true passing rate, we are 95% confident that the fluctuation range remains higher range, ensuring the reliability of our LLM-based evaluation data.

$$CI = p \pm z \sqrt{\frac{p(1-p)}{n}} \quad (24)$$

where p is the pass rate, n is the sample size, and z is the 95% confidence level.

4.1.2 Ablation analysis of LLM evaluation strategies

To evaluate the effectiveness of the proposed LLM evaluation strategy, ablation experiments were conducted on 600 representative interaction samples. Five LLM settings were compared, including raw GPT-4o, GPT-4o without DLMM, DeepSeek-V4 with KA, Gemini-2.5 with KA, and our proposed method. For each setting, three repeated evaluations were conducted under the same temperature and generation parameters. Human expert evaluation is used as reference.

Five indicators are selected from three perspectives. 1) MAE, RMSE, and Pearson correlation are used to measure the consistency between LLM evaluations and human expert judgments. 2) The Intraclass Correlation Coefficient (ICC) is adopted to evaluate the stability of repeated LLM evaluations under the same parameter settings. 3) Jensen-Shannon Divergence (JSD) is used to quantify the distributional consistency between LLM-generated scores and expert scores, indicating whether the overall evaluation tendency of the LLM is aligned with human experts.

As shown in Table 2, the proposed strategy achieves the best overall performance in terms of agreement with expert evaluations and distributional consistency. Compared with the raw GPT-4o, our method reduces MAE from 0.976 to 0.763 and RMSE from 1.326 to 1.078, while increasing the Pearson correlation from 0.804 to 0.863. This indicates that knowledge augmentation improves the alignment between LLM evaluations and expert judgments. When DLMM is removed and raw trajectory information is directly provided to the knowledge-augmented GPT-4o, the MAE and RMSE increase substantially to 1.399 and 1.823, respectively, demonstrating that DLMM is essential for converting trajectory data into interpretable descriptions for LLM-based evaluation. Although DeepSeek-v4 pro achieves the highest ICC of 0.920, its MAE and RMSE reach 1.566 and 2.030, suggesting that high repeated consistency alone does not guarantee expert-aligned evaluation. In contrast, our method maintains a high ICC of 0.901 and obtains the lowest JSD of 0.136, indicating both stable repeated evaluations and the closest score distribution to human experts. These results verify the effectiveness and necessity of the proposed LLM evaluation strategy.

4.2. Effectiveness analysis of the proposed evaluation model

4.2.1 Metric Interrelationship and Interpretability Analysis

This section further analyzes the effectiveness and interpretability of the proposed evaluation model. In applications, $d = 16$, with $d_k = 8$ and $d_v = 8$, neurons of the regression head are set to 20, 40, 20, and 10, batch size B is set to 256, and training epoch is set to 122. Fig. 6 shows the training results, along with the heatmap of the metric interrelationships obtained through the self-attention mechanism.

Table 2. Ablation results of LLM evaluation strategies

Model	Strategy		Consistency with expert evaluations			Distributional consistency	Evaluation stability
	KA	DLMM	MAE↓	RMSE↓	Pearson↑	JSD↓	ICC↑
GPT-4o (Ours)	√	√	0.763	1.078	0.863	0.136	0.901
GPT-4o	-	√	0.976	1.326	0.804	0.304	0.699
GPT-4o	√	-	1.385	1.765	0.634	0.161	0.752
DeepSeek-v4 pro	√	√	1.566	2.030	0.812	0.332	0.920
Gemini-2.5 pro	√	√	1.337	1.755	0.820	0.419	0.823

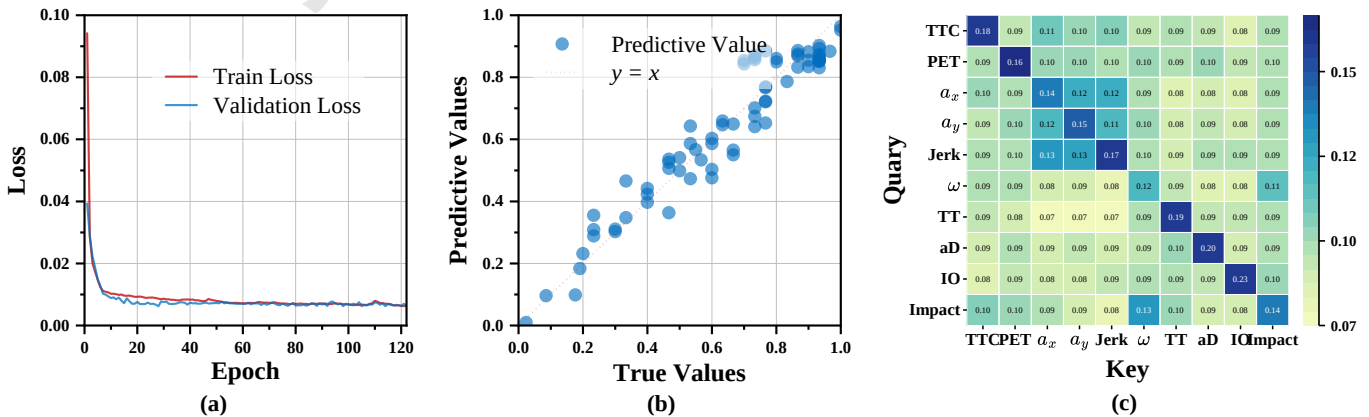


Fig. 6. Training performance and attention-based metric interrelationship analysis of the evaluation model. (a) Training and validation loss. (b) Predicted scores versus true scores. (c) Attention-based metric interrelationship matrix.

The attention matrix in Fig. 6 (c) quantifies how each metric attends to others during aggregation. The diagonal values are generally higher than most off-diagonal values, ranging from 0.147 to 0.232, indicating that each metric retains a dominant self-contribution. Among them, IO shows the highest self-attention value (0.232), followed by Impact (0.199), aD (0.198), and TT (0.187). Meanwhile, the non-negligible off-diagonal values indicate that the final intelligence score is not determined by isolated metrics, but by their coupled effects, supporting the necessity of the attention mechanism over fixed independent weights.

In addition to self-attention, clear interactions can be observed among comfort-related metrics. The attention values among a_x , a_y , and Jerk are relatively high, mostly around 0.114–0.130. This is consistent with vehicle motion characteristics, since Jerk directly reflects acceleration variation, while acceleration and yaw-related changes jointly affect motion smoothness. Safety-related metrics, such as TTC and PET, also interact with vehicle-motion metrics. TTC assigns relatively high attention to longitudinal acceleration a_x and Jerk, with attention values of 0.107 and 0.103, respectively. This is because acceleration and Jerk can directly change relative speed and following distance, thereby influencing the evolution of TTC. Efficiency and social-interaction metrics exhibit meaningful coupling. TT shows relatively high attention to aD and IO, with values of 0.089 and 0.090, respectively, while aD attends TT with a value of 0.098. This suggests that the model captures the relationship between overall task completion and local delay.

Furthermore, SHAP analysis was introduced as a post-hoc interpretability method to quantify the contribution of each metric to the final predicted intelligence score, and the result is shown in Fig. 7.

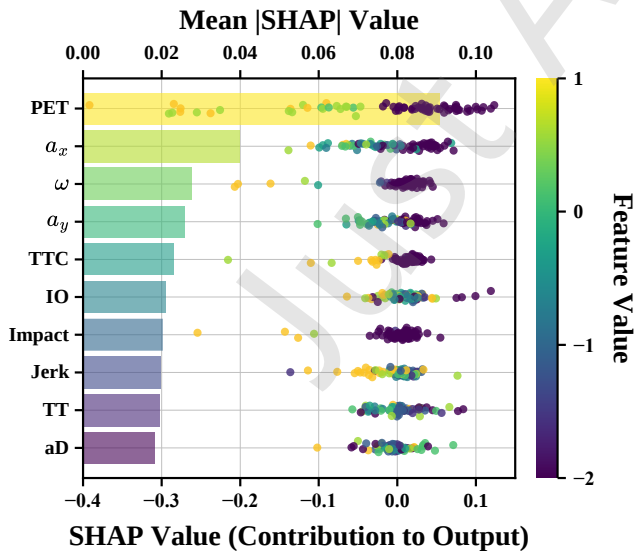


Fig. 7. SHAP-based feature contribution analysis of the proposed evaluation model.

Fig. 7 shows that PET exhibits the largest overall influence on the model output, followed by a_x , a_y , ω , and Impact, indicating that safety and vehicle-motion-related metrics play dominant roles in the final evaluation. For most metrics, higher normalized metric values generally correspond to positive SHAP values, while lower values tend to produce negative contributions. Low PET values

lead to strongly negative SHAP values, suggesting that conflict-zone safety is a critical factor that can significantly reduce the final intelligence score. Meanwhile, acceleration-related metrics and yaw rate also show notable contributions, confirming that the model considers motion stability and comfort as important evidence in evaluating ADS intelligence. These results are consistent with the attention-based analysis, further demonstrating that the proposed model can provide interpretable evaluation results.

4.2.2 Ablation analysis of the evaluation model

To validate the effectiveness of the evaluation model and the contribution of each dimension in the five-dimensional evaluation framework to the final intelligence assessment, ablation experiments were conducted in this section. Seven configurations were compared with the proposed model, including the model trained without augmented low-quality data and the models trained by removing one evaluation dimension at a time. MAE, RMSE, Pearson correlation, and R^2 were adopted for evaluation. Specifically, MAE and RMSE were used to quantify prediction errors, while Pearson correlation and R^2 were used to measure the consistency and explanatory ability of the predicted scores. The results are shown in Table 3.

As shown in Table 3, the model trained without augmented low-score data shows clear performance degradation compared with the proposed model. Specifically, MAE increases from 0.068 to 0.077, RMSE increases from 0.087 to 0.098, Pearson correlation decreases from 0.940 to 0.868, and R^2 decreases from 0.880 to 0.752. This indicates that the augmented low-score samples help the model learn a broader range of intelligence levels, especially low-intelligence behaviors that are underrepresented in naturalistic datasets. Therefore, introducing low-score data improves the robustness and generalization ability of the evaluation model.

Table 3. Ablation results of the proposed evaluation model

Model	MAE↓	RMSE↓	Pearson↑	R^2 ↑
Ours	0.068	0.087	0.940	0.880
w/o low-score data	0.077	0.098	0.868	0.752
w/o safety	0.106	0.147	0.810	0.652
w/o comfort	0.104	0.138	0.857	0.694
w/o efficiency	0.077	0.092	0.934	0.863
w/o interaction	0.073	0.087	0.941	0.878
w/o impact	0.077	0.095	0.931	0.854

The dimension-ablation results further demonstrate the necessity of the five-dimensional evaluation framework. Removing any dimension leads to a decline in model performance, indicating that safety, comfort, efficiency, social interaction, and surrounding traffic impact provide complementary information for comprehensive ADS intelligence evaluation. Among them, removing safety and comfort causes the most significant degradation, with MAE increasing 55.88% and 52.94%, RMSE increasing 68.97% and 58.62%, and R^2 decreasing 25.91% and 21.14%, respectively. This suggests that risk control and motion smoothness are particularly important for distinguishing ADS intelligence. Although removing efficiency, interaction, or impact

results in smaller changes, their performance remains lower than that of the complete model, confirming that all five dimensions contribute to the final intelligence prediction.

As shown in Fig. 8, the prediction distribution of the proposed model is highly consistent with the original score distribution, whereas the model trained without low-score data shows a clear distributional shift toward higher scores. Quantitatively, the original scores have a mean value of 0.65 and a standard deviation of 0.25, while the proposed model obtains almost identical statistics, with $\mu = 0.64$ and $\sigma = 0.25$. This indicates that our model can preserve both the central tendency and the dispersion of the original evaluation scores. In contrast, the model without low-score data produces a higher mean score of 0.71 and a much smaller standard deviation of 0.17, suggesting that its predictions are compressed into a narrower high-score range. This phenomenon indicates that, without low-score samples, the model tends to overestimate low-intelligence interactions and loses sensitivity to rare but critical failure behaviors. Therefore, the augmented low-score data are essential for improving the model's ability to distinguish poor ADS performance and for maintaining a realistic score distribution across different intelligence levels.

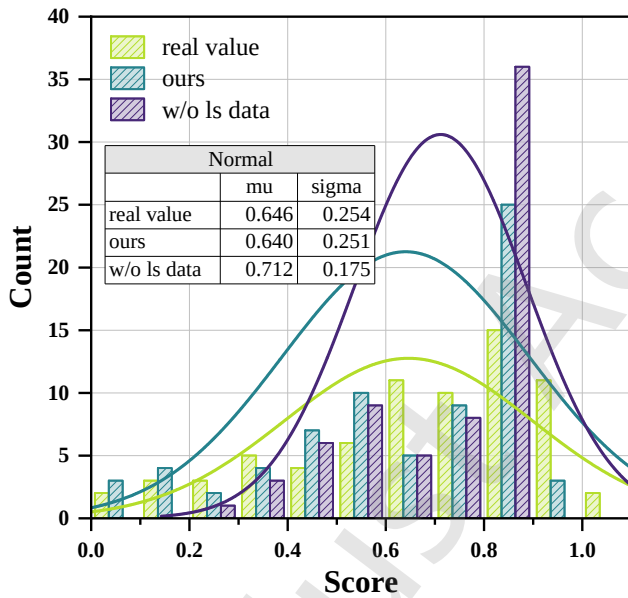


Fig. 8. Predicted score distributions of the proposed model and the model without low-score data.

5 Application and Analysis

To validate the application effectiveness of the proposed evaluation model, this study applies it to the AV simulation scenario-based testing to obtain evaluation results. These results are then compared with those from two traditional evaluation methods to assess the advantages of our model the in ADS intelligence evaluation.

5.1. Experiment settings

Five typical AV algorithms are selected from the OnSite Autonomous Driving Challenge Competition (Sun et al., 2024), each containing different decision-making, planning, and control algorithms. Table 4 presents the categories of these algorithms. Additionally, six traffic scenarios are selected, including ramp, highway following, and intersection straight and left driving.

The application scenarios used in the OnSite Autonomous

Driving Challenge are constructed based on the road structures and background-vehicle trajectories from the Waymo Open Dataset. Although the training data and application data are not identical in scenario organization and behavior distribution, they are related domains rather than independent or irrelevant datasets. Both contain typical traffic interactions, and both can be represented through the same high-level behavioral indicators extracted from vehicle interactions. In this sense, the OnSite experiments provide a related-domain validation of the proposed SOME framework.

Table 4. Decision-making and planning algorithms for AVs

Algorithm	Decision-making and planning	Control
A1	Integrated planning and control	
A2	Second-order programming	MPC
A3	Reinforcement learning	PID
A4	IDM model	PID
A5	Lattice planner	MPC

5.2. Comparative evaluation method

Two evaluation methods are selected for comparison: subjective evaluation by human experts and AHP-based evaluation. To ensure the reliability of the comparison evaluation results, three experts from China Automotive Engineering Research Institute and an associate professor from Tongji University in autonomous driving domain are invited to conduct Expert Evaluation and AHP metric weighting.

In the AHP metric weighting, four experts are required to assess the importance of the five evaluation dimensions and metrics, and judgment matrices are constructed based on these assessments:

$$\mathbf{A} = (a_{ij})_{n \times n} \quad (25)$$

$$\omega'_i = \frac{\omega_i}{\sum_{i=1}^n \omega_i} \quad (26)$$

where $\mathbf{A}$ is the judgment matrix, and its element a_{ij} represents the importance of metric i to metric j , satisfying $a_{ij} \cdot a_{ji} = 1$. n is the number of metrics, and ω_i and ω'_i are the eigenvector and the weight of metric i , respectively. Santy 1-9 Scale Method is used in importance assessment, and the results are shown in Table 5.

Table 5. Dimension and metric AHP weighting results

Dimension	Weight	Metric	Weight
Safety	0.52	TTC	0.69
		PET	0.31
Comfort	0.20	a_x	0.29
		a_y	0.28
		Jerk	0.22
		ω	0.20
Efficiency	0.13	TT	0.76
		aD	0.24
Social interaction	0.09	IO	1.00
Surrounding traffic impact	0.06	Impact	1.00

In the Expert Evaluation, testing progress is presented to four experts in video format, supplemented with test data and visual

results. Experts are required to evaluate the intelligence of each algorithm in different scenarios on a 0-10 scale and provide justifications.

5.3. Results analysis

Table 6 presents the evaluation results of AV algorithms by our evaluation model and the comparison evaluation methods, and Fig. 9 displays the evaluation results from the three evaluation methods. The specific result for all metrics is presented in Appendix C.

The algorithms were evaluated using the model proposed in this study. As shown in Table 6, the scores of most algorithms are concentrated around 0.8, but there are still cases with both higher and lower scores. For example, Algorithm 1 performs better in Scenario 4 and Scenario 5, while Algorithm 2 performs relatively worse in Scenario 3 and Scenario 5. This indicates that our evaluation model can effectively distinguish significant

performance differences among different algorithms across multiple scenarios.

By comparing it with the rankings from Expert Evaluation, it can be observed that the overall ranking trends are consistent with the judgments of human experts. The rankings of Algorithms 1 to 3 remain the same, indicating that the evaluation model has effectively learned the judgment methods and logic of human experts, providing results close to those of human experts.

However, a detailed analysis of the discrepancies between the evaluation model's results and human expert assessments reveals some insightful observations. Specifically, our evaluation model considers Algorithm 4 to have a higher intelligence level, with scores consistently around 0.8 across five scenarios. However, three experts assigned low scores to Algorithm 4 in Scenarios 4 and 6, with scores of 5, 3, 5 and 4.5, 5, 5, respectively.

Table 6. Comparison of results from three evaluation methods

Method	Subjective-Objective Mapping Evaluation Model								Expert Evaluation	AHP
Scenario	S1	S2	S3	S4	S5	S6	Score	Ranking	Ranking	Ranking
A1	0.82	0.85	0.88	0.91	0.92	0.71	0.85	1	1	2
A2	0.81	0.85	0.56	0.75	0.54	0.84	0.73	5	5	4
A3	0.91	0.82	0.87	0.83	0.77	0.82	0.84	2	2	1
A4	0.84	0.79	0.83	0.82	0.81	0.75	0.80	3	4	3
A5	0.72	0.84	0.68	0.83	0.81	0.82	0.78	4	3	5

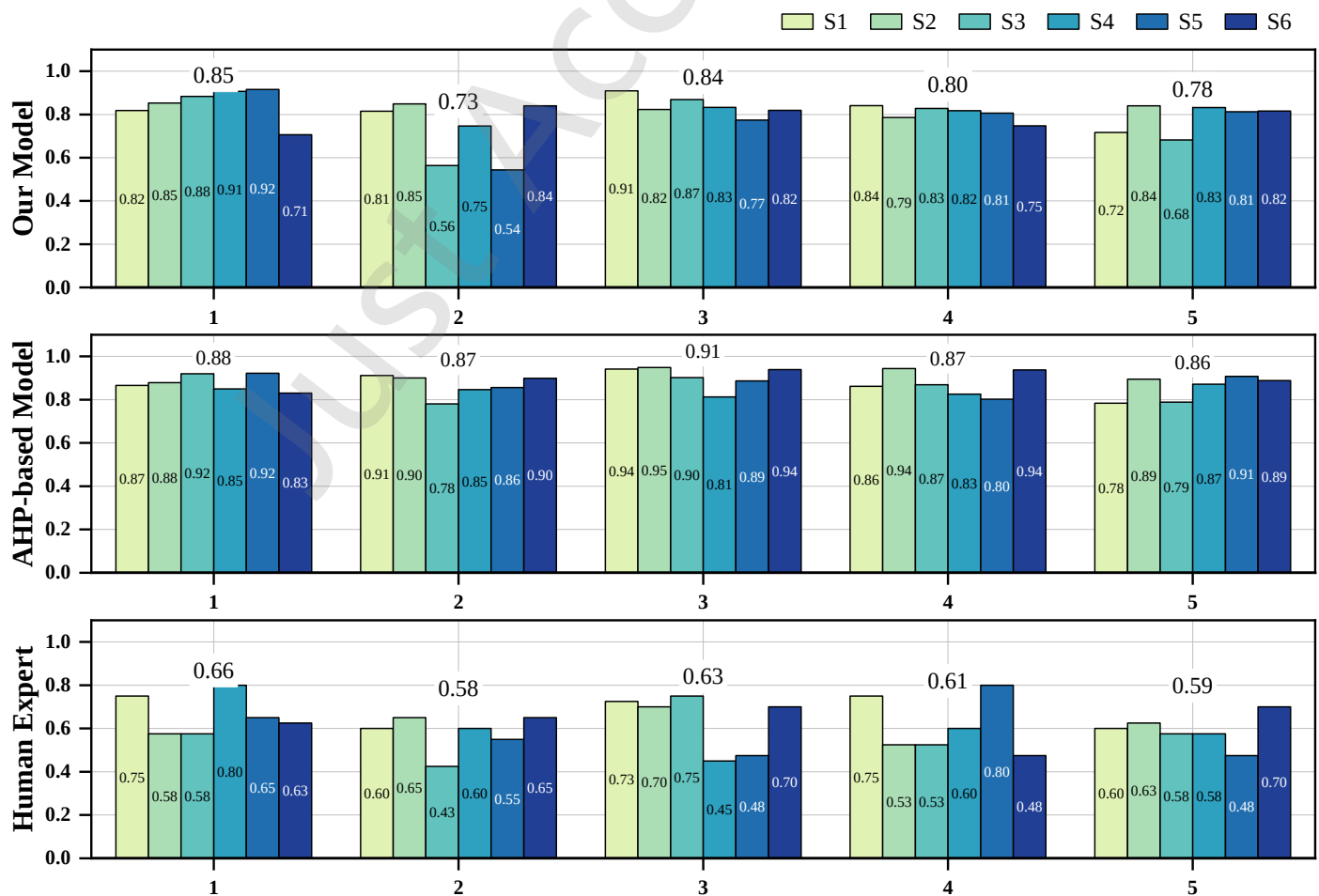


Fig. 9. Comparison of evaluation results among the proposed model, AHP-based model, and human experts.

Fig. 10 illustrates the key moments of Algorithm 4 in these two scenarios. As shown in the figure, in Scenario 4, the vehicle maintains a low speed for over 9 seconds at the intersection approach. Experts argue that this may reduce the overall efficiency of the vehicle and potentially interfere with surrounding and following traffic. However, our model considers the overall speed of the background traffic, and from this perspective, the vehicle does not exhibit a significant efficiency issue. Moreover, intentionally reducing speed at the intersection is seen as a safety measure, ensuring safer navigation through the intersection. In Scenario 6, the vehicle's decision to temporarily leave the lane boundary to optimize efficiency is an interesting aspect. While this

maneuver may introduce a potential risk when the vehicle re-enters the lane, the decision to exceed the lane boundary could be interpreted as a strategy to improve the overall flow and minimize delays, leading to a higher efficiency score.

These observations indicate that while the evaluation model offers a robust framework for assessing autonomous vehicle behaviors, the model's sensitivity to specific, scenario-dependent factors—such as waiting times and lane boundary considerations—may lead to varying interpretations of performance, which also underscores the importance of continuously refining the model to better capture nuanced aspects.

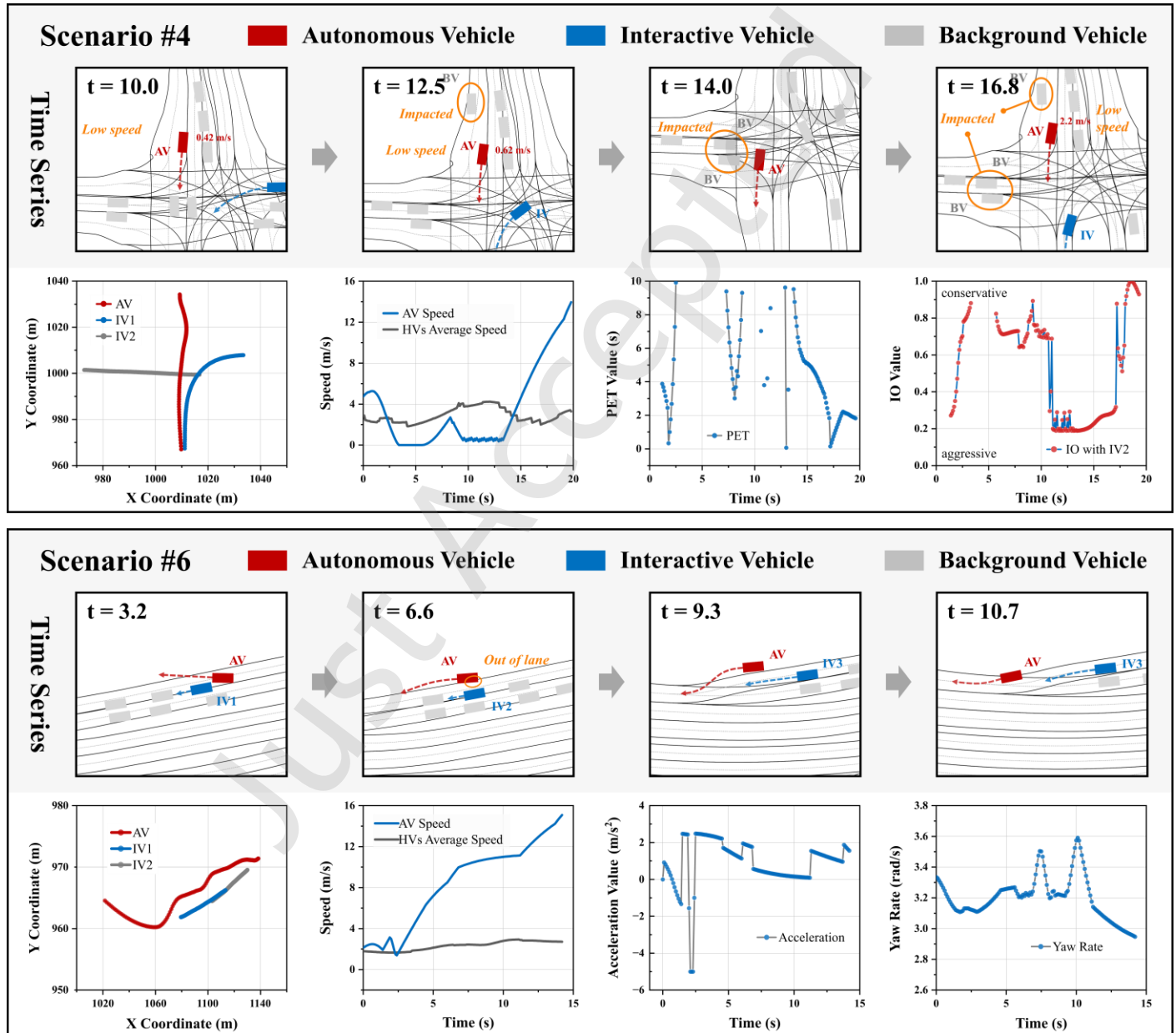


Fig. 10. Key interaction moments and metric variations of Algorithm 4 in Scenarios 4 and 6.

Furthermore, by comparing with AHP evaluation method, notable differences in rankings can be observed. Specifically, the rankings of Algorithms 1 and 3 are swapped, the same as Algorithms 2 and 5. Taking the first two as examples, Table 7 presents the evaluation details of them. For Algorithm 1, the score

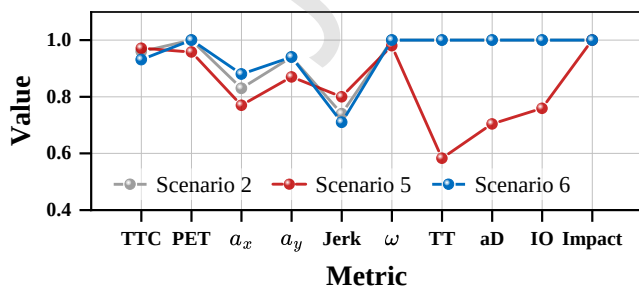
differences between the two evaluation methods are relatively small, with only Scenario 6 showing a large discrepancy. However, Algorithm 3 achieves significantly higher scores than our evaluation model in Scenarios 2, 5 and 6 using the AHP evaluation method, with differences exceeding 0.13 at maximum.

Table 7. A1 and A3 evaluation scores from our model and AHP

Algorithm	Method	S1	S2	S3	S4	S5	S6
A1	Ours	0.82	0.85	0.88	0.91	0.92	0.71
	AHP	0.87	0.88	0.92	0.85	0.92	0.83
A3	Ours	0.91	0.82	0.87	0.83	0.77	0.82
	AHP	0.94	0.95	0.90	0.81	0.89	0.94

Fig. 11 shows the quantified metrics of Algorithm 3 in these three scenarios. It can be observed that both TTC and PET are close to 1.0, indicating strong safety performance, whereas three of the efficiency metrics fall within the poor range. The high weight assigned to the safety dimension in the AHP results (0.52) leads to a high overall score, while the influence of the other dimensions on intelligence is largely overlooked. In particular, the TT and aD scores in Scenario 5 are only 0.58 and 0.70, and the IO metric score is 0.76, suggesting clear deficiencies in the vehicle's efficiency and decision-making. However, because the weights for these two dimensions are low (0.13 and 0.09, respectively), these issues are overshadowed. This demonstrates that fixed weights may cause the evaluation to deviate from actual performance, and it remains challenging to determine a set of weights that truly reflect real behavior.

In addition, Fig. 11 shows that Algorithm 3 exhibits generally poor performance in both longitudinal acceleration, lateral acceleration, and lateral speed, indicating frequent exceedance of the corresponding thresholds. As is well known, vehicle acceleration and Jerk not only directly affect its ride comfort but also indirectly impact safety. Sudden acceleration may lead to rear-end collisions or other latent safety hazards, while excessive lateral speed or acceleration may cause lateral collisions and skidding. However, in AHP evaluation method, a linear weighting approach cannot incorporate these metrics into the safety dimension and essentially ignores the relationships among metrics. In contrast, the evaluation model proposed in this study can capture such nuances. Consequently, during the evaluation, it avoids assigning excessively high overall intelligence scores to ADS solely due to their high safety metrics, which is precisely the key advantage of our model.

**Fig. 11.** Metrics' score of Algorithm 3 in Scenarios 2, 5, and 6.

6 Conclusion

The limitations of existing ADS evaluation methods make it difficult to accurately assess the comprehensive intelligence of high-level autonomous vehicles in complex interaction scenarios. To address this issue, this paper proposes a Subjective-Objective

Mapping Evaluation (SOME) framework, which integrates LLM-based expert-like evaluation with data-driven neural network modeling. In this framework, ADS intelligence is represented through five complementary behavioral dimensions, including safety, comfort, efficiency, social interaction, and impact on surrounding traffic.

The reliability of the LLM-based evaluation results is first validated. The Turing Test achieves a passing rate of 86.75%, indicating that the LLM-generated evaluations are highly consistent with human expert reasoning. The ablation experiments further show that the proposed LLM evaluation strategy achieves better agreement with expert evaluations and the closest score distribution to human experts, with MAE = 0.763, RMSE = 1.078, Pearson = 0.863, ICC = 0.901, and JSD = 0.136. These results demonstrate the necessity of DLMM and knowledge augmentation in improving LLM evaluation reliability.

Furthermore, an attention-based neural network is developed to learn the nonlinear mapping between objective metric scores and LLM-generated intelligence scores. The interpretability analysis shows that the self-attention mechanism can reveal meaningful interrelationships among evaluation metrics, while SHAP analysis further quantifies the contribution of each metric to the final prediction. The model ablation results show that the proposed evaluation model achieves the best performance. Removing low-score data or any evaluation dimension leads to performance degradation, confirming the importance of low-score sample augmentation and the necessity of the five-dimensional evaluation framework.

Finally, the proposed model is applied to scenario-based testing using algorithms and scenarios from the OnSite Autonomous Driving Challenge. The evaluation results are generally consistent with human expert rankings and provide more reasonable assessments than the AHP-based method in cases where fixed weights overlook metric interactions.

The current evaluation framework mainly focuses on trajectory-level interaction data and high-level behavioral metrics, while richer multimodal information, such as video, perception outputs, and more detailed environmental context, has not yet been fully incorporated. In addition, although the OnSite application provides preliminary cross-domain evidence, broader validation across more datasets, regions, traffic cultures, and testing platforms is still needed. Future work will further improve the generalization ability, interpretability, and practical deployment value of the proposed evaluation framework.

Appendix A

This section discusses the generation and rationality validation of the low-score interaction data used for model training. Based on the method proposed in our team's previous study, typical interaction data from the baseline dataset used in this study are expanded to supplement the low-score distribution of the training dataset. The purpose of this augmentation is not to randomly create unrealistic "bad cases", but to generate physically feasible and interaction-consistent counterfactual trajectories that cover rare but critical low-intelligence behaviors.

The scenario generation process has been briefly introduced in

Section 3.2, while the detailed methodology can be found in Jiang et al. (2025b). Based on this method, this study conducts a dedicated expansion of the Waymo Open Dataset. While preserving the physical road structure, traffic participant types, and behavioral characteristics of the original scenes, low-score AV interaction behaviors that satisfy vehicle dynamic constraints and traffic rules are generated. To ensure that the generated data can capture typical failure modes in a reliable manner, the distributional consistency between the generated data and the real-world dataset is further validated.

A consistency analysis was conducted on the distributions of key ego-vehicle features in the original data and the generated low-score data, as shown in Fig. B2. The results show that the generated data of acceleration, jerk, and yaw rate largely overlap with the original data and show similar central tendencies and dispersion ranges. This indicates that the generated trajectories do not introduce abnormal vehicle dynamics or unrealistic motion fluctuations. The speed distribution of the generated data remains within a similar range to the original data, although slight deviations can be observed, which is reasonable because low-score interaction generation may introduce more conservative or risky speed-control behaviors. The generated low-score data preserves the basic kinematic characteristics of naturalistic driving while enriching interaction behaviors with lower intelligence scores.

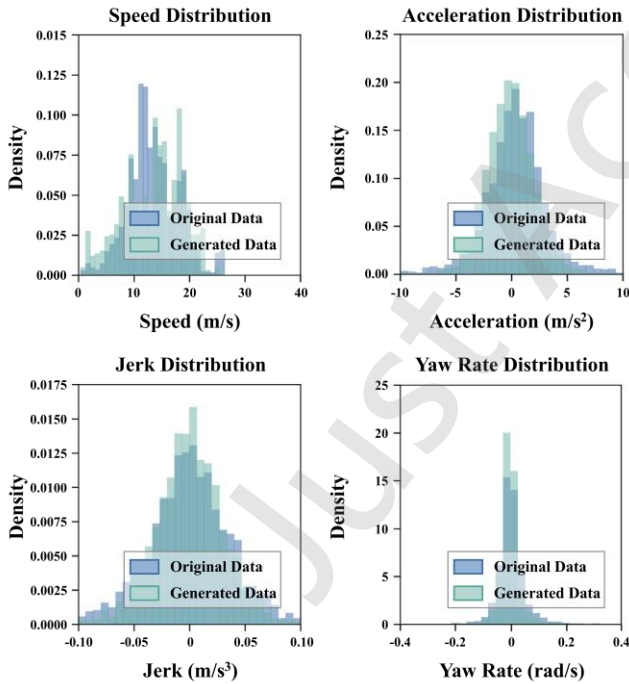


Fig. B2. Distribution between original and generated data.

Finally, Fig. B3 presents the distributions of LLM-based intelligence scores for the original data and the generated low-score data. Compared with the original data, whose scores are mainly concentrated in the high-score range, the generated data show a clear extension toward the lower-score region. This indicates that the generated samples provide additional low-intelligence interaction cases that are underrepresented in the original dataset. Therefore, the augmented data help balance the training score distribution and improve the model's ability to learn typical failure modes.

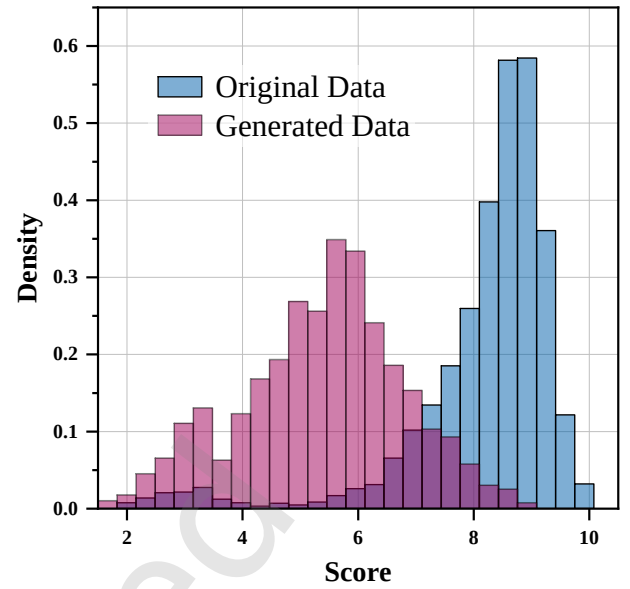


Fig. B3. Distribution comparison of LLM-based intelligence scores between original and generated data.

Appendix B

This section discusses how the quantified metrics are standardized into metric scores and mapped to the [0,1] interval, where higher scores indicate better performance. The metrics requiring standardization include safety, comfort, and social interaction metrics, while the quantified results of the remaining metrics can be directly used as metric scores.

For safety related metrics, continuous risk potential functions are constructed to standardize TTC and PET. The detailed processing methods are as follows:

$$S_s = 1 - \frac{1}{T} \sum_{t=1}^T e^{-\frac{m(t)}{\tau}} \quad (27)$$

where S_s denotes the standardized safety metric scores, $m(t)$ represents the quantified metric value at time t , τ is the metric threshold, and T is the total duration of the interaction. Both TTC and PET are standardized using the above formulation. In this study, the safety thresholds of TTC and PET are set to 3.0 and 2.0 seconds, respectively. This method preserves the continuous risk information of TTC and PET over time, avoids the information loss caused by traditional hard-threshold methods, and provides a smoother representation of the safety level during interaction.

For comfort related metrics, standardization is performed based on the thresholds listed in Table 1. The detailed processing method is as follows:

$$S_c = 1 - \frac{1}{T} \sum_{t=1}^T c(t) \quad (28)$$

$$c(t) = \begin{cases} 1, & 0 \leq m(t) \leq m_c \\ 1 - k[m(t) - m_c], & m_c < m(t) \leq m_u \\ c_u \exp\left(-\frac{m(t) - m_u}{m_u - m_c}\right), & m(t) > m_u \end{cases} \quad (29)$$

where k is the linear decay slope, representing the sensitivity of the comfort metric. A larger k indicates that the model is more sensitive to this metric, meaning that an increase in the metric value within the acceptable range leads to a faster decrease in the comfort score. c_u ensures the continuity of the function at m_u , and is calculated as

$c_u = 1 - k(m_u - m_c)$. m_c and m_u denote the thresholds of the comfortable and uncomfortable ranges for the corresponding comfort metric, respectively. This formulation preserves the continuous variation of comfort-related indicators, avoids abrupt changes caused by hard thresholding, and imposes stronger penalties on severe discomfort.

For social interaction related metric, the legal right-of-way is used to standardize IO. Based on the InterHub toolkit, the right-of-way of the AV and HV in the interaction is annotated, and the rationality of the AV's social interaction decision is then determined accordingly. The detailed processing method is as follows:

$$S_i = \frac{1}{T} \sum_{t=1}^T [1 - (i(t) - p(t))]^2 \quad (30)$$

$$p(t) = \begin{cases} 0, & \text{AV has right} \\ 1, & \text{HV has right} \\ 0.5, & \text{equal} \end{cases} \quad (31)$$

where $i(t)$ represents the quantified IO value at time t , and $p(t)$

denotes the right-of-way label at time t . The term $[i(t) - p(t)]^2$ measures the deviation between the AV's actual interaction tendency and the expected decision under the legal right-of-way relationship. Therefore, a smaller deviation indicates a more reasonable social interaction decision, resulting in a higher standardized score.

This formulation incorporates legal right-of-way constraints into the standardization of IO and transforms interaction rationality into a continuous score. Compared with simple rule-based classification, it can quantify the degree of mismatch between AV behavior and expected interaction decisions, thereby providing a more informative evaluation of social interaction performance.

Appendix C

Appendix C presents the metric scores achieved by the five autonomous driving algorithms across the same six test scenarios, as shown in Table C1.

Table C1. 5 algorithms' metrics scores in 6 test scenarios.

Algorithm	Scenario	Metric score									
		TTC	PET	a_x	a_y	Jerk	ω	TT	aD	IO	Impact
A1	S1	0.95	0.69	0.92	0.92	0.89	1.00	0.87	0.56	0.44	0.87
	S2	0.92	0.95	1.00	0.98	1.00	0.91	0.35	0.80	0.91	1.00
	S3	1.00	1.00	0.86	0.94	1.00	0.84	0.70	0.24	1.00	1.00
	S4	0.85	0.94	0.99	1.00	1.00	1.00	0.50	0.42	1.00	0.86
	S5	0.94	0.94	0.86	0.91	1.00	0.98	0.80	1.00	0.58	0.92
	S6	0.96	0.96	0.68	0.88	0.91	0.80	0.28	0.36	0.85	1.00
A2	S1	1.00	0.91	0.65	0.93	0.74	0.94	1.00	0.77	0.42	0.83
	S2	1.00	0.91	0.95	0.97	0.96	1.00	0.36	0.84	0.91	1.00
	S3	0.95	1.00	0.25	0.26	0.08	1.00	0.72	0.31	1.00	1.00
	S4	0.97	0.91	0.68	0.99	0.73	0.95	0.54	0.46	1.00	0.85
	S5	0.96	0.93	0.44	0.94	0.19	0.89	1.00	1.00	0.93	0.86
	S6	1.00	0.98	0.99	1.00	0.97	1.00	0.31	0.52	0.95	1.00
A3	S1	1.00	0.94	0.93	0.97	0.91	1.00	0.93	0.60	0.01	0.87
	S2	0.96	1.00	0.83	0.94	0.74	1.00	1.00	1.00	1.00	1.00
	S3	1.00	0.75	0.99	0.93	1.00	1.00	0.74	0.42	1.00	1.00
	S4	0.81	0.96	0.92	0.73	0.94	1.00	0.46	0.38	1.00	0.92
	S5	0.97	0.96	0.77	0.87	0.80	0.98	0.58	0.70	0.76	1.00
	S6	0.93	1.00	0.88	0.94	0.71	1.00	1.00	1.00	1.00	1.00
A4	S1	1.00	0.94	0.94	0.68	0.92	1.00	0.43	0.20	1.00	0.88
	S2	1.00	1.00	0.76	0.90	0.68	0.95	0.94	1.00	1.00	1.00
	S3	1.00	0.75	0.97	0.40	1.00	1.00	0.75	0.45	1.00	1.00
	S4	0.86	0.98	0.95	0.80	0.93	1.00	0.39	0.32	1.00	0.93
	S5	0.69	0.97	0.97	0.84	0.84	1.00	0.61	0.75	0.85	0.94
	S6	1.00	1.00	0.77	0.94	0.58	0.91	0.93	1.00	1.00	1.00
A5	S1	0.74	0.81	0.75	0.80	0.72	0.93	0.92	0.60	0.85	0.87
	S2	0.98	1.00	0.99	0.71	0.97	1.00	0.36	0.84	0.96	1.00
	S3	1.00	0.37	0.97	0.42	0.99	1.00	0.70	0.23	1.00	1.00
	S4	1.00	0.95	0.91	0.63	0.89	1.00	0.50	0.42	0.96	0.89
	S5	1.00	0.93	0.78	0.72	0.72	0.98	0.80	1.00	1.00	0.92
	S6	1.00	0.99	1.00	0.82	0.98	1.00	0.30	0.47	0.91	1.00

Author contributions

Jiarui Zhang: Writing - original draft, Visualization, Validation, Methodology, Formal analysis, Software. **Jierui Chen:** Data Curation, Resources. **Shiyu Fang:** Writing - Review Editing. **Chao Huang:** Writing - Review Editing. **Peng Hang:** Writing - Review Editing, Conceptualization of this study, Supervision. **Jian Sun:** Supervision, Project administration.

Replication and data sharing

The data and codes used in this study are available at <https://doi.org/10.26599/ETSD.2026.9190028>

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (52472451), the Shanghai Scientific Innovation

Foundation (No.23DZ1203400), and the Fundamental Research Funds for the Central Universities.

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.

References

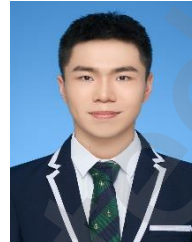
- Chen, D., Ahn, S., Chitturi, M., & Noyce, D. A. (2017). Towards vehicle automation: Roadway capacity formulation for traffic mixed with regular and automated vehicles. *Transportation research part B: methodological*, 100, 196-221.
- Chen, Z., Liu, Y., Ni, W., Hai, H., Huang, C., Xu, B., ... & Li, J. (2025). Predicting driving comfort in autonomous vehicles using road information and multi-head attention models. *Nature Communications*, 16(1), 2709.
- Darko, A., Chan, A. P. C., Ameyaw, E. E., Owusu, E. K., Pärn, E., & Edwards, D. J. (2019). Review of application of analytic hierarchy process (AHP) in construction. *International journal of construction management*, 19(5), 436-452.
- De Winkel, K. N., Irmak, T., Happee, R., & Shyrokau, B. (2023). Standards for passenger comfort in automated vehicles: Acceleration and jerk. *Applied Ergonomics*, 106, 103881.
- Du, F., Zhao, Y., Gao, L., & Wang, W. (2017, August). Evaluation of the autonomous parking system based on BP neural network. In *2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)* (Vol. 2, pp. 324-327). IEEE.
- Du, Y., Liu, C., & Li, Y. (2018). Velocity control strategies to improve automated vehicle driving comfort. *IEEE Intelligent transportation systems magazine*, 10(1), 8-18.
- Fang, S., Liu, J., Xu, C., Lv, C., Hang, P., & Sun, J. (2025). Interact, instruct to improve: A llm-driven parallel actor-reasoner framework for enhancing autonomous vehicle interactions. *IEEE Transactions on Intelligent Transportation Systems*.
- Grosse, K., & Alahi, A. (2024). A qualitative AI security risk assessment of autonomous vehicles. *Transportation Research Part C: Emerging Technologies*, 169, 104797.
- Guo, J., Kurup, U., & Shah, M. (2019). Is it safe to drive? An overview of factors, metrics, and datasets for driveability assessment in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 21(8), 3135-3151.
- Hu, L., Li, H., Yi, P., Huang, J., Lin, M., & Wang, H. (2022). Investigation on AEB key parameters for improving car to two-wheeler collision safety using in-depth traffic accident data. *IEEE Transactions on Vehicular Technology*, 72(1), 113-124.
- Jahangirova, G., Stocco, A., & Tonella, P. (2021, April). Quality metrics and oracles for autonomous vehicles testing. In *2021 14th IEEE conference on software testing, verification and validation (ICST)* (pp. 194-204). IEEE.
- Jiang, X., Zhao, X., Liu, Y., Li, Z., Hang, P., Xiong, L., & Sun, J. (2025). A naturalistic trajectory dataset with dense interaction for autonomous driving. *Scientific Data*, 12(1), 1084.
- Jiang, X., Zhao, X., Liu, Y., Hang, P., & Sun, J. (2025, November). Every Scene All at Once: Exhaustive Multi-Agent Interaction Generation with Controlled Diffusion Model. In *2025 IEEE 28th International Conference on Intelligent Transportation Systems (ITSC)* (pp. 1-8). IEEE.
- Johansson, M., Ekman, F., Karlsson, M., Strömberg, H., Jonsson, J., & Faleke, M. (2023). Automation as an enabler: passengers' experience of travelling with a full-length automated bus and their expectations of a future public transport system. *Transportation Research Procedia*, 72, 957-964.
- Lee, D., Lee, S., Chen, Z., Park, B. B., & Shim, D. H. (2021). Design and field evaluation of cooperative adaptive cruise control with unconnected vehicle in the loop. *Transportation Research Part C: Emerging Technologies*, 132, 103364.
- Li, S., Zhou, R., & Huang, H. (2025). Multidimensional Evaluation of Autonomous Driving Test Scenarios Based on AHP-EWN-TOPSIS Models: S. Li et al. *Automotive Innovation*, 8(2), 237-251.
- Li, J., & Kwan, R. S. (2003). A fuzzy genetic algorithm for driver scheduling. *European Journal of Operational Research*, 147(2), 334-344.
- Liu, J., Qi, X., Hang, P., & Sun, J. (2024). Enhancing social decision-making of autonomous vehicles: A mixed-strategy game approach with interaction orientation identification. *IEEE Transactions on Vehicular Technology*, 73(9), 12385-12398.
- Liu, Y., Wu, F., Liu, Z., Wang, K., Wang, F., & Qu, X. (2023). Can language models be used for real-world urban-delivery route optimization?. *The Innovation*, 4(6).
- Liu, Y., Zhou, Y., Liu, Y., Xu, Z., & He, Y. (2025). Intelligent fault diagnosis for CNC through the integration of large language models and domain knowledge graphs. *Engineering*.
- Mao, J., Qian, Y., Ye, J., Zhao, H., & Wang, Y. (2023). Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*.
- Montanaro, U., Dixit, S., Fallah, S., Dianati, M., Stevens, A., Oxtoby, D., & Mouzakitis, A. (2019). Towards connected autonomous driving: review of use-cases. *Vehicle system dynamics*, 57(6), 779-814.
- Nie, T., Sun, J., & Ma, W. (2025). Exploring the roles of large language models in reshaping transportation systems: A survey, framework, and roadmap. *Artificial Intelligence for Transportation*, 1, 100003.
- Qu, X., Lin, H., & Liu, Y. (2023). Envisioning the future of transportation: Inspiration of ChatGPT and large models. *Communications in Transportation Research*, 3, 100103.
- Sun, B., Zhang, Q., Wu, Z., & Mao, H. (2025). How do early-stage active smart facilities influence traffic efficiency and energy consumption? A quantitative inquiry through panel driving data. *Green Energy and Intelligent Transportation*, 100287.
- Sun, J., Huang, Y., Tian, Y., Yu, R., & Zhao, X. (2024). Development and Application of a Public Service Platform for Autonomous Driving Testing. *China Journal of Highway and Transport*, 37(8), 248-258.
- Sun, J., Zhang, H., Zhou, H., Yu, R., & Tian, Y. (2021). Scenario-based test automation for highly automated vehicles: A review and paving the way for systematic safety assurance. *IEEE transactions on intelligent transportation systems*, 23(9), 14088-14103.
- Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., ... & Anguelov, D. (2020). Scalability in perception for

- autonomous driving: Waymo open dataset. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2446-2454).
- Uzair, M., & Jamil, N. (2020, November). Effects of hidden layers on the efficiency of neural networks. In 2020 IEEE 23rd international multitopic conference (INMIC) (pp. 1-6). IEEE.
- Wang, S., Ni, Y., Miao, C., Sun, J., & Sun, J. (2025). A multiagent social interaction model for autonomous vehicle testing. *Communications in Transportation Research*, 5, 100183.
- Wang, W., Li, W. B., Qu, F. F., Dong, T., Wang, G. Y., Wu, L. G., & Shi, C. Y. (2024). An improved grey relational theory evaluation method: considering the comprehensive performance of autonomous vehicles in virtual test. *International Journal of Automotive Technology*, 25(5), 1235-1249.
- Wei, S., Pfeffer, P. E., & Edelmann, J. (2023). State of the art: Ongoing research in assessment methods for lane keeping assistance systems. *IEEE Transactions on Intelligent Vehicles*.
- Woo, J., & Ahn, C. (2024, June). Towards Human-Like Autonomous Vehicles: A Qualitative Evaluation and Design Perspective. In 2024 IEEE Intelligent Vehicles Symposium (IV) (pp. 403-408). IEEE.
- You, S., Luo, X., Liang, X., Yu, J., Zheng, C., & Gong, J. (2025, May). A comprehensive llm-powered framework for driving intelligence evaluation. In 2025 IEEE International Conference on Robotics and Automation (ICRA) (pp. 7028-7035). IEEE.
- Zhao, G., Wang, X., Zhu, Z., Chen, X., Huang, G., Bao, X., & Wang, X. (2025, April). Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 10, pp. 10412-10420).
- Zhao, X., Wang, M., Fang, S., & Sun, J. (2023, June). Towards active motion planning in interactive driving scenarios: A generic utility term of interaction activeness. In 2023 IEEE Intelligent Vehicles Symposium (IV) (pp. 1-6). IEEE.
- Zhao, X., Sun, J., & Wang, M. (2024). Measuring sociality in driving interaction. *IEEE Transactions on Intelligent Transportation Systems*, 25(8), 9224-9237.
- Zhao, Y., Zhou, J., Bi, D., Mihalj, T., Hu, J., & Eichberger, A. (2026). A survey on the application of large language models in scenario-based testing of automated driving systems. *IEEE Transactions on Intelligent Transportation Systems*.
- Zhang, S., Huang, W., Gao, Z., Chen, H., & Lv, C. (2024). Wisead: Knowledge augmented end-to-end autonomous driving with vision-language model. *arXiv preprint arXiv:2412.09951*.
- Zhu, M., Wang, Y., Pu, Z., Hu, J., Wang, X., & Ke, R. (2020). Safe, efficient, and comfortable velocity control based on reinforcement learning for autonomous driving. *Transportation Research Part C: Emerging Technologies*, 117, 102662.

Author biography



Jiarui Zhang received the B.S. degree in College of Civil Aviation from Nanjing University of Aeronautics and Astronautics, Nanjing, China. He is currently pursuing the Ph.D. degree with the Department of Traffic Engineering, Tongji University, Shanghai, China. His main research interests include autonomous driving testing and evaluation, and testing scenario generation.



Jierui Chen is undergraduate student of the Class of 2023 at the Guohao School of Tongji University, majoring in Transportation Engineering. His research interests include computer vision for autonomous-driving perception and the evaluation of autonomous-driving intelligence.



autonomous vehicles.

Shiyu Fang received the B.S. degree in transportation engineering from Jilin University, Changchun, China. He is currently pursuing the Ph.D. degree with the Department of Traffic Engineering, Tongji University, Shanghai, China. His main research interests include decision making and motion planning for



Chao Huang received the Ph.D. degree from the University of Wollongong, Wollongong, NSW, Australia, in 2018. She is currently a Senior Lecturer with the School of Electrical and Mechanical Engineering, College of Engineering and Information Technology, The University of Adelaide, South Australia, Australia. Her research interests are human-machine collaboration, mobile robots, and path planning. Dr. Huang serves as an Associate Editor for *IEEE Transactions on Transportation Electrification*, *Engineering Applications of Artificial Intelligence*, and *Control Engineering Practice*.



Peng Hang is a Research Professor at the Department of Traffic Engineering, Tongji University, Shanghai, China. He received the Ph.D. degree with the School of Automotive Studies, Tongji University, Shanghai, China, in 2019. He was a Visiting Researcher with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, in 2018. From 2020 to 2022, he served as a Research Fellow with the School of Mechanical and Aerospace Engineering, Nanyang Technological University,

Singapore. His research interests include vehicle dynamics and control, decision making, motion planning and motion control for autonomous vehicles. He serves as an Associate Editor of IEEE Transactions on Vehicular Technology, Journal of Field Robotics, IET Smart Cities, and SAE International Journal of Vehicle Dynamics, Stability, and NVH.



Jian Sun received the Ph.D. degree from Tongji University in 2006. Subsequently, he was at Tongji University as a Lecturer and then promoted to the position as a professor in 2011, where he is currently a Professor with the College of Transportation Engineering and the Dean of the Department of Traffic Engineering. His

main research interests include traffic flow theory, traffic simulation, connected vehicle-infrastructure system, and intelligent transportation systems

Just Accepted