

A Two-stage Based Social Preference Recognition in Multi-Agent Autonomous Driving System

Jintao Xue*, Dongkun Zhang*, Rong Xiong, Yue Wang, and Eryun Liu

Abstract—Multi-Agent Reinforcement Learning (MARL) has become a promising solution for constructing a multi-agent autonomous driving system (MADS) in complex and dense scenarios. But most methods consider agents acting selfishly, which leads to conflict behaviors. Some existing works incorporate the concept of social value orientation (SVO) to promote coordination, but they lack the knowledge of other agents' SVOs, resulting in conservative maneuvers. In this paper, we aim to tackle the mentioned problem by enabling the agents to understand other agents' SVOs. To accomplish this, we propose a two-stage system framework. Firstly, we train a policy by allowing the agents to share their ground truth SVOs to establish a coordinated traffic flow. Secondly, we develop a recognition network that estimates agents' SVOs and integrates it with the policy trained in the first stage. Experiments demonstrate that our developed method significantly improves the performance of the driving policy in MADS compared to two state-of-the-art MARL algorithms.

I. INTRODUCTION

The self-driving technology is widely regarded as a means to improve the safety and efficiency of transportation, leading to an increasing number of autonomous vehicles undergoing tests in the context of multi-agent autonomous driving systems (MADSs) [1]. However, current MADSs encounter difficulties operating in complex environments, including highway merging and bottleneck situations. In these scenarios, road users' actions affect others' behaviors, and a tiny conflict could lead to the decay of traffic efficiency. We aim to address the challenge and design an efficient and safe MADS in these traffic environments.

To this end, we summarize several learning paradigms. The rule-based approaches use manual rules or classical traffic models [2], [3] but have limitations in complex traffic scenarios. Multi-Agent Reinforcement Learning (MARL) holds great potential and demonstrates positive outcomes [4], [1], [5]. As depicted in Fig. 1, most MARL-based MADSs consider agents acting selfishly, producing egoistic behaviors that harm the whole efficiency. To address this issue, several works [6], [7] introduce the concept of Social Value Orientation (SVO) [8], which measures the degree of selfishness or altruism of the agent by weighting its rewards with those of others, to promote socially compatible behavior. However, these methods do not know SVOs of other agents and produce conservative behaviors. It is

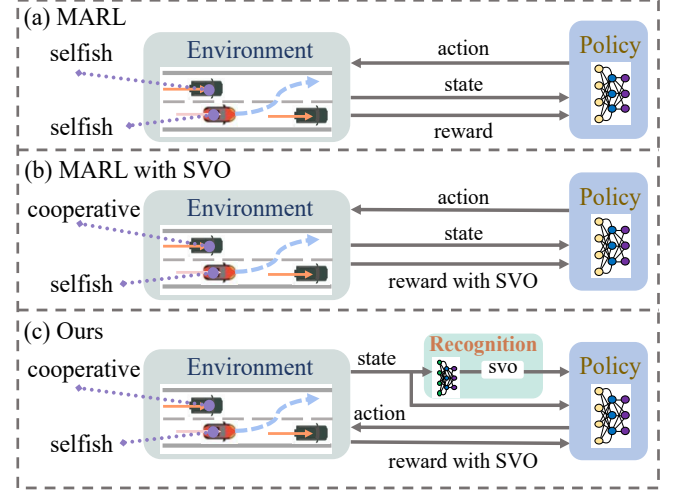


Fig. 1: Illustration comparing (a) Classical MARL-based approach, (b) MARL with SVO, and (c) our recognition-based policy to have knowledge of other agents' SVOs, thus letting agents make SVO-informed decisions.

important to note that social preferences, referred to as SVOs, can vary among individuals and significantly affect the interactions between agents [9], and neglecting to account for the SVO can negatively impact the safety and efficiency of traffic flow [8].

Due to this limitation, in our method, agents are able to understand surrounding agents' SVOs, thereby enabling themselves to make SVO-informed driving, as shown in Fig. 1 (c). We start by adopting the MARL approach and incorporating SVO to model MADS. Specifically, we introduce a two-stage training framework. Firstly, we train a policy of actual SVOs, which are the inner parameters of agents to build up a coordinated traffic flow. However, obtaining SVOs from other agents poses a challenge as they are generally considered private information. Additionally, this paper considers the scenario of communication breakdown. Hence, in the second stage, we train a recognition policy that can recognize other agents' SVOs and integrate the recognition policy with the policy learned in the first stage. The performance of the two-stage policy is close to that of directly knowing the true SVOs. Our experiments explore the effect of varying levels of knowledge among agents on system-level performance metrics, revealing that knowing SVOs leads to more effective and coordinated driving among agents.

To summarize, the main contributions of this paper include the following:

*The authors are equally contributed.

Jintao Xue and Eryun Liu are with the Dept of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. Dongkun Zhang, Rong Xiong and Yue Wang are with the Dept of Control Science and Engineering, Zhejiang University, Hangzhou, China. Eryun Liu is the corresponding author, eryunliu@zju.edu.cn.

- A two-stage training procedure by knowing the SVOs of agents in advance to build up a first-stage MADS, then estimating the SVOs to build up the final MADS.
- A SVO recognition framework that leverages the power of self-attention models to tackle complex driving environments. This framework seamlessly integrates multiple sources of information to obtain accurate estimations.
- Our proposed approach for MADS is validated on a simulation environment and compared against two state-of-the-art MARL-based methods. The evaluation results demonstrate that our method outperforms others in terms of performance.

II. RELATED WORKS

In this section, we review relevant MARL approaches in the context of MADS. We also discuss recent techniques used to generate heterogeneous driving behavior and online parameter estimation approaches for navigation.

A. Deep Reinforcement Learning in MADS

The application of Multi-Agent Reinforcement Learning (MARL) produces promising outcomes in MADS. Authors of [4] apply the DRL method to the problem of forming long-term driving strategies while ensuring functional safety, and a hierarchical temporal abstraction is introduced to reduce the variance of the gradient estimation. Palanisamy *et al.* [5] provide a simulation platform and a taxonomy of multi-agent learning environments to help further research. Wang *et al.* [10] utilize graph attention networks in the navigation setting of MARL for mixed-autonomy cooperation. Liu *et al.* [11] propose a distributed training framework for deep Q-networks to deal with the multi-vehicle platooning problem. [12], [13], [7] incorporate SVO to acquire socially compliant behavior, and the third paper obtains better results using a multi-agent actor-critic algorithm. Dai *et al.* [14] dynamic change SVOs for interacting agents in each episode. Peng *et al.* [6] use the coordination factor to facilitate the coordination of agents at both local and global levels in fully autonomous traffic flow. We follow this approach by incorporating SVO. However, we design a recognition framework to know other agents' SVOs.

B. Heterogeneity of Vehicle Agents

Considerable methods design heterogeneous agents having diverse driving styles to reflect real-world driving scenes. The intelligent driver model (IDM) [2] and the MOBIL lane-changing model are often combined as a model of human drivers [3]. Furthermore, some approaches adjust the IDM-MOBIL model's parameters (e.g., politeness factor) to get different levels of aggressiveness. Saxena *et al.* [15] modify IDM to include a stop-and-go behavior and diverse cooperativeness. Mavrogiannis *et al.* [16] present an algorithm to conduct behavior-rich simulation consisting of egoistic and conservative agents. [17], [18], [8], [19] follow a weighted cost function, and varied weight metrics characterizes the difference between individuals,

and the weights can be manually tuned or apply Inverse Reinforcement Learning (IRL) to learn from real-human data. Schwarting *et al.* [8] also preferably employ SVO in autonomous driving, determining the degree of competitive and prosocial. And many researchers integrate the concept of SVO into their works [6], [7], [17], [20], [21], [22], [23].

C. Online Parameter Estimation

Several works estimate social preferences in driving. [24], [25] use online filtering techniques to estimate parameters in IDM. Authors of [26] use an unscented Kalman filter to iteratively update a Bayesian estimate of other agents' cost function parameters. Li *et al.* [22] identify other drivers' driving preferences by estimating the SVOs. [8], [21] estimate the SVO of the agent to improve predictions and prove essential assets for interactive driving. Wang *et al.* [17] allow agents to infer other road users' characteristics include egoism, courtesy, and confidence. However, these model-based methods assume that the agent fully knows the state transition function or other agents' objective functions. But what if the function is a black box, or more precisely, the internals of the environment will be unknown to an agent? For example, give random discrete estimation values of one objective, and each value needs to go through the black box and get the outcome, which means running a significant number of parallel processes to get the running results, which is costly to call. However, our policy can handle the black-box environment with dense agents as we do not need the objective or transition function.

III. METHOD

Based on the discussion above, we propose a two-stage strategy to solve MADS problem using MARL: (i) train a policy to coordinate traffic flow by knowing true SVOs, (ii) adopt a policy to estimate agents' SVOs. For ease of reference, we denote these two policies as the "decision policy" and the "recognition policy". First, we provide problem definitions of the two policies. We then describe the reward function and state representation used in the environment, then followed by an introduction to the overall network architecture.

A. Decision Policy Training

1) Partially Observable Stochastic Game (POSG):

We formulate the decision-making processes in MADS using a stochastic game [27] by the tuple $G = \langle \mathcal{I}, \mathcal{S}, \mathcal{A}, \mathcal{O}, P, \mathcal{R}, n, \rho_0, \gamma, T \rangle$. \mathcal{I} represents a finite set of n agents, \mathcal{S} represents the state-space of all agents, while $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \cdots \times \mathcal{A}_n$, and $\mathcal{O} = \mathcal{O}_1 \times \mathcal{O}_2 \cdots \times \mathcal{O}_n$ denote the joint action, and observation spaces, respectively. At a time agent $i \in \mathcal{I}$ receive the observation $o_i : \mathcal{S} \rightarrow \mathcal{O}_i$ and take action based on the shared policy $\pi : \mathcal{O}_i \times \mathcal{A}_i \rightarrow [0, 1]$. Consequently, the full state changes from s to s' after all agents take their actions w.r.t the state transition function $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$. The rewards denoted by an agent-specific reward functions $\mathcal{R}_i \in \mathcal{R}$ where $\mathcal{R} = \{R_0, R_1, \dots, R_{n-1}\}$. ρ_0 is the initial state distribution, $\gamma \in (0, 1]$ is the discount factor, and T is the time horizon.

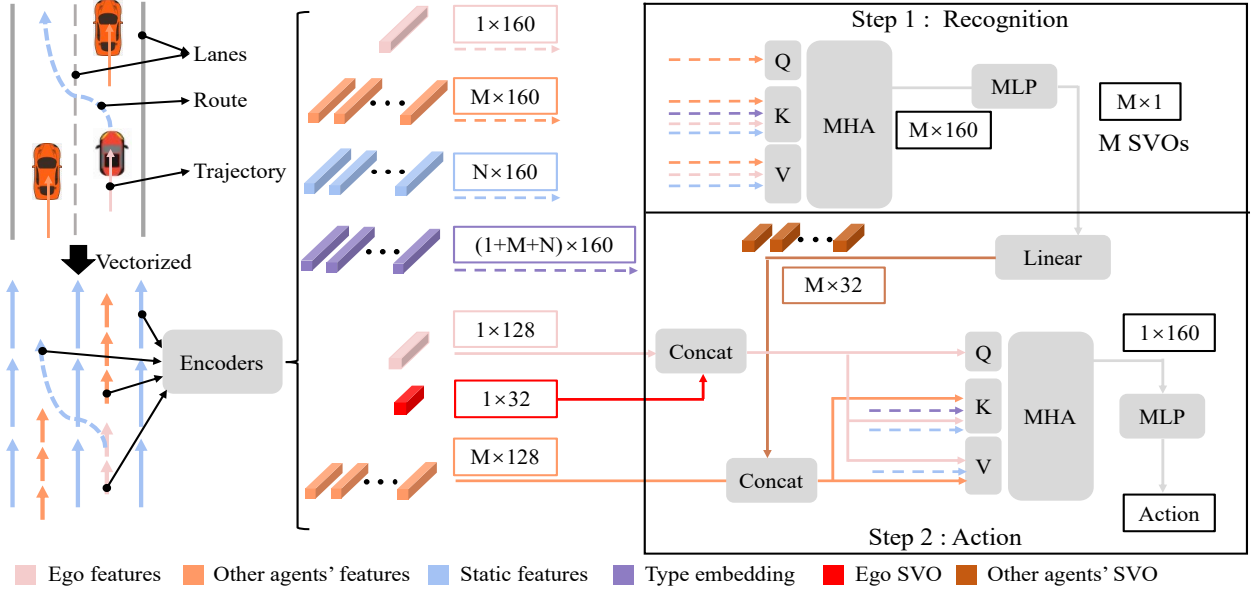


Fig. 2: The architecture of our policy consists of encoders that embed the input vectorized information into different features. These features then go through the first multi-head attention (MHA) layer on the upper right, which outputs information about M surrounding agents' SVOs. The resulting features then pass through the second MHA to produce the final output.

2) *Decentralized reward function:* Our approach leverages the SVO concept from prior research [8], [28], [6] and integrates it into our framework. By doing so, each agent is able to exhibit varying behaviors, such as aggressive or cooperative, depending on the extent to which they consider other agents' rewards:

$$R_i = \cos\left(\frac{\pi}{2} \cdot \phi_i\right) r_i + \sin\left(\frac{\pi}{2} \cdot \phi_i\right) r_i^s, \quad (1)$$

$$r_i^s = \frac{\sum_{j \in \mathcal{D}} r_j}{|\mathcal{D}|}, \quad \mathcal{D} = \{j : \|\text{Pos}(i) - \text{Pos}(j)\| \leq d\},$$

the reward r_i represents the individual driving performance, which contains metrics such as average speed and a negative reward in case of collision. r_i^s is the average of surrounding agents' utilities, and the d means each agent only perceives information from other agents within a certain Euclidean distance. $\phi_i \in [0, 1]$ is the SVO of agent i and remains constant throughout each episode.

3) *Objective Function:* We adopt the decentralized learning method to solve the optimization objective of POSG:

$$\pi^* = \arg \max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=0}^T \gamma^t R_i(s_t, a_t) \right], \quad i \in \mathcal{I}, \quad (2)$$

where $\pi : \mathcal{O}_i \times \mathcal{A}_i \rightarrow [0, 1]$, which uses experiences of all agents to learn a strategy with sharing of policy network parameters. Due to unique observations and indices by agents, diverse behaviors emerge, aligning with concepts presented in [29].

B. Recognition Policy Training

According to III-A, we can train a decision policy under the assumption of knowing other SVOs. However, it is

more reasonable to consider SVO as an inner parameter that can not be directly observed. This section presents a policy to recognize surrounding agents' SVOs, coupled with the trained decision policy. Our training approach involves directly fitting the actual and the predicted value of SVOs. Given a set of observations, including static and vehicular features, our task is to predict the values of SVOs for surrounding agents. To generate training data, we run the POSG with the trained decision policy and store the resulting observations. We denote SVO as ϕ and define the mean square error (MSE) loss function as the average of squared differences between the actual and the predicted values of SVOs:

$$\mathcal{L}_{reg} = \frac{1}{N \sum_{i=0}^N M_i} \sum_{i=0}^N \sum_{j=0}^{M_i} (\phi_{ij} - \hat{\phi}_{ij})^2. \quad (3)$$

Given N agent's observations, each agent needs to estimate the values of ϕ for its neighboring agents. M_i denote the number of surrounding agents for the i -th agent, and $\hat{\phi}_{ij}$ represent the estimated value of the j -th agent's ϕ by the i -th agent.

C. State Space Representation

1) *State Space for decision policy:* To improve computational and memory efficiency, we implement a vectorized representation strategy known as VectorNet [30]. The state space contains static and vehicular set $\chi = \{\chi^s, \chi^v\}$, and elements in χ^s and χ^v are sets of points containing corresponding features. For the static set containing road centerlines, sidelines, and routes, $\chi^s = \{\text{centerline}, \text{sideline}, \text{route}\} = \{e_0^s, e_1^s, \dots, e_i^s, \dots\}$, where $e_i^s = \{\xi_0, \xi_1, \dots, \xi_j, \dots\}$, $i \in \chi^s$. $\xi_j = [p_j, \phi_i, i, j]$,

in which $p_j = (x, y, heading)$ is the pose of point j in element i and ϕ_i is the lane width of element i . For the vehicular set cover poses and velocities of n agents, $\chi^v = \{e_0^v, e_1^v, \dots, e_{n-1}^v\}$, in which $e_i^v = \{\xi_0, \xi_1, \dots, \xi_{horizon}\}$, $i \in \chi^v$, and $\xi_j = [p_j, \phi_i, i, j]$, $j \in e_i^v$, where $p_j = (x, y, heading, speed)$ and ϕ_i denotes the SVO of agent i . In practice, setting the agent's trajectory horizon to 10 achieves good driving performance without exceeding computational resources.

2) *State space for recognition policy*: Almost the same as mentioned in III-C.1, the state space for recognition comprises static and vehicular elements. However, the recognition policy does not require knowledge of the true SVOs, hence $\xi_j = [p_j, i, j]$, where $j \in e_i^v$.

D. Reward Function Design

The driving task involves multiple attributes when defining the reward function. These attributes include factors such as comfort and compliance with traffic regulations. To address this, we design a reward function that provides continuous incentives for driving fast while imposing penalties for catastrophic failures. These failures encompass collisions with other agents, deviations from the designated driving zone, and excessive deviation from the global path.

E. Policy Architecture

The entire network architecture is depicted in Figure 2, including encoders, recognition, and decision components. The process starts with encoding observations of the ego agent into high-level features and embedding diverse features into a uniform dimensional space as described in III-E.1. The recognition policy then focuses on M surrounding agents and receives their driving behavior information, which is embedded into M SVOs as outlined in III-E.2. Finally, the decision policy receives SVOs, including the ground truth of the ego agent, and outputs the final action detailed in III-E.3.

1) *DeepSet Encoder*: Based on theorem 2 in [31], the representation element $e \subset \chi$, where $\chi = \{\chi^s, \chi^v\}$, requires a function that preserves the adjacency between elements and is permutation-invariant to the order of objects in the element. Hence, the propagation function f is defined as follows:

$$f(e) = \rho \left(\sum_{\xi \in e} \varphi(\xi) \right). \quad (4)$$

We obtain features at the element level by transforming the nodes $\xi \in e$ into a representation $\varphi(\xi)$. The sum of representations is processed using the ρ network defined by Multi-Layer Perception (MLP) network. The DeepSet architecture enables us to extract features at the polyline level while keeping the number of parameters relatively small.

2) *Recognition Policy Architecture*: Through DeepSet, the input features are embedded into a 160-dimensional space and categorized into three groups: ego, other agents, and static elements. The multi-head attention (MHA) layer is then applied, with features of M surrounding agents serving as queries $Q_r = [q_r^1, q_r^2, \dots, q_r^M] \in \mathbb{R}^{M \times d_k}$, where d_k is the

dimension of the key vectors, set to 160. All features are keys K_r or values V_r . Following the work in [32], [33], we incorporate an additional type embedding into the keys to allow the model to attend to values based on object types. Depicting the complete computation process as follows:

$$\Phi_r = \tanh(\text{Decoder}(\text{MultiHead}(Q_r, K_r, V_r))), \quad (5)$$

where MultiHead composes several Attention operations, which calculate the weighted sum of the values using the dot-product of queries and keys. The outputs of Attention are concatenated and then transformed using a linear layer to obtain the final representation. Attention is defined as:

$$\text{Attention}(Q_r, K_r, V_r) = \text{softmax} \left(\frac{Q_r K_r^T}{\sqrt{d_k}} \right) V_r. \quad (6)$$

For simplicity, we use an MLP as the decoder function. Via the decoder and tanh, the output of the MHA layer is projected to a single-dimensional space and gets the final recognition result, $\Phi_r = [\phi_r^1, \phi_r^2, \dots, \phi_r^M] \in \mathbb{R}^{M \times 1}$.

3) *Decision Policy Architecture*: As mentioned in III-C, the decision policy's definition of vehicular elements ξ^v differs from that of the recognition policy. Hence, another DeepSet-based encoder embeds the vehicular elements into a 128-dimensional feature space. Next, we project the self-true SVO and the estimated SVOs of M surrounding agents into a 32-dimensional space using a linear layer, and the resulting features are concatenated with vehicular element features and passed through another MHA network, $\text{MultiHead}(Q_a, K_a, V_a)$. Unlike the recognition policy, we only use a single query $Q_a = [q_a] \in \mathbb{R}^{1 \times 160}$ by features of ego agent. Finally, the output of MHA is decoded into the action $a \in \mathcal{A}$.

IV. EXPERIMENTS

In this section, we pursue to answer several questions. (1) Can our recognition-based method achieve superior system-level performance? (2) Can our recognition framework successfully estimate agents' SVOs? Additionally, we investigate the factors that affect the accuracy of recognition.

A. Experimental Setup

We utilize the Universe simulator [34] to simulate bottleneck and merging scenarios. To model the motion of the vehicles in the simulator, we employ the Kinematic Bicycle Model and utilize a closed-loop proportional-integral-derivative (PID) controller to translate the actions into low-level steering and acceleration control signals. To allow for a continuous representation of action, we use the $a = [speed, heading] \in \mathbb{R}^2$ notation, with values bounded by the range of $[-1, 1]$, then mapped to the speed range of $[0, 6m/s]$ and the steering angle range of $[-\pi/4, \pi/4]$, respectively. During the training phase, in each episode, the agents are randomly spawned within a range of 8 to 20, and we randomly initialize their spawn points, global paths, and SVOs. We assumed vehicles have optimal conditions for map information, perception, localization, and control to focus on

planning during the simulation. In the testing phase, we fix the number of agents at 20. All experiments are performed on a computer with an Intel i9-12900KF CPU and NVIDIA GeForce RTX 3090.

B. Training

We use Independent Policy Learning (IPL) [35] for training the decision policy. To train the IPL within single-agent reinforcement learning, we utilize Soft-Actor-Critic (SAC) [36]. We train our recognition policy using the supervised learning approach. In particular, we execute the decision policy trained from the first stage in the environment by acquiring the true value of SVOs and subsequently use generated offline data to train the recognition policy. We use the Adam optimizer [37] to optimize both policies.

C. Metrics

Our experiments are evaluated based on measures of both efficiency and safety. We evaluate safety by calculating the percentage of an episode resulting in accidents, including the frequency of departures from the designated driving zone, collisions into the wrong lane, and driving too far from the global path. For brevity, we denote the above three types of accidents as “Crash”. To measure the recognition accuracy at the system level, we consider the mean deviation error between the multi-agent recognition values and the corresponding true values.

TABLE I: The table presents the percentage of various metrics (defined in section IV-C) for the bottleneck and merge scenarios, along with the performance of our proposed method indicated by a “†”.

Methods	Bottleneck		
	Success (↑)	Crash (↓)	Speed (↑)
MACAD [5]	76.1 ± 0.3	24.1 ± 0.6	75.1 ± 0.2
CoPO [6]	80.3 ± 0.6	20.7 ± 1.1	74.5 ± 0.3
TrueSVO	83.1 ± 0.4	16.9 ± 1.0	76.3 ± 0.2
Recog†	82.3 ± 0.3	17.3 ± 1.0	76.0 ± 0.1
Methods	Merge		
	Success (↑)	Crash (↓)	Speed (↑)
MACAD [5]	66.1 ± 0.6	34.0 ± 0.9	55.0 ± 0.2
CoPO [6]	69.3 ± 0.5	30.7 ± 0.9	54.9 ± 0.2
TrueSVO	82.9 ± 0.4	17.2 ± 0.6	60.0 ± 0.1
Recog†	81.8 ± 0.3	18.4 ± 0.7	59.6 ± 0.1

D. Performance of Multi-agent Driving System

We compare our proposed approach with two MARL-based baselines, MACAD [5] and CoPO [6]. MACAD is an approach that considers each agent aiming to maximize its reward. While CoPO incorporates SVO to promote coordination among agents at both local and global levels but does not know other agents’ SVO. Our approach, where the recognition policy estimates the SVOs and passes them to the decision policy referred to as Recog. We also take our decision policy as a comparison method denoting TrueSVO,

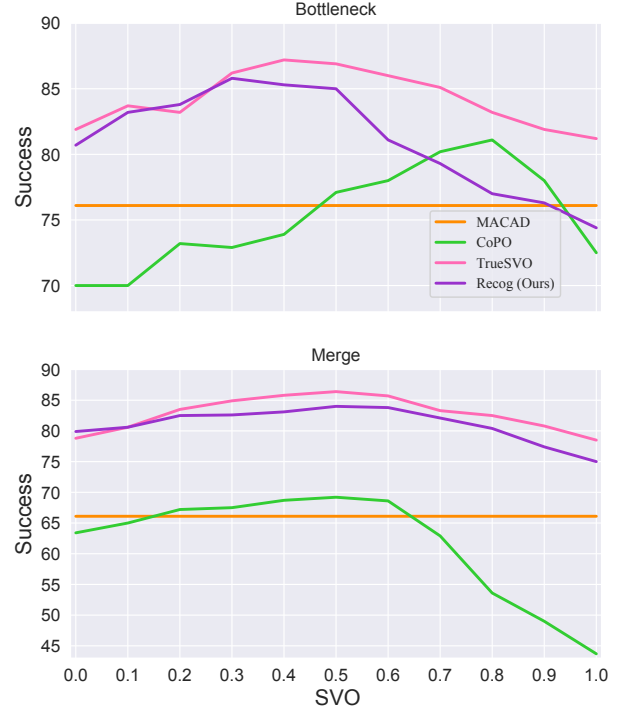


Fig. 3: Success rates. The figure shows the percentage of success rates in the bottleneck and merge. We assign a fixed SVO from 0 to 1 at regular intervals. All agents in traffic flows are given the same SVO. Each is evaluated for 200 episodes. MACAD actually does not have the concept of SVO, it is used here as a reference line.

where the TrueSVO receives true value of SVOs directly. Evaluation results shown in Table. I and Fig. 3, The MACAD approach produces individual egoistic behaviors, results in lower performance. While CoPO achieves better coordination and higher success rates in MADS with random SVOs but has lower average speeds across two scenarios. The Recog approach offers insights into other agents’ driving styles and performs better than MACAD and CoPO, but due to estimation errors, its performance is lower compared to TrueSVO. TrueSVO outperforms other approaches in all metrics, indicating that sharing driving attitude information leads to more effective and coordinated MADS.

E. Recognition Accuracy

1) *Impact of Agents Number*: As depicted in Figure 4, it can be observed that our policy demonstrates better precision convergence as the number of agents increases. When the number of agents decreases, the influence of individual driving behaviors on other agents reduces, and the demand for coordinated behaviors among agents is lower, making it more difficult for our strategy to identify the characters of the agents from the interaction. While the number increase, more interactions exist, thus the policy shows a rising performance of estimation of SVOs.

2) *Highly Interactive Period*: Fig. 5 presents a visualization of the scenes and the roll-out trajectories of one agent, with time period markings of high-interaction

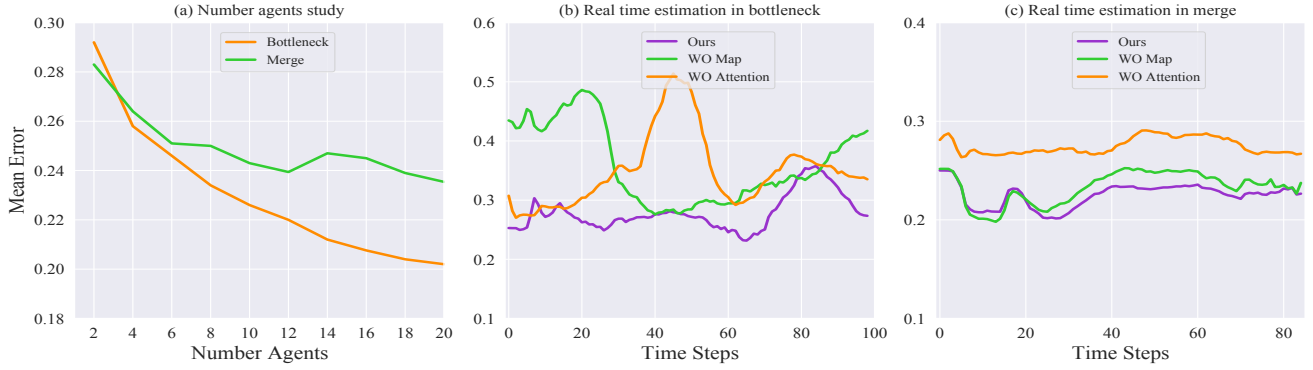


Fig. 4: Influences of accuracy. In (a), we change the number of agents in the bottleneck and merge. In (b) and (c), we change the architecture of our network, including removing the road information (WO Map) or attention model (WO Attention). We use the mean deviation error mentioned in IV-C as metric. Each is evaluated for 200 episodes.

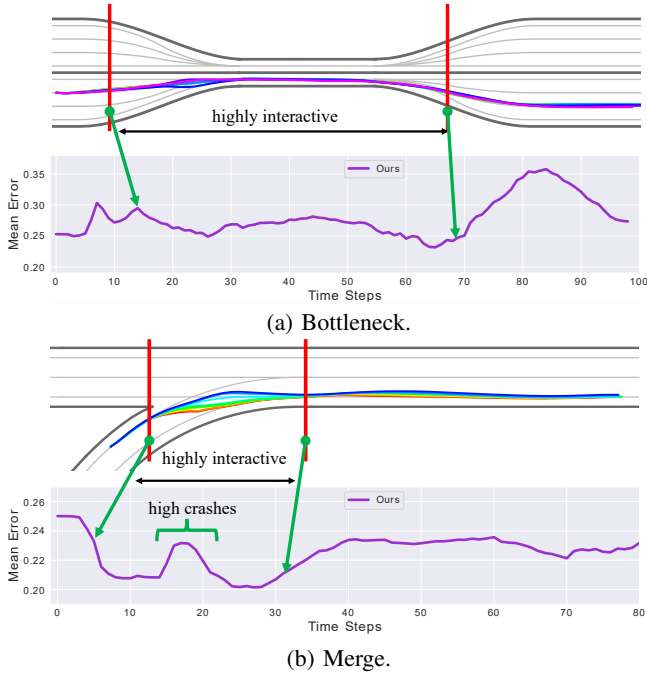


Fig. 5: Visualization of scenarios and one vehicle's trajectories.

environments. The result shows our policy achieves better accuracy in the highly interactive period. In the bottleneck, the recognition error rises at the beginning, as the agents exhibit little driving information. Furthermore, during times 15 and 70, the agents navigate through a highly interactive environment as each agent slows down into a narrow lane and interacts with others, exhibiting diverse behaviors, allowing the policy to estimate more accurately. In the merge scenario, the error declines from steps 0 to 30 because agents going straight need to avoid merging agents, providing more information for the recognition policy. However, the time steps around 18 are with high incidences of crashes, which causes increasing estimation error. In the second half of the time, the agents experience a less competitive road structure and exhibit more homogeneous

behaviors, leading to an increase in recognition error.

TABLE II: Label Interpretation.

	trajectories	road structures	attention network
Without attention	✓		
Without map (with attention)	✓		✓
With attention and map (Ours)	✓	✓	✓

3) *Importance of Attention-based Model:* Fig. 4 (b) and (c) show how the map information, such as lane and boundary data, impacts the recognition accuracy of SVOs. Table II explains the labels used in these figures. Though performing well in the merge, the recognition policy without map information performs badly in the bottleneck, due to the fact that the bottleneck has a more complex road structure that lasts for a longer time period, and the lack of map information makes it difficult for the policy to estimate the SVOs of the surrounding agents accurately. As for the policy without attention, it only knows the agents' trajectories and performs worst in both scenarios. Our method utilizes the attention model that combines map information with agents' trajectories to help to know the surrounding environments better and get the best performance. Although our method's accuracy decreases in the bottleneck during the last thirty time steps, this can be attributed to the fact that agents are in a less interactive environment, making it more challenging to estimate their characteristics accurately. These findings suggest that combining multiple sources of information can lead to more accurate SVOs recognition and enhance the effectiveness of autonomous driving systems.

V. CONCLUSIONS

This paper focuses on the challenge of designing a safe and efficient MADS and introduces a novel social preference recognition framework to handle complex driving environments. The framework can integrate multiple sources of information to achieve more accurate social preference recognition. We propose a two-stage method for MADS, which comprises a recognition policy and a decision policy that are seamlessly integrated. We evaluate our method on

two complex scenarios, namely bottleneck and merge, and compare its performance with other MARL-based methods. The results demonstrate that sharing SVOs can lead to better performance of MADs, highlighting the effectiveness of our approach.

REFERENCES

- [1] C. Wu, A. R. Kreidieh, K. Parvate, E. Vinitsky, and A. M. Bayen, "Flow: A modular learning framework for mixed autonomy traffic," *IEEE Transactions on Robotics*, 2021.
- [2] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Physical review E*, vol. 62, no. 2, p. 1805, 2000.
- [3] A. Kesting, M. Treiber, and D. Helbing, "General lane-changing model mobil for car-following models," *Transportation Research Record*, vol. 1999, no. 1, pp. 86–94, 2007.
- [4] S. Shalev-Shwartz, S. Shammah, and A. Shashua, "Safe, multi-agent, reinforcement learning for autonomous driving," *arXiv preprint arXiv:1610.03295*, 2016.
- [5] P. Palanisamy, "Multi-agent connected autonomous driving using deep reinforcement learning," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–7.
- [6] Z. Peng, Q. Li, K. M. Hui, C. Liu, and B. Zhou, "Learning to simulate self-driven particles system with coordinated policy optimization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 10 784–10 797, 2021.
- [7] B. Toghi, R. Valiente, D. Sadigh, R. Pedarsani, and Y. P. Fallah, "Altruistic maneuver planning for cooperative autonomous vehicles using multi-agent advantage actor-critic," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021.
- [8] W. Schwarting, A. Pierson, J. Alonso-Mora, S. Karaman, and D. Rus, "Social behavior for autonomous vehicles," *Proceedings of the National Academy of Sciences*, vol. 116, no. 50, pp. 24 972–24 978, 2019.
- [9] H. Wang, W. Wang, S. Yuan, X. Li, and L. Sun, "On social interactions of merging behaviors at highway on-ramps in congested traffic," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 11 237–11 248, 2021.
- [10] J. Wang, T. Shi, Y. Wu, L. Miranda-Moreno, and L. Sun, "Multi-agent graph reinforcement learning for connected automated driving," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020, pp. 1–6.
- [11] B. Liu, Z. Ding, and C. Lv, "Platoon control of connected autonomous vehicles: A distributed reinforcement learning method by consensus," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 15 241–15 246, 2020.
- [12] B. Toghi, R. Valiente, D. Sadigh, R. Pedarsani, and Y. P. Fallah, "Cooperative autonomous vehicles that sympathize with human drivers," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 4517–4524.
- [13] —, "Social coordination and altruism in autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 12, pp. 24 791–24 804, 2022.
- [14] Z. Dai, T. Zhou, K. Shao, D. H. Mguni, B. Wang, and H. Jianye, "Socially-attentive policy optimization in multi-agent self-driving system," in *6th Annual Conference on Robot Learning*.
- [15] D. M. Saxena, S. Bae, A. Nakhaei, K. Fujimura, and M. Likhachev, "Driving in dense traffic with model-free reinforcement learning," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 5385–5392.
- [16] A. Mavrogiannis, R. Chandra, and D. Manocha, "B-gap: Behavior-rich simulation and navigation for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4718–4725, 2022.
- [17] L. Wang, L. Sun, M. Tomizuka, and W. Zhan, "Socially-compatible behavior design of autonomous vehicles with verification on real human data," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3421–3428, 2021.
- [18] Z. Huang, J. Wu, and C. Lv, "Driving behavior modeling using naturalistic human driving data with inverse reinforcement learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 10 239–10 251, 2021.
- [19] L. Sun, W. Zhan, M. Tomizuka, and A. D. Dragan, "Courteous autonomous cars," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 663–670.
- [20] L. Crosato, H. P. Shum, E. S. Ho, and C. Wei, "Interaction-aware decision-making for automated vehicles using social value orientation," *IEEE Transactions on Intelligent Vehicles*, 2022.
- [21] X. Zhao, Y. Tian, and J. Sun, "Yield or rush? social-preference-aware driving interaction modeling using game-theoretic framework," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 453–459.
- [22] C. Li, T. Trinh, L. Wang, C. Liu, M. Tomizuka, and W. Zhan, "Efficient game-theoretic planning with prediction heuristic for socially-compliant autonomous driving," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 248–10 255, 2022.
- [23] N. Buckman, W. Schwarting, S. Karaman, and D. Rus, "Semi-cooperative control for autonomous emergency vehicles," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 7052–7059.
- [24] S. Hoermann, D. Stumper, and K. Dietmayer, "Probabilistic long-term prediction for autonomous vehicles," in *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2017, pp. 237–243.
- [25] R. P. Bhattacharyya, R. Senanayake, K. Brown, and M. J. Kochenderfer, "Online parameter estimation for human driver behavior prediction," in *2020 American Control Conference (ACC)*. IEEE, 2020, pp. 301–306.
- [26] S. Le Cleac'h, M. Schwager, and Z. Manchester, "Lucidgames: Online unscented inverse dynamic games for adaptive trajectory prediction and planning," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5485–5492, 2021.
- [27] F. A. Oliehoek and C. Amato, *A concise introduction to decentralized POMDPs*. Springer, 2016.
- [28] N. Buckman, A. Pierson, W. Schwarting, S. Karaman, and D. Rus, "Sharing is caring: Socially-compliant autonomous intersection negotiation," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 6136–6143.
- [29] J. K. Terry, N. Grammel, A. Hari, L. Santos, and B. Black, "Revisiting parameter sharing in multi-agent deep reinforcement learning," *arXiv preprint arXiv:2005.13625*, 2020.
- [30] J. Gao, C. Sun, H. Zhao, Y. Shen, D. Anguelov, C. Li, and C. Schmid, "Vectornet: Encoding hd maps and agent dynamics from vectorized representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 525–11 533.
- [31] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola, "Deep sets," *Advances in neural information processing systems*, vol. 30, 2017.
- [32] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 213–229.
- [33] O. Scheel, L. Bergamini, M. Wolczyk, B. Osiński, and P. Ondruska, "Urban driver: Learning to drive from real-world demonstrations using policy gradients," in *Conference on Robot Learning*. PMLR, 2022, pp. 718–728.
- [34] D. Zhang, "Universe," 2 2023. [Online]. Available: <https://github.com/alibaba-damo-academy/universe>
- [35] M. Tan, "Multi-agent reinforcement learning: Independent vs. cooperative agents," in *Proceedings of the tenth international conference on machine learning*, 1993, pp. 330–337.
- [36] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*. PMLR, 2018, pp. 1861–1870.
- [37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015.