ISSN 2616-5775 Vol. 9, Issue 9: 31-39, DOI: 10.25236/AJCIS.2026.090905

Driving Behavior Modeling of Human Drivers on Highways Based on Maximum Entropy Deep Inverse Reinforcement Learning and Non-Cooperative Game Theory

Qiyue Chang[1,a,*]

[1]*School of Transportation and Logistics, Southwest Jiaotong University, Chengdu, 611756, China*
[a]*chainback@foxmail.com*

Abstract: *Deploying autonomous vehicles (AVs) in mixed traffic requires a background traffic environment that reproduces the variability and uncertainty of human driving. This paper proposes MIG, a hierarchical highway human driving behavior model combining Maximum Entropy Deep Inverse Reinforcement Learning (MEDIRL) with non-cooperative game theory. Surrounding traffic is organized into three layers: the outer layer replays real trajectory data; the inner layer re-drives vehicles disturbed by the ego vehicle with a MEDIRL model whose nonlinear reward is a deep neural network, avoiding manual feature engineering; and the game layer selects the highest-risk neighbor and resolves the interaction as a Nash equilibrium. Trained and validated on NGSIM US-101 in a purpose-built environment (ngsim_env), MIG converges stably, generates game trajectories close to the real ones, and reproduces the speed and acceleration distributions of real traffic more faithfully than the IDM+MOBIL and IRL baselines.*

Keywords: *Deep inverse reinforcement learning, Maximum entropy inverse reinforcement learning, Game theory, Human driving behavior modeling*

1. Introduction

Deployment of autonomous vehicles (AVs) is expanding, so AV intelligence must be verified in diversified scenarios [1,2]. Simulation is now the primary testing means owing to its low cost and controllability, and AVs will share the road with human-driven vehicles (HDVs), whose flows are dynamic, stochastic and heterogeneous [3].

Validating AV algorithms in mixed traffic requires an accurate human driving model [4]. Vehicles generated from predefined parameters [5] only execute fixed maneuvers like moving obstacles, while replaying sampled real trajectories [6] yields offline scenes that cannot respond to the vehicle under test.

Classical microscopic models such as IDM [5], Newell [6], Gipps [7] and MOBIL [8] establish interaction rules, but their predefined patterns make simulated reactions mechanical and fail to capture human driving uncertainty. Clustering drivers into aggressive, conservative and normal types [9,10] still leaves the fleet deterministic.

Human driving is summarized as observing, predicting, forming trajectory candidates and deciding [11]. Physical kinematic models are interpretable but unsuitable for long-horizon multi-agent prediction, whereas deep models such as CNNs [13] and LSTMs [14] are efficient but less interpretable. Unlike AVs, HDVs cannot communicate and are limited by human perception [12].

Trajectory generation and decision making can be policy-based or reward-based. Policy-based methods are dominated by imitation learning such as generative adversarial imitation learning [15]. Game theory instead captures inter-driver interaction by modeling the mutual influence of simultaneous actions [16]; to avoid exact Nash solutions, Xu et al. reduced the problem to a two-vehicle game between the ego vehicle and its highest-risk neighbor [17], whose equilibrium is not globally optimal, reflecting human bounded rationality.

Reward-based methods are mainly reinforcement learning with hand-crafted rewards [18]. Inverse reinforcement learning (IRL) infers the reward from demonstrations rather than designing it manually, and maximum entropy IRL (MEIRL) avoids reward ambiguity [19]. Maximum entropy deep IRL

ISSN 2616-5775 Vol. 9, Issue 9: 31-39, DOI: 10.25236/AJCIS.2026.090905

(MEDIRL) replaces hand-crafted features [20] with a neural network, allowing raw high-dimensional states and nonlinear rewards [19]. Existing studies, however, address car-following only and ignore interaction among surrounding agents [21].

Human trajectories comprise a small set of basic maneuvers, typically lane changing, keeping, accelerating and braking [17]; considering such actions at every planning step fragments the learning process [17]. sampling-based trajectory sets and flexible-band path planning [17] have been proposed to estimate the partition function of MEIRL efficiently. Nevertheless, most existing works neither build a structural hypothesis of human driving behavior for reward learning nor establish the interaction behaviors of surrounding agents in the environment.

To fill these gaps, this paper proposes a hierarchical highway human driving behavior model, MIG, with three contributions:

(1) A MEDIRL-based model is built for inner-layer vehicles with a deep neural network as the nonlinear reward function, so that features are learned automatically and individual styles are preserved.

(2) A non-cooperative game captures inter-vehicle interaction in real time: the highest-risk vehicle is selected as the opponent, and decisions are obtained by eliminating dominated strategies and computing the Nash equilibrium.

(3) A simulation environment, ngsim_env, is built on highway_env with the NGSIM US-101 dataset, providing trajectory sampling, training and multi-model comparison.

2. Methodology

2.1. Maximum Entropy Deep Inverse Reinforcement Learning

2.1.1. Maximum Entropy Inverse Reinforcement Learning

IRL learns a reward function from expert demonstrations rather than using a predefined one. Marginal-type algorithms are ambiguous, since different rewards may induce the same policy [22], whereas entropy-based formulations such as relative entropy IRL [23] and Bayesian IRL avoid this. Ziebart et al. formulated MEIRL as a constrained optimization problem [24], as in (1):

$$\begin{cases} max \quad -p(\varsigma_i|\omega_{n\times 1})\, log\, p(\varsigma_i|\omega_{n\times 1}) \\ s.t. \quad \sum_{path\,\varsigma_i} p(\varsigma_i|\omega_{n\times 1})\phi_{\varsigma_i,n\times 1} = \tilde{\phi}_{n\times 1} \\ \sum p(\varsigma_i|\omega_{n\times 1}) = 1 \end{cases} \tag{1}$$

Human driving is described by candidate trajectories generated under a stochastic policy. By the maximum entropy principle the trajectory probability is proportional to the exponential of its reward:

$$P(\tau \mid \theta) = \frac{e^{R(\tau \mid \theta)}}{Z(\theta)} \tag{2}$$

where P(tau|theta) is the trajectory probability under parameters theta, r(theta, tau) the reward, and Z the partition function:

$$Z(\theta) = \int_D e^{R(\tau \mid \theta)} d\tau \tag{3}$$

where D is the set of all trajectories. As the continuous integral is intractable, Z is approximated by a sum over a finite set M of sampled trajectories.

Conditioning on the same initial state, the probability of a generated trajectory is expressed as (4):

$$P(\tau \mid \theta) \approx \frac{e^{R(\tau \mid \theta)}}{\sum_{\tau_i \in M} e^{R(\tau_i \mid \theta)}} \tag{4}$$

where tau_i is a generated trajectory sharing the expert trajectory's initial state. Because human driving is stochastic, the parameters are optimized within each discretized sub-space.

The objective of MEIRL maximizes the log-likelihood of the expert demonstration, written as (5):

$$J(\theta) = R(\tau \mid \theta) - log \sum_{\tau_i \in M} e^{R(\tau_i \mid \theta)} \tag{5}$$

2.1.2. Deep Neural Network Reward Function

Conventional IRL uses a linear reward over manually designed features that cannot fully represent

ISSN 2616-5775 Vol. 9, Issue 9: 31-39, DOI: 10.25236/AJCIS.2026.090905

the factors of human driving.

A deep neural network is therefore adopted as the nonlinear reward function, extracting features automatically and fitting the complex reward landscape of real driving.

The reward network (Fig. 1) has two hidden layers with ReLU activations and a sigmoid output layer. Its input is the lateral/longitudinal positions and speeds of the vehicle and its neighbors within a given time window, for both candidate and real trajectories; its output is the reward of each candidate. Hyperparameters and architecture are in Tables 1 and 2.

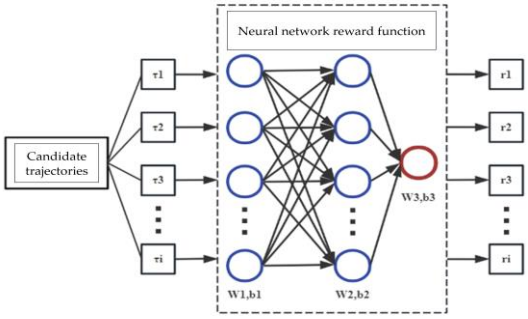


Figure 1: Deep neural network reward function.

Table 1: Training hyperparameters of the algorithm.

Hyperparameter	Value
Number of iterations	5200
Time window length	5 s
Sampling interval	0.1 s
Learning rate / step size	0.00085
Optimizer	Adam

Table 2: Architecture of the neural network.

Layer	Number of neurons	Activation function
Hidden layer 1	10	RELU
Hidden layer 2	8	RELU
Output layer	1	Sigmoid

2.1.3. Training and Application Procedure

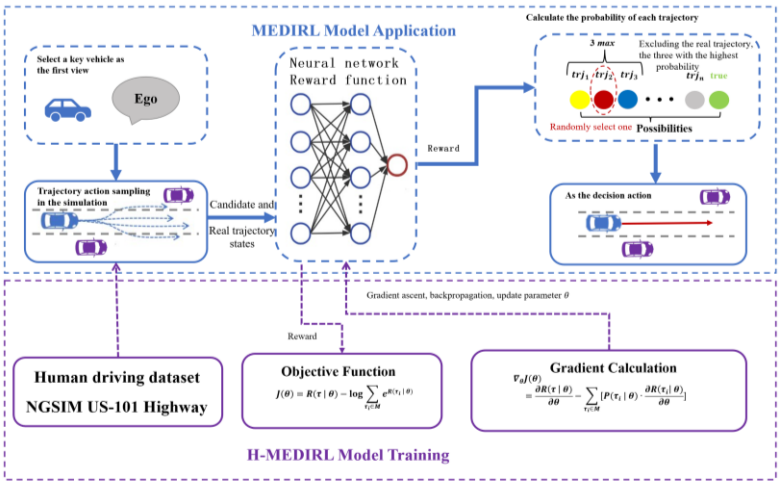


Figure 2: Basic framework of the MEDIRL model.

The training procedure (Fig. 2) samples trajectory states of a selected vehicle at a fixed interval within a 5 s window, matching the short perception horizon of human drivers, and feeds them with the real trajectory states into the reward network to compute each candidate's reward. The parameters are updated

ISSN 2616-5775 Vol. 9, Issue 9: 31-39, DOI: 10.25236/AJCIS.2026.090905

until the objective (5) converges; in application one of the three most probable trajectories is drawn randomly as the next decision.

2.2. Game-Theoretic Interactive Trajectory Generation

2.2.1. Observation Space and Influence Domain

As in real driving, a vehicle cannot access complete information about all other vehicles. Its observation space on a multi-lane highway comprises the directly front, diagonally front, directly rear and diagonally rear regions (left and right), with inclusion determined by a speed-dependent longitudinal threshold following highway safe-distance rules [25].

Based on the observation space the inner layer of the influence domain is determined dynamically in real time; its vehicles may change state at any time and are driven by the MEDIRL model of Section 2.1.

2.2.2. Action Space and Trajectory Generation

The low-level maneuvers of highway vehicles are car-following and left/right lane change laterally, and constant speed, acceleration and deceleration longitudinally; their combinations yield nine basic high-level trajectories.

Trajectory generation for both MEDIRL sampling and game equilibrium selection is constrained by the endpoints of the planning horizon: the driver plans from the current state to a terminal goal, deciding the target speed longitudinally and the lane-keeping or lane-changing tactic laterally. Enumerating all low-level actions at every step would explode the action space (420 actions for a 2 s horizon [17]). Trajectories are polynomial curves with six lateral and four longitudinal boundary conditions, expressed in (6):

$$\begin{cases} x(\tau) = a_0 + a_1\tau + a_2\tau^2 + a_3\tau^3 + a_4\tau^4 \\ y(\tau) = b_0 + b_1\tau + b_2\tau^2 + b_3\tau^3 + b_4\tau^4 + b_5\tau^5 \end{cases} \tag{6}$$

where t is the time and {a0, ..., a4} and {b0, ..., b5} are the polynomial coefficients. Given the initial state, target state and arrival time T, the boundary conditions determine the coefficients, and the position, speed and acceleration then follow analytically.

2.2.3. Game Process

Selection of the interaction vehicle. Among surrounding vehicles, the one with the highest risk is selected in real time as the game opponent. Longitudinal distance is the primary risk factor, since collision avoidance is every driver's bottom line; the longitudinal speed difference is second, and the lateral speed difference reflects lane-change intention. The risk function is designed as (7):

$$F_i = \frac{p_1}{\Delta d_i} + p_2 i_z^{\Delta v} + p_3 i_h^{\Delta v} \tag{7}$$

where d_i is the longitudinal distance between vehicle i and the ego vehicle, dv_i = v0 - v_i their longitudinal speed difference with v0 the ego speed, and |dvy_i| the absolute lateral speed difference.

The coefficients alpha, beta and gamma were tuned to 0.94, 0.01 and 0.05, completing the three-layer organization around the ego vehicle with the observation space of Section 2.2.1 and the risk function (7) (Fig. 3).

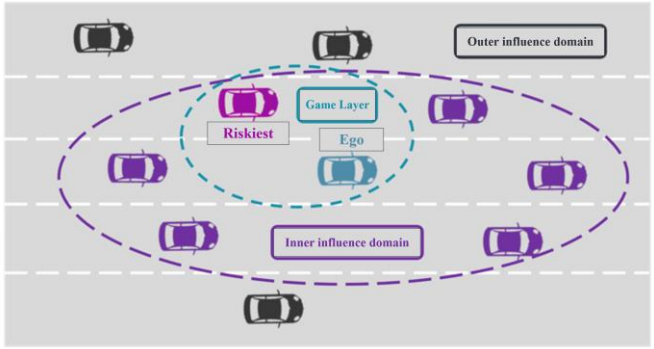


Figure 3: Overall three-layer organization of vehicle modeling.

ISSN 2616-5775 Vol. 9, Issue 9: 31-39, DOI: 10.25236/AJCIS.2026.090905

Throughout a trip the ego vehicle continuously re-evaluates its neighbors: the activated highest-risk individual interacts through the non-cooperative game, then resumes driving with the trained MEDIRL model once deactivated.

Observation and prediction. Once the high-risk individual is determined, the two players observe each other. Each player's observation set contains the opponent's longitudinal/lateral position, speed, acceleration and heading angle at time t, while a vehicle state consists of longitudinal/lateral position and speed, as defined in (8):

$$S = \{x^t, y^t, v^t\} \tag{8}$$

Each player then estimates the next-step states of itself and the opponent from the kinematic relations, with dt = 0.1 s.

Reward matrix. Each player generates a 9 x 9 payoff matrix R, whose entry (i, j) is the ego reward when the ego takes action i and the opponent takes action j within the window from t to t+n, defined as (9):

$$R = \frac{k_1 int_{t+n}^{\Delta d} + k_2 \Delta v_{self}}{self_{t+n}^{k_3 \Delta d} + L_0} \tag{9}$$

where dist is the distance between the two vehicles at the next state t+n under the action pair, dv the ego speed change relative to its original speed, T the planning-window length, and dl its lateral offset relative to the constant-speed car-following state.

Nash equilibrium computation. Combining the two payoff matrices yields a strategy-value table. For example, vehicle 1207 at step 90 identifies vehicle 1210 as the highest-risk neighbor; Table 3 lists the table, whose rows are actions of vehicle 1207, columns actions of vehicle 1210, and cells the payoffs (r_1207, r_1210) at the predicted next state.

Dominated strategies are eliminated first. For vehicle 1207 the three decelerated strategies are strictly dominated, since the opponent is behind and deceleration would shorten the gap; for vehicle 1210 all strategies involving acceleration or a left lane change are dominated. This leaves a single cell, which is itself the equilibrium: vehicle 1207 keeps constant-speed following and vehicle 1210 a constant-speed right change.

Table 3: Strategy value between vehicle 1207 and vehicle 1210 after eliminating dominated strategies.

Strategy value table after eliminating dominated strategies		**Action of vehicle 1210**
		7
Action of vehicle 1207	1	(5.90, 5.27)

Nash non-uniqueness can also be illustrated: for vehicles 333 and 326 at step 330, two equilibria remain after eliminating dominated strategies (both keep constant-speed following, or 326 changes right while 333 changes left). Although the first dominates in payoff, the second is not discarded, because no player gains by unilaterally changing its strategy.

3. Data Processing and Simulation Environment

3.1. Trajectory Preprocessing

The preprocessing pipeline organizes the CSV records into vehicle and snapshot dictionaries, sorts them by time, deduplicates within each snapshot and downsamples by the sampling rate. Because raw trajectories are noisy, a Savitzky-Golay filter [26] (third-order polynomial, 2 s window) smooths them into the demonstration trajectories for reward learning.

3.2. Simulation Environment

Based on the open-source platform highway_env, an environment named ngsim_env reproduces the NGSIM US-101 scenario. It simulates and visualizes highway vehicle driving and interaction and represents trajectories with polynomial curves covering longitudinal/lateral positions and speed/acceleration. Target states are sampled to generate candidate trajectories, and the three most probable ones are drawn randomly.

ISSN 2616-5775 Vol. 9, Issue 9: 31-39, DOI: 10.25236/AJCIS.2026.090905

4. Experiments and Comparative Analysis

4.1. Evaluation Metrics

The model is evaluated at two levels: microscopically by the agreement between the real and predicted trajectories of the same vehicle (time series of speed, acceleration and longitudinal/lateral position) under the same scenario, and macroscopically by the speed and acceleration distributions of multiple simulated vehicles against the real ones.

4.2. Experimental Design

The experiments proceed as follows: (1) ngsim_env reproduces the US-101 scenario; (2) the NGSIM data are cleaned, filtered and smoothed (Section 3); (3) the model is trained with MEDIRL to predict trajectory probabilities; (4) trajectories are generated by simulating the target vehicle's interaction with its neighbors, drawing one of the three most probable ones at random; and (5) the results are compared with the real data and other models.

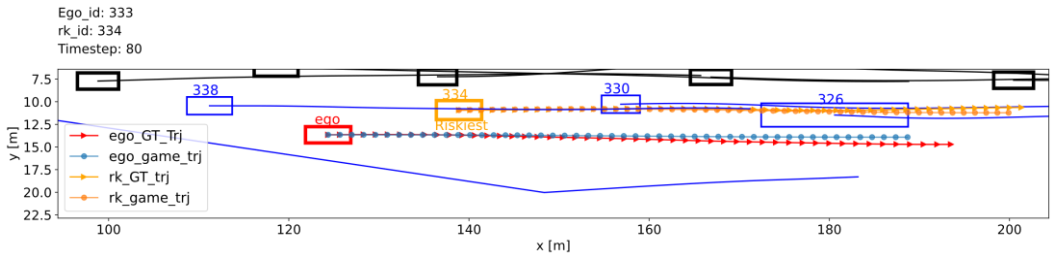
4.3. Results

4.3.1. Accuracy of the MEDIRL Algorithm

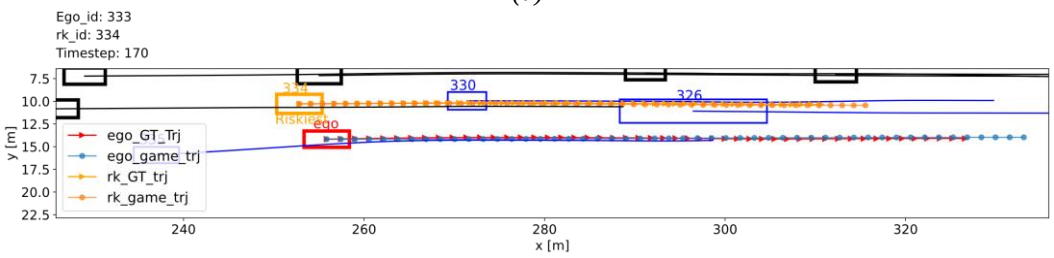
Taking vehicle 1200 as an example, multiple 5 s segments of its US-101 trip were used for training, and the objective function (5) converged for all segments. Several test segments were randomly selected: the three most probable trajectories follow the real trend and keep the current lane, while the occasional different lane choice reflects human driving randomness.

4.3.2. Accuracy of the Game-Based Trajectories

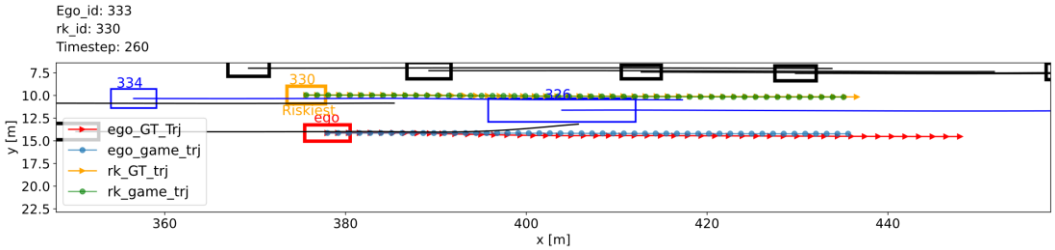
Vehicles 333 and 338 were randomly selected for individual examination. Fig. 4 shows the game-planned and real trajectories of vehicle 333 during selected periods. Most Nash equilibria during the trip consist of a single combination; at step 330 two equilibria were obtained, and comparing payoffs the both-going-straight solution yields higher rewards, consistent with reality.



(a)



(b)



(c)

ISSN 2616-5775 Vol. 9, Issue 9: 31-39, DOI: 10.25236/AJCIS.2026.090905



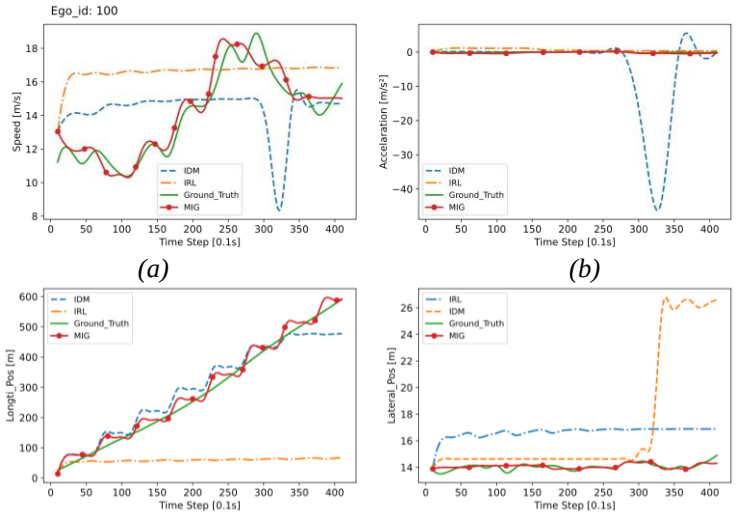
(d)

Figure 4: Game-planned trajectories of vehicle 333 and its interaction partners versus real trajectories: (a)-(d) selected periods.

For vehicle 338 the planned trajectories coincide with the real ones most of the time, the residual differences reflecting an acceptable speed prediction error. At step 220 the high-risk partner 341 changed to the left lane in reality, and the game predicted the same pattern, though the planned curvature differs slightly because polynomials are more regular than real lane-change maneuvers.

4.3.3. Comparison with Other Models

The trained model (MEDIRL + Game, i.e., MIG) was compared with the IDM+MOBIL (IDM) and IRL models at both levels against the real NGSIM data. Microscopically, Fig. 5 shows, for vehicle 100, the acceleration, speed and longitudinal/lateral position profiles of the three models and the real ones.



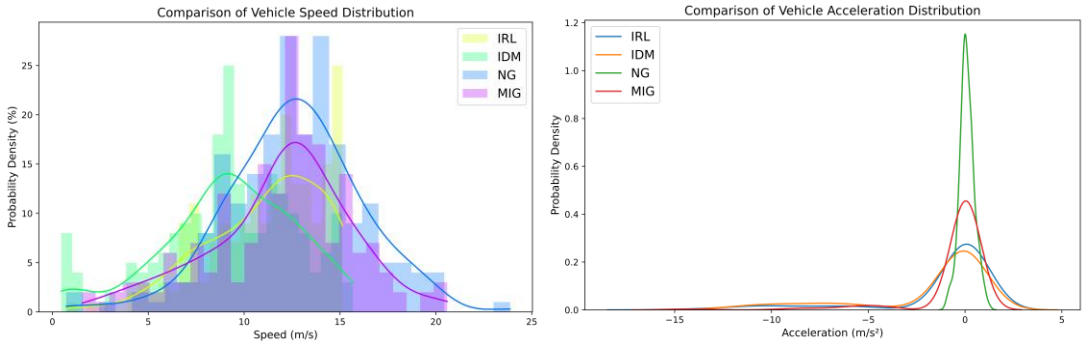
(a) *(b)*

(c) *(d)*

Figure 5: Kinematic profiles of vehicle 100 generated by different models versus the real data: (a) speed; (b) acceleration; (c) longitudinal position; (d) lateral position.

Macroscopically the speed and acceleration distributions measure the generated traffic environment. The simulation time follows the real trajectory duration, 100 trajectories are run for each model, and 20 randomly selected vehicles provide the reference distributions (Figs. 6).

ISSN 2616-5775 Vol. 9, Issue 9: 31-39, DOI: 10.25236/AJCIS.2026.090905



(a) *(b)*

Figure 6: Comparison of the speed (a) and acceleration (b) distributions of the vehicle trajectories generated by different models.

The speed and acceleration distributions of MIG are closer to the real NGSIM distributions than those of IDM+MOBIL and IRL in both shape and median. IDM+MOBIL deviates most in the speed distribution, while the acceleration distribution of MIG is more concentrated, closer to the real dataset.

5. Conclusions

This paper proposed MIG, a highway human driving behavior model combining maximum entropy deep inverse reinforcement learning (MEDIRL) with non-cooperative game theory. Trained on the NGSIM dataset, it preserves individual driving styles in three layers: the outer layer replays real driving data, the inner layer reproduces human-like behavior after disturbances with MEDIRL, and the game layer captures real-time interaction between the ego vehicle and its highest-risk neighbor. MEDIRL converges stably, the game model predicts interactive trajectories faithfully, and MIG outperforms the IDM+MOBIL and IRL baselines, especially in speed fitting.

Several limitations remain: the algorithm is restricted to discrete state-space problems; the learned features reduce interpretability; there is no definitive reference for verifying that the modeled randomness matches reality; only non-cooperative games are used, ignoring willing cooperators; and layer switching may cause instability. Future work will address cooperative games, reward interpretability and layer-switching stability.

The proposed model supports human driving behavior simulation in mixed traffic, providing a realistic background traffic environment for AV research and a theoretical basis for autonomous driving system design.

References

[1] Wang M, Zhang L, Wang Z, et al. A Real-Time Dynamic Trajectory Planning for Autonomous Driving Vehicles. 2019 3rd Conference on Vehicle Control and Intelligence (CVCI). Hefei, China: IEEE, 2019: 1-6.

[2] Feng S, Yan X, Sun H, et al. Intelligent Driving Intelligence Test for Autonomous Vehicles with Naturalistic and Adversarial Environment. Nature Communications, 2021, 12(1): 748.

[3] Nallaperuma D, De Silva D, Alahakoon D, et al. Intelligent Detection of Driver Behavior Changes for Effective Coordination Between Autonomous and Human Driven Vehicles. IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society. Washington, DC: IEEE, 2018: 3120-3125.

[4] Kalra N, Paddock S M. Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability? Transportation Research Part A: Policy and Practice, 2016, 94: 182-193.

[5] Treiber M, Hennecke A, Helbing D. Congested Traffic States in Empirical Observations and Microscopic Simulations. Physical Review E, 2000, 62(2): 1805-1824.

[6] Newell G F. A Simplified Car-Following Theory: A Lower Order Model. Transportation Research Part B: Methodological, 2002, 36(3): 195-205.

[7] Gipps P G. A Behavioural Car-Following Model for Computer Simulation. Transportation Research Part B: Methodological, 1981, 15(2): 105-111.

[8] Kesting A, Treiber M, Helbing D. General Lane-Changing Model MOBIL for Car-Following Models. Transportation Research Record, 2007, 1999(1): 86-94.

ISSN 2616-5775 Vol. 9, Issue 9: 31-39, DOI: 10.25236/AJCIS.2026.090905

[9] Ikotun A M, Ezugwu A E, Abualigah L, et al. K-means Clustering Algorithms: A Comprehensive Review, Variants Analysis, and Advances in the Era of Big Data. Information Sciences, 2023, 622: 178-210.
[10] Siami M, Naderpour M, Lu J. A Mobile Telematics Pattern Recognition Framework for Driving Behavior Extraction. IEEE Transactions on Intelligent Transportation Systems, 2021, 22(3): 1459-1472.
[11] Brown K, Driggs-Campbell K, Kochenderfer M J. A Taxonomy and Review of Algorithms for Modeling and Predicting Human Driver Behavior. arXiv preprint, 2020, arXiv:2006.12188.
[12] Li N, Oyler D W, Zhang M, et al. Game Theoretic Modeling of Driver and Vehicle Interactions for Verification and Validation of Autonomous Vehicle Control Systems. IEEE Transactions on Control Systems Technology, 2018, 26(5): 1782-1797.
[13] Krizhevsky A, Sutskever I, Hinton G E. ImageNet Classification with Deep Convolutional Neural Networks. Communications of the ACM, 2017, 60(6): 84-90.
[14] Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation, 1997, 9(8): 1735-1780.
[15] Kebria P M, Khosravi A, Salaken S M, et al. Deep Imitation Learning for Autonomous Vehicles Based on Convolutional Neural Networks. IEEE/CAA Journal of Automatica Sinica, 2020, 7(1): 82-95.
[16] Georgakopoulos G F, Kavvadias D J, Sioutis L G. Nash Equilibria in All-Optical Networks. Discrete Mathematics, 2009, 309(13): 4332-4342.
[17] Xu C, Zhao W, Li L, et al. A Nash Q-Learning Based Motion Decision Algorithm with Considering Interaction to Traffic Participants. IEEE Transactions on Vehicular Technology, 2020, 69(11): 12621-12634.
[18] Zhu M, Wang X, Wang Y. Human-Like Autonomous Car-Following Model with Deep Reinforcement Learning. Transportation Research Part C: Emerging Technologies, 2018, 97: 348-368.
[19] Wulfmeier M, Ondruska P, Posner I. Maximum Entropy Deep Inverse Reinforcement Learning. arXiv preprint, 2016, arXiv:1507.04888.
[20] Huang Z, Wu J, Lv C. Driving Behavior Modeling Using Naturalistic Human Driving Data with Inverse Reinforcement Learning. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(8): 10239-10251.
[21] Zhou Y, Fu R, Wang C. Learning the Car-Following Behavior of Drivers Using Maximum Entropy Deep Inverse Reinforcement Learning. Journal of Advanced Transportation, 2020, 2020: 1-13.
[22] Abbeel P, Ng A Y. Apprenticeship Learning via Inverse Reinforcement Learning. Twenty-First International Conference on Machine Learning (ICML '04). Banff, Alberta, Canada: ACM Press, 2004: 1.
[23] Boularias A, Kober J, Peters J. Relative Entropy Inverse Reinforcement Learning. Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS). 2011: 182-189.
[24] Ziebart B D, Maas A, Bagnell J A, et al. Maximum Entropy Inverse Reinforcement Learning. Proceedings of the 25th International Conference on Machine Learning (ICML 2008). Helsinki, Finland: ACM Press, 2008: 1433-1438.
[25] Road Traffic Safety Law of the People's Republic of China (2021 Revision). Beijing Traffic Administration. Available online: https://jtgl.beijing.gov.cn/jgj/jgxx/flfg/fl/205308/index.html (accessed on 13 May 2024).
[26] Schafer R. What Is a Savitzky-Golay Filter? [Lecture Notes]. IEEE Signal Processing Magazine, 2011, 28(4): 111-117.