Cooperative Autonomous Vehicles that Sympathize with Human Drivers

Behrad Toghi[1], Rodolfo Valiente[1], Dorsa Sadigh[2], Ramtin Pedarsani[3], Yaser P. Fallah[1]

***Abstract*— Widespread adoption of autonomous vehicles will not become a reality until solutions are developed that enable these intelligent agents to co-exist with humans. This includes safely and efficiently interacting with human-driven vehicles, especially in both conflictive and competitive scenarios. We build up on the prior work on socially-aware navigation and borrow the concept of social value orientation from psychology —that formalizes how much importance a person allocates to the welfare of others— in order to induce altruistic behavior in autonomous driving. In contrast with existing works that explicitly model the behavior of human drivers and rely on their expected response to create opportunities for cooperation, our Sympathetic Cooperative Driving (SymCoDrive) paradigm trains altruistic agents that realize safe and smooth traffic flow in competitive driving scenarios only from experiential learning and without any explicit coordination. We demonstrate a significant improvement in both safety and traffic-level metrics as a result of this altruistic behavior and importantly conclude that the level of altruism in agents requires proper tuning as agents that are too altruistic also lead to sub-optimal traffic flow. The code and supplementary material are available at: https://symcodrive.github.io/**

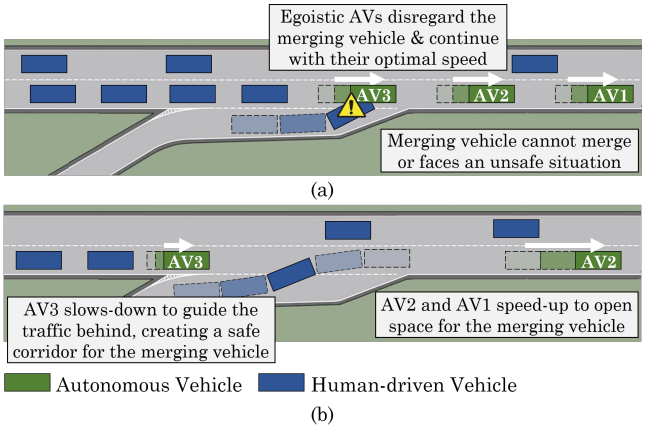

Fig. 1: Seamless and safe highway merging requires all AVs working together and accounting for the human-driven vehicles' utility. *(top)* Egoistic AVs solely optimize for their own utility, *(bottom)* Altruistic AVs compromise on their welfare to account for the human-driven vehicles.

I. Introduction

The next generation of transportation systems will be safer and more efficient with connected autonomous vehicles. Vehicle-to-vehicle (V2V) communication enables autonomous vehicles (AVs) to constitute a form of mass intelligence and overcome the limitations of a single agent planning in a decentralized fashion [**?**]. If all vehicles on the road were connected and autonomous, V2V could allow them to coordinate and handle complex driving scenarios that require selflessness, e.g., merging to and exiting a highway, and crossing intersections. However, a road shared by AVs and human-driven vehicles (HVs) naturally becomes a competitive scene due to their different levels of maneuverability and reaction time. In contrast with the full-autonomy case, here the coordination between HVs and AVs is not as straightforward since AVs do not have an explicit means of harmonizing with humans and therefor require to locally account for the other HVs and AVs in their proximity.

To further elaborate on this need, assume the merging scenario depicted in Figure 1. The merging vehicle, either HV or AV, faces a mixed group of AVs and HVs on the highway and needs them to slow-down to allow it to merge. If AVs act selfishly, it will be up to the HVs in the highway to allow for merging. Relying only on the human drivers can lead to sub-optimal or even unsafe situations due to their hard-to-predict and differing behaviors. In this particular example, assuming egoistic AVs, the merging vehicle will either get stuck in the merging ramp and not be able to merge or will wait for an HV and risk on cutting into the highway without knowing if the HV will slow-down or not. On the other hand, altruistic AVs can work together and guide the traffic on the highway, e.g., by slowing down the vehicles behind as AV3 does in Figure 1(b), in order to enable a seamless and safe merging. Such altruistic autonomous agents can create societally desirable outcomes in conflictive driving scenarios, without relying on or making assumptions about the behavior of human drivers.

Altruistic behavior of autonomous cars can be formalized by quantifying the willingness of each vehicle to incorporate the utility of others, whether an HV or an AV, into its local utility function. This notion is defined as *social value orientation (SVO)*, which has recently been adopted from the psychology literature to robotics and artificial intelligence research [**?**]. SVO determines the degree to which an agent acts egoistic or altruistic in the presence of others. Figure 1(b) demonstrates an example of altruistic behavior by AVs where they create a safe corridor for the merging HV and enable a seamless merging. In a mixed-autonomy scenario, agents either are homogeneous with the same SVO or can directly obtain each other's SVO (via V2V). However, the utility and

*This material is based upon work supported by the National Science Foundation under Grant No. CNS-1932037.

[1]Connected & Autonomous Vehicle Research Lab (CAVREL), University of Central Florida, Orlando, FL, USA. toghi@knights.ucf.edu

[2] Intelligent and Interactive Autonomous Systems Group (ILIAD), Stanford University, Stanford, CA, USA.

[3] Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA, USA.

SVO of an HV are unknowns, as they are subjective and inconstant and therefore cannot be communicated to the AVs.

The existing social navigation works model a human driver's SVO either by predicting their behavior [**?**] and avoiding conflicts with them or relying on the assumption that humans are naturally willing or can be incentivized to cooperate [**?**]. By explicitly modeling human behavior, agents can exploit cooperation opportunities in order to achieve a social goal that favors both humans and autonomous agents. However, modeling human behaviors is often challenging due to time-varying changes in the model affected by fatigue, distraction, and stress as well as scalability of belief modeling techniques over other agent's behaviors, hence limiting the practicality of the above approach. Methods based on model-predictive control (MPC) generally require an engineered cost function and a centralized coordinator [**?**]. As such, they are not suitable for cooperative autonomous driving, where central coordination is not viable. On the other hand, data-driven solutions such as reinforcement learning are challenged in mixed-autonomy multi-agent systems, mainly due to the non-stationary environment in which agents are evolving concurrently.

Considering these shortcomings, the notion of altruism in AVs can be divided into *cooperation* within autonomous agents and *sympathy* among autonomous agents and human drivers. Dissociating the two components helps us to separately probe their influence on achieving a social goal. Our key insight is that defining a social utility function can induce altruism in decentralized autonomous agents and incentivize them to cooperate with each other and to sympathize with human drivers with no explicit coordination or information about the humans' SVO. The core differentiating idea that we rely on is that AVs trained to reach an optimal solution for all vehicles, learn to implicitly model the decision-making process of humans only from experience. We study the behavior of altruistic AVs in scenarios that would turn into safety threats if either of sympathy and cooperation components is absent. In other words, we perform our experiments in scenarios with a similar nature to the one depicted in Figure 1 that essentially require all agents to work together and success cannot be achieved by any of them individually. Our main contributions are as follows:

- We propose a data-driven framework, Sympathetic Cooperative Driving (SymCoDrive), that incorporates a decentralized reward structure to model cooperation and sympathy and employ a 3D convolutional deep reinforcement learning (DRL) architecture to capture the temporal information in driving data,
- We demonstrate how tuning the level of altruism in AVs leads to different emerging behaviors and affects the traffic flow and driving safety,
- We experiment with a highway merging scenario and demonstrate that our approach results in improved driving safety and societally desirable behaviors compared to egoistic autonomous agents.

II. Related Work

Multi-agent Reinforcement Learning. A major challenge for multi-agent reinforcement learning (MARL) is the inherent non-stationarity of the environment. Foerster et. al. suggest a novel learning rule to address this issue [**?**]. Additionally, the idea of decorrelating training samples by drawing them from an experience replay buffer becomes obsolete and a multi-agent derivation of importance sampling can be employed to remove the outdated samples from the replay buffer [**?**]. Xie et. al. have also attempted to mitigate this problem by using latent representations of partner strategies to enable a more scalable MARL and partner modeling [**?**].

The *counterfactual multi-agent (COMA)* algorithm proposed by Foerster et. al. uses a centralized critic and decentralized actors to tackle the problem of credit assignment in multi-agent environments [**?**]. In the case of centralized control, deep Q-networks with full observability over the environment can be used to control the joint-actions of a group of agents [**?**]. Within the context of mixed-autonomy, the existing literature focuses on solving cooperative and competitive problems by making assumptions on the nature of interactions between autonomous agents (or autonomous agents and humans) [**?**]. Contrary to these works, we assume partial observability and a decentralized reward function and aim to train sympathetic cooperative autonomous agents with no assumption on humans' behavior.

Autonomous Driving in Mixed-autonomy. Driving styles of humans can be learned from demonstration through inverse RL or employing statistical models [**?**]. Modeling human driver behavior assists autonomous vehicles to identify potentials for creating cooperation and interaction opportunities with humans in order to realize safe and efficient navigation [**?**]. Moreover, human drivers are able to intuitively anticipate next actions of neighboring vehicles through observing slight changes in their trajectories and leverage the prediction to move proactively if required. Inspired by this fact, Sadigh et. al. reveal how autonomous vehicles can exploit this farsighted behavior of humans to shape and affect their actions [**?**]. On a macro-traffic level, prior works have demonstrated emerging human behaviors within mixed-autonomy scenarios and studied how these patterns can be utilized to control and stabilize the traffic flow [**?**], [**?**]. Closely related to our topic, recent works in social robot navigation have shown the potential for collaborative planning and interaction with humans as well [**?**], [**?**], [**?**]. In contrast, we seek altruistic behaviors without having an explicit model of the human driver's behavior or relying on their aid.

III. Preliminaries

Partially Observable Stochastic Games (POSG). We formulate the problem of multi-vehicle interaction using a stochastic game defined by the tuple $\mathcal{M}_{\mathrm{G}} := (\mathcal{I}, \mathcal{S}, [\mathcal{A}_i], [\mathcal{O}_i], P, [r_i])$ for $i = 1, \ldots, N$, in which $\mathcal{I}$ is a finite set of agents and $\mathcal{S}$ represents the state-space including all possible formations that the N agents can adopt. At a given time the agent receives a local observation $o_i : \mathcal{S} \rightarrow \mathcal{O}_i$

and takes an action within the action-space $a_i \in \mathcal{A}_i$ based on a stochastic policy $\pi_i : \mathcal{O}_i \times \mathcal{A}_i \to [0, 1]$. Consequently, the agent transits to a new state s_i' which is determined based on the state transition function $P(s'|s, a) : \mathcal{S} \times \mathcal{A}_1 \times ... \times \mathcal{A}_N \to \mathcal{S}$ and receives a reward $r_i : \mathcal{S} \times \mathcal{A}_i \to \mathbb{R}$. The goal is to derive an optimal policy π^* that maximizes the discounted sum of future rewards over an infinite time horizon.

In a partially-observable stochastic game (POSG), the state transition and reward functions are usually not known and an agent only has access to a local observation which is correlated with the state. Employing multi-agent reinforcement learning, independent MARL agents can work together to overcome the physical limitations of a single agent and outperform them [**?**]. In a multi-vehicle problem, controlling vehicles by a centralized MARL controller that has full observability over the environment and assigns a centralized joint reward ($\forall i, j : r_i \equiv r_j$) to all vehicles is rather straightforward. However, such assumptions are not feasible in real-world autonomous driving applications and we rather focus on the decentralized case where vehicles have partial observability and are not aware of each other's actions. Coordination among agents in such settings is expected to arise from the decentralized reward function that we introduce which uses the local observations to estimate the utility of other vehicles.

Deep Q-networks (DQN). Q-learning, which has been widely applied in reinforcement learning problems with large state-spaces, defines a state-value function $Q^\pi(s, a) := \mathbb{E}[\sum_{i=1}^{\infty} \gamma^i r(s_i, \pi(s_i))|s_0 = s, a_0 = a]$ to derive the optimal policy $\pi^*(s) = \arg\max_a Q^*(s, a)$ where $\gamma \in [0, 1)$ is a discount factor. DQN [**?**] uses a neural network with weights $\mathbf{w}$ to estimate the state-action value function by performing mini-batch gradient descent steps as $\mathbf{w}_{i+1} = \mathbf{w}_i + \alpha_i \hat{\nabla}_{\mathbf{w}} \mathcal{L}(\mathbf{w}_i)$, where the loss function is defined as

$$\mathcal{L}(\mathbf{w}_i) = \mathbb{E}[(r + \gamma \max_a Q^*(s', a'; \mathbf{w}^\circ) - Q^*(s, a; \mathbf{w})^2] \quad (1)$$

and the $\hat{\nabla}_{\mathbf{w}}$ operator is an estimate of the gradient at $\mathbf{w}_i$ and $\mathbf{w}^\circ$ is the target network's weights which get updated periodically in training. Sets of (s, a, r, s') are randomly drawn from an *experience replay buffer* to de-correlate the training samples in Equation (1). This experience replay process particularly becomes problematic in the case of MARL since, $P(s'|s, a, \pi_1, ..., \pi_N) \neq P(s'|s, a, \pi_1', ..., \pi_N')$ if $\exists i.\pi_i \neq \pi_i'$, i.e., if any of the agents' policies evolve during the training.

IV. Sympathetic Cooperative Driving

Highway Merging Scenario. Our base scenario is a highway merging ramp where a merging vehicle (either HV or AV) attempts to join a mixed platoon of HVs and AVs, as illustrated in Figure 1. We specifically choose this scenario due to its inherent competitive nature, since the local utility of the merging vehicle is conflictive with that of the cruising vehicles. We ensure that only one AV yielding to the merging vehicle will not make the merge possible and for it to happen, essentially all AVs require to work together. In Figure 1(b), AV3 must slow down and guide the vehicles in behind, which perhaps are not able to see the merging vehicle, while AV2 and AV1 speed-up to open space for the merging vehicle. If any of the vehicles do not cooperate or act selfishly, traffic safety and efficiency will be compromised.

Formalism. Consider a road section as shown in Figure 1 with a set of autonomous vehicles $\mathcal{I}$, a set of human-driven vehicles $\mathcal{V}$, and a *mission vehicle*, $M \in \mathcal{I} \cup \mathcal{V}$ that can be either AV or HV and is attempting to merge into the highway. HVs normally have a limited perception range restricted by occlusion and obstacles. In the case of AVs, although we assume no explicit coordination and no information about the actions of the others, autonomous agents are connected through V2V communication which allows them to share their situational awareness. Leveraging this extended situational awareness, agents can broaden their range of perception and overcome occlusion and line-of-sight visibility limitations. Therefore, while each AV has a unique partial observation of the environment, they can see all vehicles within their extended perception range, i.e., they can see a subset of AVs $\widetilde{\mathcal{I}} \subset \mathcal{I}$, and a subset of HVs $\widetilde{\mathcal{V}} \subset \mathcal{V}$.

In order to model a mixed-autonomy scenario, we deploy a mixed group of HVs and AVs to cruise on a highway and target to maximize their speed while maintaining safety. The contrast between humans and autonomous agents is that humans are solely concerned about their own safety while the altruistic autonomous agents attempt to optimize for the safety and efficiency of the group. Social value orientation gauges the level of altruism in an agent's behavior. In order to systematically study the interaction between agents and humans, we decouple the notion of *sympathy* and *cooperation* in SVO. Specifically, we consider the altruistic behavior of an agent with humans as sympathy and refer to the altruistic behavior among agents themselves as cooperation. One rationale behind this definition is the fact that the two are different in nature as the sympathetic behavior can be one-sided when humans are not necessarily willing to help the agents. Cooperation, however, is a symmetric quality since the same policy is deployed in all AVs and as we will see in our experiments, social goal of the group can be achieved regardless of the humans' willingness to cooperate.

Decentralized Reward Structure. The local reward received by agent $I_i \in \mathcal{I}$ can be decomposed to

$$\begin{aligned} R_i(s_i, a_i) &= R^E + R^C + R^S \\ &= \lambda^E r_i^E(s_i, a_i) + \\ &\quad \lambda^C \sum_j r_{i,j}^C(s_i, a_i) + \lambda^S \sum_k r_{i,k}^S(s_i, a_i) \end{aligned} \quad (2)$$

in which $i \in \widetilde{\mathcal{I}}$, $j \in \widetilde{\mathcal{I}} \setminus \{I_i\}$, $k \in (\widetilde{\mathcal{V}} \cup \{M\}) \setminus (\mathcal{I} \cap \{M\})$. The level of altruism or egoism can be tuned by λ^E, λ^C, and λ^S coefficients. The r_i^E component in Equation (2) denotes the local driving performance reward derived from metrics such as distance traveled, average speed, and a negative cost for changes in acceleration to promote a smooth and efficient movement by the vehicle. The cooperative reward term, $r_{i,j}^C$ accounts for the utility of the observer agent's allies, i.e.,

other AVs in the perception range except for I_i. It is important to note that I_i only requires the V2V information to compute R^C and not any explicit coordination or knowledge of the actions of the other agents. The sympathetic reward term, $r^S_{i,k}$ is defined as

$$r^S_{i,k} = r^M_k + \sum_k \frac{1}{\eta d^\gamma_{i,k}} u_k, \tag{3}$$

where u_k denotes an HV's utility, e.g., its speed, $d_{i,k}$ is the distance between the observer autonomous agent and the HV, and η and γ are dimensionless coefficients. Moreover, the sparse scenario-specific *mission reward* term r^M_k in the case of our driving scenario is representing the success or failure of the merging maneuver, formally

$$r^M_k = \begin{cases} 1, & \text{if } V_k \text{ is the mission vehicle and has merged} \\ 0, & \text{o.w.} \end{cases} \tag{4}$$

During training, each agent optimizes for this decentralized reward function using Deep RL and learns to drive on the highway and work with its allies to create societally desirable formations that benefits both AVs and HVs.

State-space and Action-space. The robot navigation problem can be viewed from multiple levels of abstraction: from the low-level continuous control problem to the higher level meta-action planning. Our purpose in this work is to study the inter-agent and agent-human interactions as well as the behavioral aspects of mixed-autonomy driving. Thus, we choose a more abstract level and define the action-space as a set of discrete meta-actions $a_i \in \mathbb{R}^n$.

We experiment with two different local state representations to find the most suitable one for our problem. The *multi-channel VelocityMap* representation separates AVs and HVs into two channels and embeds their relative speed in the pixel values. Figure 2 illustrates an example of this multi-channel representation. A clipped logarithmic function is used to map the relative speed of the vehicles into pixel values as it showed a better performance compared to the linear mapping, i.e.,

$$Z_j = 1 - \beta \log(\alpha |v^{(l)}_j|) \mathbb{1}(|v^{(l)}_j| - v_0) \tag{5}$$

where Z_j is the pixel value of the jth vehicle in the state representation, $v^{(l)}$ is its relative Frenet longitudinal speed from the kth vehicle's point-of-view, i.e., $\dot{l_j} - \dot{l_k}$, v_0 is speed threshold, α and β are dimensionless coefficients, and $\mathbb{1}(.)$ is the Heaviside step function. Such non-linear mapping gives more importance to neighboring vehicles with smaller $|v^{(l)}|$ and almost disregards the ones that are moving either much faster or much slower than the ego. We add three more channels that embed 1) the road layout, 2) an attention map to emphasize on the location of the ego, and 3) the mission vehicle.

The other candidate is an *occupancy grid* representation that directly embeds the information as elements of a 3-dimensional tensor $o_i \in \mathcal{O}_i$. Theoretically, this representation is very similar to the previous VelocityMap and what contrasts them is that the occupancy grid removes the shapes

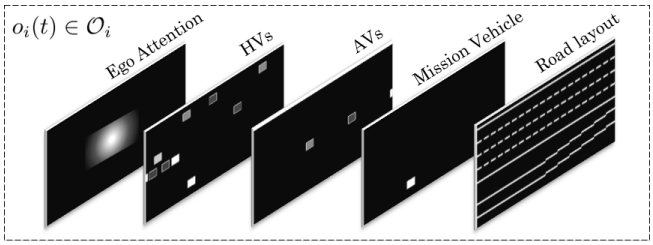

Fig. 2: Multi-channel VelocityMap state representation embeds the speed of the vehicle in pixel values.

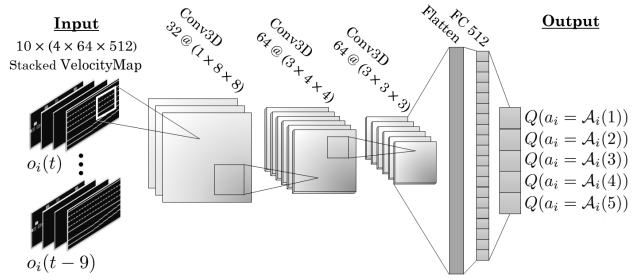

Fig. 3: Our deep Q-network with 3D Convolutional Architecture.

and visual features such as edges and corners and directly feeds the network with sparse numbers. More specifically, consider a tensor of size $W \times H \times F$, in which the nth channel is a $W \times F$ matrix defined as

$$o_{(n,,)} \in \mathbb{R}^2 = \begin{cases} f(n), & \text{if } f(1) = 1 \\ 0, & \text{o.w.} \end{cases} \tag{6}$$

where $f = [p, l, d, v^{(l)}, v^{(d)}, \sin\delta, \cos\delta]$ is the feature set, p is a binary variable showing the presence of a vehicle, l and d are relative Frenet coordinates, $v^{(l)}$ and $v^{(d)}$ are relative Frenet speeds, and δ is the yaw angle measured with respect to a global reference.

Training with Deep MARL. We experiment with 2 existing architectures proposed in the literature by Mnih et. al. and Egorov et. al. [**?**], [**?**] in addition to a simple 2-layer multi-layer perceptron (MLP) network as function approximators for our Q-learning problem. Additionally, we implemented a 3D convolutional network that captures the temporal dependencies in a training episode as shown in Figure 3. The input to our network is a stack of 10 VelocityMap observations, i.e., a $10 \times (4 \times 512 \times 64)$ tensor, which capture the last 10 time-steps in the episode. The performance of the architectures is compared in Section V-G.

We train a single neural network offline and deploy the learned policy into all agents for distributed independent execution in real-time. In order to cope with the non-stationarity issue in MARL, agents are trained in a semi-sequential manner, as illustrated in Figure 4. Each agent is trained separately for k episodes while the policies of its allies, $\mathbf{w}^-$, are frozen. The new policy, $\mathbf{w}^+$, is then disseminated to all agents to update their neural networks.

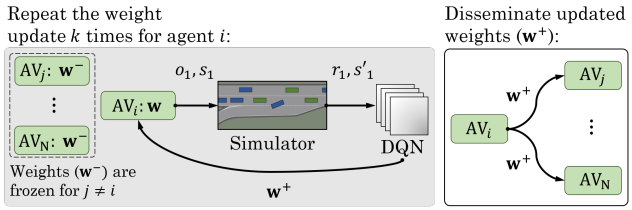

Fig. 4: The multi-agent training and policy dissemination process.

Additionally, inspired by [**?**], we employ a novel experience replay mechanism to compensate for our highly skewed training data. A training episode can be semantically divided into two sections, cruising on a straight highway and highway merging. The ratio of the latter to the former in the experience replay buffer is a small number since the latter occurs in only a short time period of each episode. Consequently, uniformly sampling from the experience replay buffer leads to too few training samples relating to highway merging. Instead, we set the probability of a sample being drawn from the buffer proportional to its last resulted reward and its spatial distance with the merging point on the road. Balancing skewed training datasets is a common practice in computer vision and machine learning and appeared to be beneficial in our MARL problem as well.

V. Experiments

A. Driving Simulator Setup

We customize an OpenAI Gym environment [**?**] to simulate the highway driving and merging scenarios. In the framework of our simulator, a Kinematic Bicycle Model describes the motion of the vehicles and a closed-loop proportional–integral–derivative (PID) controller is employed for translating the meta-actions to low-level steering and acceleration control signals. Particularly, we choose a set of $n = 5$ abstract actions as $a_i \in \mathcal{A}_i = [\texttt{Lane Left}, \texttt{Idle}, \texttt{Lane Right}, \texttt{Accelerate}, \texttt{Decelerate}]^T$. As a common practice in the autonomous driving space, we express road segments and vehicles' motion in the Frenet-Serret coordinate frame which helps us to take the road curvature out of our equations and break-down the control problem to lateral and longitudinal components. In our simulated environment, the behavior of HVs is governed by lateral and longitudinal driver models proposed by Treiber et al. and Kesting et al [**?**], [**?**].

In order to ensure the generalization capability of our learned policies, we draw the initial position of all vehicles from a clipped Gaussian distribution with mean and variance tuned to ensure that the initialized simulations fall into our desired merging scenario configuration. We further randomize the speed and initial position of the vehicles during the testing phase to probe the agents' ability to handle unseen and more challenging cases.

B. Computational Details

A single training iteration in the PyTorch implementation of SymCoDrive takes about 440ms using a NVIDIA Tesla V100 GPU and a Xeon 6126 CPU @ 2.60GHz. We have repeated the training process multiple times to ensure all runs converge to similar emerging behaviors and policy. Training the Conv3D network for 15,000 episodes took approximately 33 hours on our hardware. The policy execution frequency is set to 1Hz and an online query of the network in the testing phase takes approximately 10ms. We spent ∼4,650 GPU-hours to tune the neural networks and reward coefficients for the purpose of our experiments.

C. Independent Variables

We conducted a set of experiments to study how *sympathy* and *cooperation* components of the reward function impact the behavior of autonomous agents and the overall safety/efficiency metrics. We compare the case in which the mission vehicle—merging vehicle in the example in Fig. 1—is *autonomous* to its dual scenario with a *human-driven* mission vehicle. We define 2x4 settings, in which the mission vehicle is either an AV or HV, and the other autonomous agents follow an egoistic, cooperative-only, sympathetic-only, or sympathetic cooperative objectives:

- **HV+E.** The mission vehicle is *human-driven* and autonomous agents act *egoistically*,
- **HV+C.** The mission vehicle is *human-driven* and autonomous agents only have a *cooperation* component (R^C) in their reward,
- **HV+S.** The mission vehicle is *human-driven* and autonomous agents only have the *sympathy* (R^S) element,
- **HV+SC.** The mission vehicle is *human-driven* and autonomous agents have both *sympathy* (R^S) and *cooperation* (R^C) components in their reward,
- **AV+E/C/S/SC.** Similar to the cases above with the difference of mission vehicle being *autonomous*.

D. Dependent Measures

Performance of our experiments can be gauged in terms of efficiency and safety. The average distance traveled by each vehicle within the duration of a simulation episode is a traffic-level measure for efficiency. The percentage of the episodes that experienced a crash indicates the safety of the policy. Counting the number of scenarios with no crashes and successful missions (merging to the highway) gives us an idea about our solution's overall efficacy.

E. Hypotheses

We examine three key hypotheses:

- **H1.** *In the absence of both cooperation and sympathy, a HV will not be able to safely merge into the highway. Thus, we anticipate to witness a better performance in **HV+SC** compared to **HV+C** and **HV+E**.*
- **H2.** *An autonomous mission vehicle only requires altruism from its allies to successfully merge. We do not expect to see a significant difference between **AV+SC** and **AV+C** scenarios; however, we hypothesize that they both will outperform **AV+E**.*
- **H3.** *Tuning the level of altruism in agents leads to different emerging behaviors that contrast in their impact on efficiency and safety. Increasing the level of altruism*

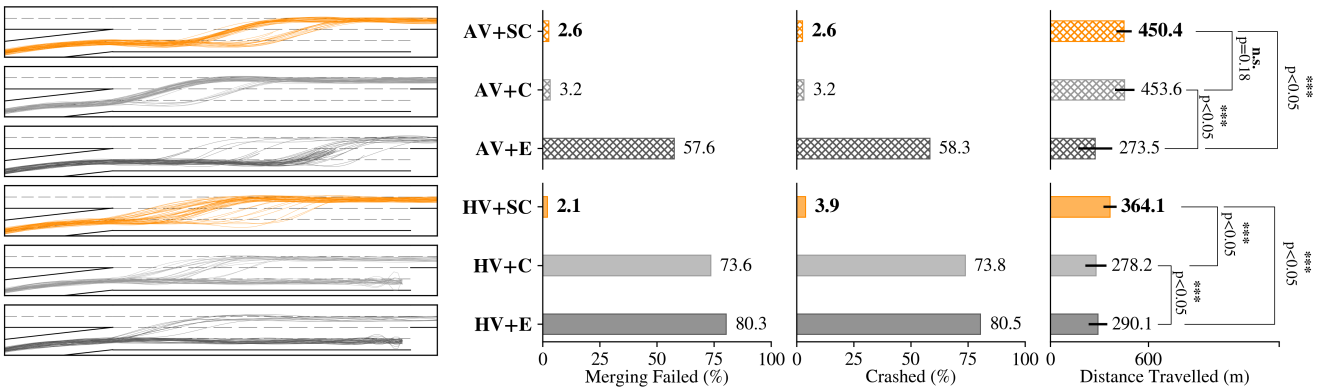

Fig. 5: Comparison between egoistic, cooperative-only, and sympathetic cooperative autonomous agents and how they interact with an autonomous (*top*) or human-driven (*bottom*) mission vehicle. A set of sampled mission vehicle's trajectories are illustrated on the left-side, relating to each of the 6 experiment setups defined in Section V-C.

can become self-defeating as it jeopardizes the agent's ability to learn the basic driving skills.

F. Results

We train SymCoDrive agents for 15,000 episodes in randomly initialized scenarios with a small standard deviation and average the performance metrics over 3,000 test episodes with 4x larger initialization range to ensure that our agents are not over-fitting on the seen training episodes.

1) Cooperation & Sympathy: To examine our hypothesis **H1**, we focus on scenarios with a human-driven mission vehicle, i.e., **HV+E**, **HV+C**, and **HV+SC**. The bottom row in Figure 5 illustrates our observations for these scenarios. It is evident that agents that integrate cooperation and sympathy elements (**SC**) in their reward functions show superior performance compared to solely cooperative (**C**) or egoistic (**E**) agents. This insight is also reflected in the bar plots that measure the average distance traveled by vehicles on the bottom right-most side. As a result of fair and efficient traffic flow, vehicles in the **HV+SC** scenario clearly succeed to travel a longer distance whereas in the **HV+C** and **HV+E** scenarios failed merging attempts and possible crashes deteriorate the performance. The left-most column in Figure 5 visualizes a set of sampled mission vehicle trajectories. It is clear that in the majority of episodes, cooperative sympathetic agents successfully merge to the highway while the other (**C**) and (**E**) agents fail in most of their attempts. Figure 6 provides further intuition on our discussion by comparing a set of mission vehicle's trajectories extracted from a **HV+E** scenario to the trajectories from the **HV+SC** scenario. Evidently, cooperative sympathetic agents enable successful merging while the other egoistic and solely-cooperative agents fail to do so, supporting our hypothesis **H1**.

It is imperative to repeat the experiments above for scenarios with an autonomous mission vehicle as one can argue that the failed missions and crashes in **HV+C** and **HV+E** are due to inadequacy of the driver model we have chosen for HVs. To precisely address this argument, **AV+E**, **AV+C**, and **AV+SC** scenarios are illustrated in the top row of Figure 5. First, a comparison between two scenarios with egoistic agents, i.e., **AV+E** and **HV+E**, unveils that an autonomous mission vehicle acts more creatively and explores different ways of merging to the highway, hence the more spread trajectory samples in **AV+E** compared to **HV+E**. Next, comparing the performance of an egoistic autonomous mission vehicle with a human-driven mission vehicle in terms of crashes and failed merges shows the autonomous agent is generally more capable to find a way to merge into the platoon of humans and egoistic agents. However, it still fails in more than half of its merging attempts. Figure 5 verifies our hypothesis **H2** as we can observe that adding only a cooperation component to the agents, i.e., **AV+C** scenario, enables the mission vehicle to merge to the highway almost in all of its attempts. Adding the sympathy element in **AV+SC** slightly improves the safety as it incentivizes the agents to be aware of the humans that are not in the direct risk of collision with them. We consider cooperation as an enabler for sympathy and did not conduct any experiment with sympathetic-only setting as its results can be inferred from a comparison between (**SC**) and (**C**).

2) Tuning Altruism & Emerging Behaviors: To investigate hypothesis **H3**, we train a set of agents and vary their reward coefficients, i.e., λ_E, λ_C, λ_S, to adjust their level of sympathy and cooperation. Revisiting our driving scenario depicted in Figure 1, we particularly witness two critical emerging behaviors in agents. Strongly sympathetic agents

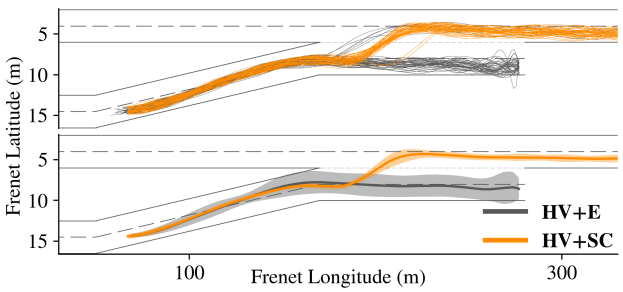

Fig. 6: A set of sample trajectories of the merging vehicle shows mostly successful merging attempts in **HV+SC**, compared to the failed attempts in **HV+E**.

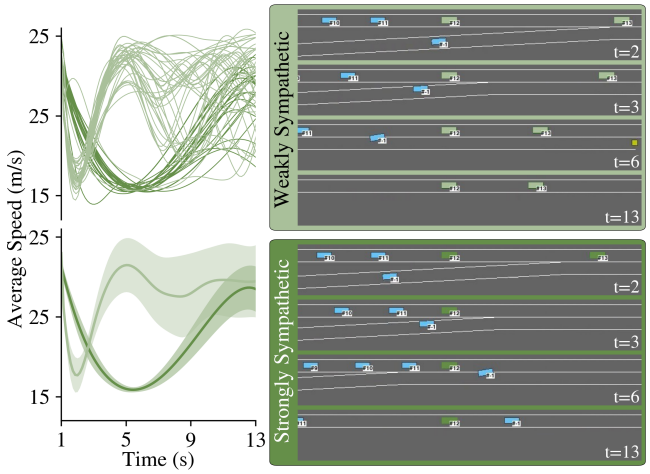

Fig. 7: Comparing weakly and strongly sympathetic autonomous agents: (*left*) Speed profiles of the "guide AV" (consider AV3 in Figure 1(b)) and (*right*) Sample snapshots.

that are trained with a high ratio of $\lambda_S/(\lambda_C+\lambda_E)$, naturally prioritize the benefit of humans over their own. Figure 7 shows a set of snapshots extracted from two scenarios with strongly sympathetic and weakly sympathetic agents. A strongly sympathetic agent (consider AV3 in Figure 1(b)) slows down and blocks the group of vehicles behind it to ensure that the mission vehicle gets a safe pathway to merge. On the other hand, the weakly sympathetic agent initially brakes to slow down the group of the vehicles behind it and then prioritizes its own benefit, speeds up, and passes the mission vehicle. Although both behaviors enable the mission vehicle to successfully merge, the speed profiles of the agent in Figure 7 depict how a strongly sympathetic agent compromises on its traveled distance (the area under the speed curve) to maximize the mission vehicle's safety. Motivated by this observation, we thoroughly studied the effect that tuning the reward coefficients in Equation (2) makes on the performance of SymCoDrive agents. As illustrated in Figure 8, we empirically observe that an optimal point between *caring about others* and *being selfish* exists that eventually benefits all the vehicles in the group.

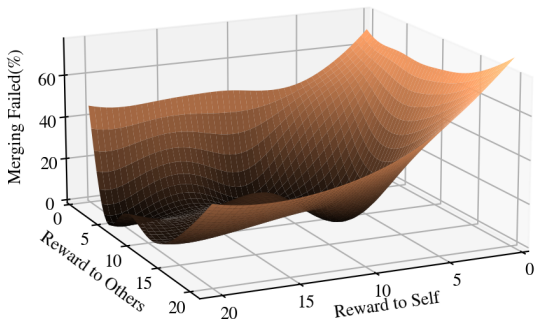

Fig. 8: Tuning SVO for autonomous agents reveal that an optimal point between *caring about others* and *being selfish* exists that eventually benefits all the vehicles in the group.

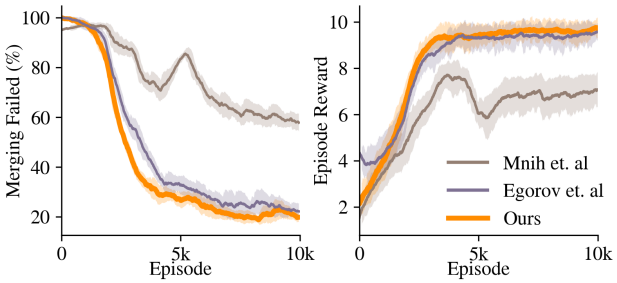

Fig. 9: Training performance of the three benchmark network architectures.

G. Deep Networks and Generalization

We trained the network architectures introduced in Section IV and examined their ability in generalizing to test episodes with 4x wider range of initialization randomness, figure 9 shows the training performance of the networks. When tested in episodes with the same range of initialization randomness as training, all networks showed acceptable performance. However, their performance quickly depreciated when the range of randomness was increased and agents faced episodes different than what they had seen during the training, as noted in Table II. While the other networks (particularly MLP) over-fitted on the training episodes, our Conv3D architecture significantly outperformed them in the more diverse test scenarios. We conclude that using VelocityMaps and our Conv3D architecture, agents learn to handle more complex unseen driving scenarios. Table I lists the hyper-parameters we have used to train our Conv3D architecture.

The Occupancy Grid state-space representation, defined in Equation (6) showed an inferior performance in all neural network architectures compared to the VelocityMap representation in our particular driving problem. We speculate that this is due to the fact that the Occupancy Grid representation does not benefit from the road layout and visual cues embedded in the VelocityMap state representation. All of our experiments discussed earlier are performed with VelocityMap representation, unless stated otherwise. After tuning the VelocityMaps, we concluded that integrating a hard ego-attention map in the state representation did not make a significant enhancement and decided to drop this channel, reducing the number of channels to 4. Instead, we aligned the center of VelocityMaps with regards to the ego such that 30% of the observation frame reflects the range behind the ego and the rest shows the range in front. We

TABLE I: List of hyper-parameters of our Conv3D Q-Network

Hyper-param	Value	Hyper-param	Value
Training iterations	720,000	Initial exploration	1.0
Batch size	32	Final exploration	0.1
Replay buffer size	10,000	ϵ decay	Linear
Learning rate	0.0005	Optimizer	ADAM
Target network update	200	Discount factor γ	0.95

TABLE II: Performance comparison of related architectures. Our Conv3D architecture outperformed the others as the level of randomness increases and agents face episodes different than what they had seen during the training.

	Low Randomness			Medium Randomness			High Randomness		
Models	C (%)	MF (%)	DT (m)	C (%)	MF (%)	DT (m)	C (%)	MF (%)	DT (m)
MLP	6.2	**0**	288	65.2	65.2	304	78.9	31.4	212
Mnih et. al. [**?**]	9.6	7.2	**350**	41.2	41.2	240	12.9	10.8	344
Egorov et. al. [**?**]	19.7	9.0	312	7.3	1.7	366	18.9	8.4	313
Conv3D (Ours)	**3.3**	0.2	334	**2.4**	**0.4**	**373**	**4.8**	**1.0**	**351**

C: *Crashed*, MF: *Merging Failed*, DT: *Distance Travelled*

noticed that this parameter plays an important role in training convergence and the resulted behaviors as it enables the agent to see the mission vehicle and other vehicles before they get to its close proximity.

VI. Concluding Remarks

Summary. We tackle the problem of autonomous driving in mixed-autonomy environments where autonomous vehicles interact with vehicles driven by humans. We incorporate a cooperative sympathetic reward structure into our MARL framework and train agents that cooperate with each other, sympathize with human-driven vehicles, and consequently demonstrate superior performance in competitive driving scenarios, such as highway merging, compared to egoistically trained agents.

Limitations and Future Work. Our current reward structure includes a hand-crafted marker that depends on the driving scenario, e.g., merging or exiting a highway. Given diverse driving episodes, this marker can also be learned from interaction data, cutting the need for a mission-specific reward term. We believe the merging scenario is representative of many common interaction scenarios we observe including other behaviors that require the two agents regulating their speeds and coordinating with each other such as exiting a highway. We have only experimented with training and testing agents in the same scenario and have not cross-validated them across different scenarios. We hope to extend this work to other scenarios in the future. We believe, given a large enough training data, an agent is expected to learn the same altruistic behavior in general driving scenarios. As a minor limitation, we have assumed perfect V2V communication which is not too unrealistic as the current V2V solutions easily handle hundreds of vehicles with almost no packet drop.