

# Evaluation of MPC-based Imitation Learning for Human-like Autonomous Driving

Flavia Sofia Acerbo<sup>\*,\*\*,\*</sup> Jan Swevers<sup>\*,\*\*\*</sup>  
Tinne Tuytelaars<sup>\*\*\*</sup> Tong Duy Son<sup>\*</sup>

<sup>\*</sup> *Siemens Digital Industries Software, Leuven, Belgium*  
(e-mail: [flavia.acerbo@siemens.com](mailto:flavia.acerbo@siemens.com)).

<sup>\*\*</sup> *KU Leuven, Mechanical Engineering Department, Leuven, Belgium.*

<sup>\*\*\*</sup> *KU Leuven, Electrical Engineering Department, Leuven, Belgium.*

<sup>\*\*\*\*</sup> *DMMS-M core-lab Flanders Make, Leuven, Belgium.*

**Abstract:** This work evaluates and analyzes the combination of imitation learning (IL) and differentiable model predictive control (MPC) for the application of human-like autonomous driving. We combine MPC with a hierarchical learning-based policy, and measure its performance in open-loop and closed-loop with metrics related to safety, comfort and similarity to human driving characteristics. We also demonstrate the value of augmenting open-loop behavioral cloning with closed-loop training for a more robust learning, approximating the policy gradient through time with the state space model used by the MPC. We perform experimental evaluations on a lane keeping control system, learned from demonstrations collected on a fixed-base driving simulator, and show that our imitative policies approach the human driving style preferences.

**Keywords:** Autonomous Vehicles, Learning and adaptation in autonomous vehicles, Human and vehicle interaction.

## 1. INTRODUCTION

Human-like autonomous driving is a promising solution to increase user acceptance of autonomous vehicles (AVs). Current research is primarily focusing on data-driven approaches, such as imitation learning (IL), which can obtain policies mimicking good driving demonstrations. In this domain, various solutions from both academia and industry show great potential on learning policies end-to-end, i.e. from sensory data to control actions, as in Hu et al. (2022); Scheel et al. (2021). However, the intensive use of deep learning in these approaches limits the safety, comfort and stability properties of the closed-loop system. Moreover, these IL policies are commonly evaluated either on open-loop accuracy of the control actions or in closed-loop on the accomplishment of the main task, number of collisions or on similarity to human decision-making in specific situations, as in Bronstein et al. (2022). However, it is only marginally evaluated whether the learned behavior is human-like, in the sense of driving characteristics that distinguish a human driving style from a robotic one, and that can influence the passenger comfort. Recent works have also proposed a hierarchical framework for these learning-based control schemes, where they have been combined

with model-based components such as model predictive control (MPC) to incorporate previous domain knowledge, explicitly handle safety constraints or plan vehicle dynamics conformant trajectories (Hewing et al. (2020), Acerbo et al. (2020)).

In this work, we investigate the use of differentiable MPC in IL for human-like AVs through experimental evaluations with human demonstrations. We identify the best design choice for robust learning of a human-like IL policy in the mix of MPC control, hierarchical policies and closed-loop training, where imitation losses are computed over simulations of policy and dynamics. Our contribution is threefold:

- We assess the impact of a differentiable parametric MPC policy compared to a pure learning-based one when training in open-loop, showing less tendency to the covariate shift. This enables us to evaluate our policy ability to imitate the human, under comfort and human-like metrics that can influence the passenger perception.
- We propose a hierarchical framework that combines MPC with a learning-based policy to better imitate some dynamic human-like behaviors, difficult to be incorporated in the MPC policy alone.
- We highlight how the MPC design can influence the closed-loop performance of the hierarchical policy trained in open-loop, by worsening the causal confusion. Hence, we demonstrate how closed-loop training can drastically reduce this effect, approximating the

<sup>\*</sup> This work has been submitted to IFAC for possible publication. This project has received funding from the Flemish Agency for Innovation and Entrepreneurship (VLAIO) under the Baekeland Mandaat No. HBC.2020.2263 (MIMIC) and has been supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 953348 (ELO-X).

policy gradient over time with the state space model used by the MPC.

The paper is structured as follows: in Section 2, we provide an introduction on related works concerning IL for AVs and MPC with learning-based schemes. In Section 3, we present the generic formulation of the proposed IL approach combining MPC differentiability, hierarchical decomposition and closed-loop training. Finally, experimental results are provided in Section 4, applied to the design of a human-like lane keeping controller from human demonstrations.

## 2. BACKGROUND AND RELATED WORK

### 2.1 Imitation Learning for Autonomous Driving

Let us consider a fully observable Markov Decision Process characterized by a set of continuous states  $s \in \mathcal{S} \subset \mathbb{R}^{n_s}$ , which can be controlled with a set of continuous actions  $a \in \mathcal{A} \subset \mathbb{R}^{n_a}$  and having a state transition probability distribution  $s_{t+1} \sim P(\cdot|s_t, a_t)$ . An agent can control the system with a policy  $\pi(a_t|s_t)$ . We denote the *state occupancy measure*  $\rho_\pi^s$  induced by a policy  $\pi$  as the density of occurrence of state  $s$ , while following that policy, over an infinite time horizon, discounted by the factor  $\gamma$ :  $\rho_\pi^s = \sum_{t=0}^{\infty} \gamma^t P(s|\pi)$ . Given an expert policy  $\pi_E$  to be imitated, IL tries to find a policy  $\pi$  minimizing the occupancy measure distance formulated as:  $\min_{\pi} \mathbb{E}_{s \sim \rho_{\pi_E}^s} [\mathcal{L}(\pi_E, \pi)] = \mathcal{L}(\rho_{\pi_E}^s, \rho_{\pi}^s)$ .

In the context of AVs, behavioral cloning (BC) is the IL algorithm with the most promising history of success, e.g. Bansal et al. (2018). Standard BC provides a straightforward solution to the IL problem, by learning the policy in a supervised way through direct maximum likelihood, i.e. by minimizing the distance between the action distributions under the expert state occupancy measure as:  $\min_{\pi} \mathbb{E}_{s \sim \rho_{\pi_E}^s} [\mathcal{L}(\pi_E(s), \pi(s))]$ . BC is an *open-loop (OL) learning* algorithm, as it does not require further environment interactions. For this reason, BC suffers from the covariate shift problem, i.e. the policy  $\pi$  can become unpredictable in its own induced distribution of states, if far from the expert one. Among different solutions to mitigate this problem, it has been shown by Ghasemipour et al. (2020), that learning the marginal state distribution of the expert in addition to the action distribution, can heavily influence the results. Hence, there have been proposed algorithms training the policy in *closed-loop (CL)*, where imitation losses are computed over states obtained by simulations of the closed-loop system (e.g., Bronstein et al. (2022); Suo et al. (2021)). Nevertheless, CL learning is more expensive than OL, since it requires (1) to simulate the policy in the environment during the learning loop, (2) to backpropagate through time and (3) it needs solutions to estimate the policy gradient from the states.

### 2.2 Model Predictive Control and Reinforcement Learning

Model predictive control (MPC) is an optimal controller based on a receding horizon strategy, which can be written in its most general form as:

$$\begin{aligned} a = \arg \min_{x, u} \quad & \sum_{k=0}^{N-1} l_k(x_k, u_k, \theta) + l_N(x_N, \theta) = l(x, u, \theta) \\ \text{s. t.} \quad & x_0 = \hat{x} \\ & x_{k+1} = f(x_k, u_k, \theta) \\ & h(x_k, u_k, \theta) \leq 0 \\ & a = u_0. \end{aligned} \quad (1)$$

At each time step, the current state of the system is estimated as  $\hat{x}$ . Then, MPC minimizes a cost function  $l$  across a prediction horizon  $N$ , where the evolution of the system is predicted with the state space model  $f$  and where states and controls should not violate the constraints specified by  $h$ . MPC was recently proposed as a solution for safety and sample efficiency issues in IL and reinforcement learning (RL). Gros and Zanon (2020) showed that using a highly parametric MPC as a differentiable function approximator for RL results in safe and stable policies and allows MPC to optimize its closed-loop performance, even when its underlying model  $f$  is inaccurate and system identification is not performed. Amos et al. (2018) have benchmarked the use of differentiable MPC for classical IL/RL problems showing the advantages in terms of sample efficiency and learning flexibility with respect to generic IL and system identification.

## 3. MPC-BASED IMITATION LEARNING

This section presents the formulation and generic algorithmic framework of our method combining MPC and IL. Here, differentiable MPC is used as the final control layer of a hierarchical policy. This means that MPC parameters are given by preceding learning-based components and the overall hierarchical policy is trained with imitation learning. Moreover, we present the use of MPC for closed-loop state cloning. The overall methodology is graphically summarized in Fig. 1.

### 3.1 Problem Formulation

We consider IL as the minimization of the occupancy measure distance, according to a certain metric  $\mathcal{L}$ . The system under study is controlled by a human with an unknown stochastic policy  $\pi_H$  as  $a_t \sim \pi_H(\cdot|s_t)$ . Instead, the mimicking agent controls the system with a deterministic policy parametrized by  $\theta$  as  $a_t = \pi_\theta(s_t)$ . We now formulate this problem in the context of *inverse optimal control*, where the policy is represented by an MPC, optimizing its control action  $a$  according to the state  $s$  and the parameters  $\theta$  that are part of the objective function, model and/or constraints:

$$\begin{aligned} \min_{\theta} \quad & \mathbb{E}_{s \sim \rho_{\pi_H}^s} [\mathcal{L}(\pi_H(s), \pi_\theta(s))] = \mathcal{L}(\rho_{\pi_H}^s, \rho_{\pi_\theta}^s) \\ \text{with} \quad & a(s) = \pi_\theta(s) \in \arg \min_{x, u} l(x, u, \theta) \\ \text{s. t.} \quad & x_0 = s \\ & x_{k+1} = f(x_k, u_k, \theta) \\ & h(x_k, u_k, \theta) \leq 0 \\ & a = u_0. \end{aligned} \quad (2)$$

Equation (2) describes a *bilevel optimization problem*. The upper-level problem entails realizing that policy  $\pi$  imitates the human policy  $\pi_H$ , while the lower-level problem entails



of the L2 state error  $L(s_t, s_t^*) = \|s_t - s_t^*\|_2^2$  between the human and policy on sequences of induced states, when starting from the same initial state  $s_0^*$ . These sequences are obtained by rolling out the policy on the black-box simulated system  $F : s_{t+1} = F(s_t, a_t)$ . We propose to compute an approximation of the system gradients  $\nabla_{s_t, a_t} F$  through the MPC model  $f$  by mapping  $s, a \rightarrow x, u \Rightarrow x_{k+1} = f(x_k, u_k) \rightarrow s_{t+1} = f(s_t, a_t)$ . Under this assumption, we compute  $\nabla_{\theta} L(s_{t+1})$  with backpropagation through time (BPTT) as:

$$\begin{aligned} \nabla_{\theta} L|_{s=s_{t+1}} &= \nabla_s L|_{s=s_{t+1}} (\nabla_a F \nabla_{\theta} a|_{a=a_t} + \nabla_s F \nabla_{\theta} s|_{s=s_t}) \\ &\approx \nabla_s L|_{s=s_{t+1}} (\nabla_u f \nabla_{\theta} a|_{u=a_t} + \nabla_x f \nabla_{\theta} s|_{x=s_t}). \end{aligned} \quad (5)$$

With BPTT future state errors are accounted by the earlier actions taken by the policy. We define  $\mathcal{L}(\theta) = \sum_{t=0}^T L(s_t, s_t^*)$  the sum of the L2 pose errors along a trajectory where  $s_0 = s_0^*$ . The gradient of  $\mathcal{L}(\theta)$  is computed by applying Equation (5) recursively, starting from  $t = T$  down to  $t = 0$ .

#### 4. EXPERIMENTS

This section presents experimental results of the methodologies presented in Section 3 for the design of a lane keeping control system from human demonstrations. On this application, we test different policies and imitation algorithms combinations. The evaluated policies are the following: MPC, with different designs, multilayer perceptron (NN) and a hierarchical combination of the two. Imitation losses are computed with: • OL behavioral cloning, i.e. L2 error between the policy actions and the recorded human ones, • OL supervised learning, i.e. L2 error between the high-level NN output in the hierarchical policy and the recorded human pose at a future time instant, or • CL state cloning, as detailed in Section 3.4.

##### 4.1 Human Demonstrations

Human demonstrations of lane keeping control are collected on a fixed-base driving simulator, as shown in Figure 1. First, the track and the visual feedback presented to the driver are designed with Simcenter Prescan. The track is 1200m long and made of 7 curved roads with different lengths and curvatures. Second, the ego vehicle is modelled via Simcenter Amesim as a 15DOF high-fidelity model of a Ford Focus. During the simulation, the driver controls the ego vehicle via a steering wheel with force feedback, while the longitudinal speed is kept constant at 50km/h. In this setting, 10 laps are collected around the track. These laps are then divided into trajectories of length  $T$  seconds for learning, with an overlap of  $T/2$  between them. Then, the trajectories are split into batches for stochastic gradient descent. For OL learning,  $T = 2$  seconds, which results into 2257 samples, divided into batches of size 64, and where one batch is used for validation. For CL learning,  $T = 10$  seconds, for a total of 140 trajectories, divided into batches of size 10.

##### 4.2 Vehicle Modelling

The state space model of the system used by the MPC considers Frenet coordinates to describe the vehicle kinematics with respect to the road centerline. The coordinate

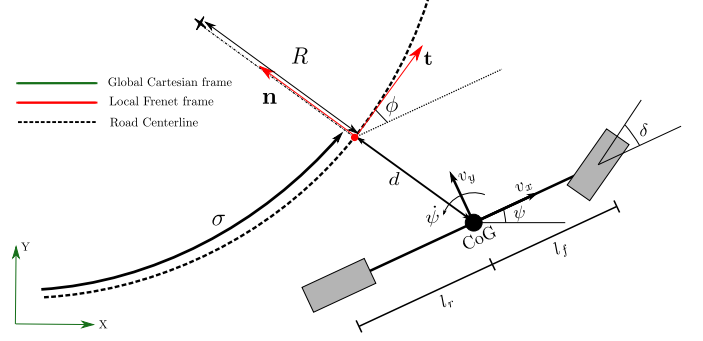


Fig. 2. Bicycle Model in Frenet coordinates.

system is defined with the following variables, shown in Figure 2: (1) the arc length  $\sigma$ , representing the travelled distance along the road, (2) the centerline deviation  $d$ , representing the lateral signed position on the road with respect to the centerline, according to the  $\mathbf{n} - \mathbf{t}$  frame (with positive  $\mathbf{t}$  according to vehicle direction of travel) and (3) the heading error  $\phi$ , representing the difference between the yaw angle of the vehicle and the heading of the road. The kinematics of the vehicle evolve according to the changing curvature of the road  $\kappa(\sigma)$ , that we assume as known. From Qian et al. (2016), the state equations for the kinematics are:

$$\begin{aligned} \dot{\sigma} &= \frac{v_x \cos \phi - v_y \sin \phi}{1 - \kappa(\sigma) d} \\ \dot{d} &= v_x \sin \phi + v_y \cos \phi \\ \dot{\phi} &= \dot{\psi} - \kappa(\sigma) \frac{v_x \cos \phi - v_y \sin \phi}{1 - \kappa(\sigma) d}. \end{aligned} \quad (6)$$

As to the dynamics, we employ a bicycle model, with vehicle sideslip angle  $\beta$  and yaw rate  $\dot{\psi}$  as states (see Milliken et al. (1995) for details). The steering angle at the wheel is denoted as  $\delta$ .

##### 4.3 MPC Formulation

Based on the model chosen in the previous section, we can define the states and controls of the MPC as:  $x = (\beta, \psi, \sigma, d, \phi)$ ,  $u = \delta$ . The MPC is formulated as in the following equations:

$$\min_{x, u} \sum_{k=0}^{N-1} l_k(x_k, u_k, \theta) \quad (7)$$

$$\text{s. t. } x_{k+1} = f(x_k, u_k) \quad (8)$$

$$-w/2 \leq d \leq w/2, \quad (9)$$

where: eq. (8) is the 4th order Runge-Kutta discretization (sample time  $dt = 0.1s$ ) of the vehicle model presented in Section 4.2; eq. (9) represents the safety constraints related to the lane boundaries, whose width is  $w$ ; eq. (7) is the objective function whose parameters are learned with IL. The other states and controls are also box-constrained according to physical limits. The control horizon is chosen such that the controller has a lookahead distance  $d_{la}$ , therefore  $N = \frac{d_{la}}{v_x dt}$ . The MPC is implemented using Rockit, by Gillis et al. (2020). and then translated into a PyTorch module.

#### 4.4 Metrics

We measure policy performance using selected metrics belonging to four different areas:

(1) **Imitation:**

- Open-loop (OL) error: mean L2 error, in rad or rad/s, between the policy control action  $\pi_\theta(s_t^*)$  and the human control action  $a_t^*$ , where  $s_t^*$  is taken from the validation batch of the human demonstrations.
- Closed-loop (CL) likelihood: the human lateral deviation collected among the 10 laps  $d^*$  is modeled as a Gaussian Process with independent variable  $\sigma$ , such that  $d^*(\sigma) \sim \mathcal{N}(\mu_\sigma, \text{Var}_\sigma)$ . The policy  $\pi_\theta$  is rolled out on the track starting from a neutral position on the centerline, i.e.  $s_0 = \mathbf{0}$  for  $t = 1, \dots, T$  timesteps. Then, the CL imitation performance is computed as the mean likelihood of each  $d_t$  point as:

$$\frac{t}{T} \sum_{t=1}^T \left[ \frac{1}{\sqrt{\text{Var}_{\sigma_t} 2\pi}} e^{-\frac{(d_t - \mu_{\sigma_t})^2}{2\text{Var}_{\sigma_t}}} \right].$$

(2) **Safety.** On the policy closed-loop simulation, the number of timesteps where there is a violation of the lane boundaries constraint.

(3) **Comfort.** On the policy closed-loop simulation, mean and standard deviation of the lateral jerk  $j_y$ , in  $\text{m/s}^3$ .

(4) **Human-like.** Driving characteristics computed on the closed-loop simulations, to be compared with the ones computed from the human demonstrations:

- Lateral deviation from the centerline ( $d$ ): mean and standard deviation, in meters.
- Steering Reversal Rate (SRR): the number, per minute, of steering wheel angle reversals larger than 5 degrees, as defined by Markkula and Engström (2006). The steering signal is filtered as 0.6Hz.

#### 4.5 Results

In this section we present results of the different policies and imitation algorithms combinations. We assess the impact of a differentiable MPC policy, evaluate advantages and potential hidden pitfalls of a hierarchical MPC policy and demonstrate the importance of closed-loop state cloning to limit covariate shift. All the evaluation metrics are summarized in Table 1.

**Effect of the Differentiable MPC Policy** We evaluate the effect of training a MPC policy in open-loop (OL), i.e. with behavioral cloning (BC). A feedforward neural network policy (NN) is considered as our baseline: it takes a feature vector  $\chi_t = (d_t, \phi_t, \kappa(\sigma_t), \kappa(\sigma_t + 5), \kappa(\sigma_t + 10), \dots, \kappa(\sigma_t + 30))$  and outputs the corresponding steering angle to be applied to the vehicle. Intuitively, it considers the current relative position on the road and the curvature of the road at the current point and ahead of 5, 10, ..., 30 meters. We concatenate these 9 features and connect them to the output via 4 dense layers (with hidden sizes 64, 32 and 16) and ReLU activation. The MPC policy controls the steering angle too, with a quadratic objective function as:  $\min_{x,u} \sum_{k=0}^{N-1} W_d(d_k - \bar{d})^2 + W_\phi \phi_k^2 + W_\delta \delta_k^2$ . Its learnable parameters are  $(W_d, W_\phi, W_\delta, \bar{d})$ . The weights

$(W_d, W_\phi, W_\delta)$  are enforced to be non-negative by mapping them in logarithmic space such that the parameter vector results in  $\theta = (\log(W_d), \log(W_\phi), \log(W_\delta), \bar{d})$ . Its initial value before BC is a zero-vector. For fair comparison, its prediction horizon equals the one of the NN, hence  $d_{la} = 30$  meters. For both policies we optimize the L2 error between their output and the recorded human steering angle, with the recorded human states as input i.e.,  $L = \|\pi_\theta(s_t) - a_t^*\|_2^2$ . We minimize this loss using Adam with a learning rate of  $10^{-4}$  for the NN parameters and  $10^{-2}$  for the MPC parameters.

The NN policy reaches the reported open-loop imitation loss after 25 epochs. When in closed-loop, it suffers from the covariate shift and drives off the lane after 10 seconds, without being able to come back to a stable state. With the same amount of data and training epochs as NN, the MPC parameters converge to  $(-0.28, 0.22, 0.28, -0.27)$ . The MPC policy shows a higher open-loop loss but also a much more stable closed-loop behavior, which makes it possible to evaluate its similarity to the human behavior. The most relevant human driving preference captured by the MPC parameters is the mean preferred  $d$ , which happens to be more on the right side of the road. This is expected as this objective is taken into account in the MPC structure by  $\bar{d}$ . However, by considering other metrics, such as the standard deviation of  $d$  and SRR, the MPC similarity to the human decreases.

**Effect of the Hierarchical Decomposition** We evaluate the effect of a hierarchical variant of the policy, modifying the MPC objective function to track changing setpoints in  $d$  and  $\phi$  provided by the NN. The NN takes the feature vector  $\chi_t$ , introduced in the previous paragraph, and outputs the desired setpoints for time instant  $t$ :  $(\bar{d}_t, \bar{\phi}_t)$ . Subsequently, the MPC objective function is changed to track these setpoints while minimizing the control action as:  $\min_{x,u} \sum_{k=0}^{N-1} (d_k - \bar{d}_t)^2 + (\phi_k - \bar{\phi}_t)^2 + W_\delta \delta_k^2$ . By keeping  $W_\delta = 1$  fixed, the decomposition allows us to directly learn a mapping that imitates the observed human goals. This is done through supervised learning (SL) on the output setpoints of the NN:  $L = \|g(\chi_t^*) - (d_{t+Ndt}^*, \phi_{t+Ndt}^*)\|_2^2$ , where  $g(\chi_t^*)$  is the output of NN given the human recorded state features  $\chi_t^*(s_t^*)$  and  $Ndt$  is the prediction horizon length of the MPC in seconds. This yields a more human-like  $d$  in its dynamic content, i.e. in the standard deviation. Since the policy still shows high SRR and  $j_y$ , we propose to augment the state as  $x = (\beta, \dot{\psi}, \sigma, d, \phi, \delta)$  and define the steering angle rate as our control action  $u = \dot{\delta}$ . This allows us to penalize high frequency steering rate changes, and so indirectly optimize for comfort. Hence, the objective function becomes:  $\min_{x,u} \sum_{k=0}^{N-1} (d_k - \bar{d}_t)^2 + (\phi_k - \bar{\phi}_t)^2 + W_\delta \dot{\delta}_k^2$ . This combination yields the best overall performance: the CL likelihood of the  $d$  trajectory increases, together with less lateral jerk and more human-like SRR.

However, the hierarchical decomposition poses more risk than the MPC policy alone as the NN can still suffer from covariate shift. This becomes evident when we reduce the prediction horizon of the MPC by decreasing  $d_{la}$ . By doing so,  $t + Ndt$  gets closer to  $t$  and hence the optimal setpoint  $(d_{t+Ndt}^*, \phi_{t+Ndt}^*)$  gets closer to the NN input  $(d_t, \phi_t)$  itself, generating spurious correlations between the

Table 1. Evaluation metrics for different policy and algorithm configurations. The tested policies differ in: type, i.e. MPC, neural network (NN) or a hierarchical combination of the two (NN-MPC); control action, i.e. steering angle  $\delta$  or steering rate  $\dot{\delta}$ ; length of the MPC prediction horizon, in meters ( $d_{la}$ ); imitation algorithm, either in open-loop behavioral cloning or supervised learning (BC, SL) or augmented with closed-loop state cloning (SC).

| Configuration |           |                |          | Imitation                         |             | Safety   | Comfort                           | Human-like                         |              |
|---------------|-----------|----------------|----------|-----------------------------------|-------------|----------|-----------------------------------|------------------------------------|--------------|
| Policy        | Algorithm | $a$            | $d_{la}$ | OL                                | CL          | Off-Road | $j_y$                             | $d$                                | SRR          |
| NN            | BC        | $\delta$       | /        | <b>0.01 <math>\pm</math> 0.00</b> | 0.02        | 622      | <b>0.14 <math>\pm</math> 0.14</b> | 29.32 $\pm$ 12.43                  | <b>28.33</b> |
| MPC           | BC        | $\delta$       | 30       | 0.66 $\pm$ 0.20                   | <b>0.93</b> | <b>0</b> | <b>0.13 <math>\pm</math> 0.19</b> | -0.58 $\pm$ 0.04                   | 52.90        |
| NN-MPC        | SL        | $\delta$       | 30       | 0.39 $\pm$ 0.11                   | <b>0.93</b> | <b>0</b> | 0.24 $\pm$ 0.42                   | <b>-0.57 <math>\pm</math> 0.30</b> | 62.5         |
| NN-MPC        | SL        | $\dot{\delta}$ | 30       | 0.62 $\pm$ 0.18                   | <b>0.98</b> | <b>0</b> | <b>0.06 <math>\pm</math> 0.09</b> | <b>-0.53 <math>\pm</math> 0.29</b> | <b>25.83</b> |
| NN-MPC        | SL        | $\dot{\delta}$ | 9.72     | <b>0.11 <math>\pm</math> 0.02</b> | 0.48        | 2        | <b>0.09 <math>\pm</math> 0.13</b> | -0.50 $\pm$ 0.76                   | <b>17.50</b> |
| NN-MPC        | SL + BC   | $\dot{\delta}$ | 9.72     | <b>0.09 <math>\pm</math> 0.01</b> | 0.02        | 678      | 0.08 $\pm$ 0.30                   | -153.85 $\pm$ 111.34               | 5.83         |
| NN            | BC + SC   | $\delta$       | /        | <b>0.05 <math>\pm</math> 0.01</b> | 0.80        | <b>0</b> | 0.11 $\pm$ 0.24                   | -0.71 $\pm$ 0.26                   | 67.50        |
| NN-MPC        | SL + SC   | $\dot{\delta}$ | 9.72     | 0.30 $\pm$ 0.07                   | <b>0.93</b> | <b>0</b> | <b>0.07 <math>\pm</math> 0.09</b> | <b>-0.46 <math>\pm</math> 0.17</b> | <b>23.33</b> |
| Human         |           |                |          | /                                 | 1.22        | 0        | 0.09 $\pm$ 0.13                   | -0.54 $\pm$ 0.39                   | 19.5         |

input features and overall worsening the CL performance. As in De Haan et al. (2019), this phenomenon is called *causal confusion*: adding more information (e.g. history) as input to a policy trained with BC yields to a worse covariate shift. We test this phenomenon by decreasing  $d_{la}$  to 9.72 meters (0.7 seconds). This reduction causes a performance drop of 51% in the CL likelihood (from 0.98 to 0.48), which can also be seen qualitatively in Figure 3. In addition, the policy even slightly violates the lane boundaries constraint. We also test the effect of BC training on this policy. We differentiate through the MPC to learn the mapping to directly imitate  $\dot{\delta}^*$ , this time learning  $W_{\dot{\delta}}$  too. With BC the CL performance of the policy substantially drops, causing an unstable behavior, as in the case of NN alone. This suggests that learning on the human distribution of states may not be robust enough when policies with NN are involved, even with hierarchical decomposition.

**Effect of the Closed-Loop State Cloning** To reduce the effect of causal confusion, we augment SL and BC with state cloning (SC) that, as detailed in Section 3.4, employs the MPC model to do closed-loop training on the states generated by the policy. In this case, the parameter  $W_{\dot{\delta}}$  is also learned by the SC algorithm. We set the imitation loss on the lateral deviation  $d$ , such that  $L(s_t, s_t^*) = \|d_t - d_t^*\|_2^2$ . SC significantly improves the performance for the NN policy alone, which shows a more stable behavior and can imitate some human attributes, such as the standard deviation of  $d$ . However, compared to the NN-MPC policy variants shown in the previous paragraph, we notice that the NN policy has worse comfort performance, e.g. on the lateral jerk, also due to the high SRR. This is unsurprising as we do not explicitly optimize for these attributes in the imitation loss or in the policy itself. On the other hand, SC on the NN-MPC policy with  $d = 9.72$  yields a 94% improvement (from 0.48 to 0.93) after 100 epochs on the CL imitation metric compared to SL, and no violations of the lane boundaries. Moreover, compared to the NN policy trained with SC, it performs better in terms of comfort and SRR, since they are indirectly optimized for in the MPC objective function. The improvement brought by SC can be seen qualitatively on the lateral deviation  $d$  in Figure 3. It is worth reporting that employing SC directly without warm-starting with OL learning (BC/SL) leads to local minima that don’t yield the same level of

performance, e.g. the NN-MPC policy with  $d_{la} = 9.72$  initialized with random NN parameters starts from an unstable closed-loop behavior that SC cannot improve. This suggests that a combination of both OL and CL losses should be considered.

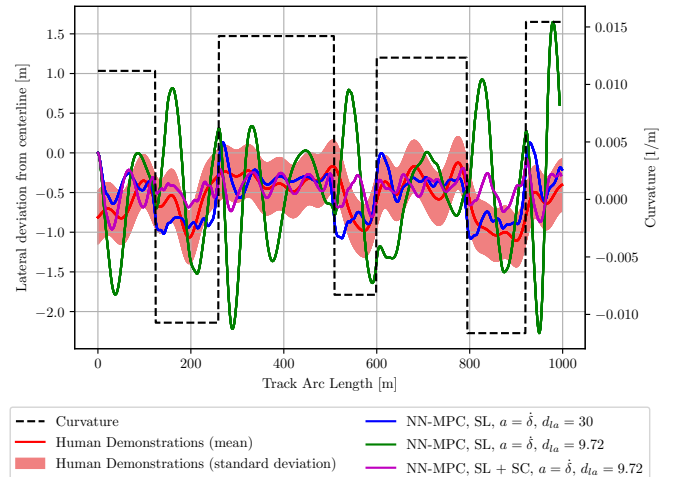


Fig. 3. Lateral centerline deviation for different policy and algorithm configurations.

## 5. CONCLUSIONS

We evaluated a hierarchical MPC-based imitation learning method in its ability to imitate human driving behavior during a lane keeping task. We have evaluated different possible configurations of policy types and training algorithms with metrics that go beyond the accomplishment of the given task and that can correlate to passenger perception, such as the standard deviation of the lateral centerline deviations or the steering reversal rate. We have demonstrated that our method can approach the human preferences under these metrics, even without directly optimizing for them, by careful MPC design and its combination with more flexible learning-based components. We also highlighted the importance of closed-loop training to reduce the effect of causal confusion, which can show up when designing MPC with shorter prediction horizons. Further works will include human in the loop subjective testing of the demonstrated policies, to investigate passenger perception of the objective metrics, and the use



of camera images as inputs to the policy, to compare our method to end-to-end approaches.

## REFERENCES

- Acerbo, F.S., der Auweraer, H.V., and Son, T.D. (2020). Safe and computational efficient imitation learning for autonomous vehicle driving. In *2020 American Control Conference (ACC)*. IEEE.
- Amos, B., Jimenez, I., Sacks, J., Boots, B., and Kolter, J.Z. (2018). Differentiable mpc for end-to-end planning and control. *Advances in neural information processing systems*, 31.
- Andersson, J.A.E., Gillis, J., Horn, G., Rawlings, J.B., and Diehl, M. (2018). CasADi: a software framework for nonlinear optimization and optimal control. *Mathematical Programming Computation*, 11(1), 1–36.
- Andersson, J.A. and Rawlings, J.B. (2018). Sensitivity analysis for nonlinear programming in CasADi. *IFAC-PapersOnLine*, 51(20), 331–336.
- Bansal, M., Krizhevsky, A., and Ogale, A. (2018). Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. *arXiv preprint arXiv:1812.03079*.
- Bronstein, E., Palatucci, M., Notz, D., White, B., Kuefler, A., Lu, Y., Paul, S., Nikdel, P., Mougin, P., Chen, H., et al. (2022). Hierarchical model-based imitation learning for planning in autonomous driving. *arXiv preprint arXiv:2210.09539*.
- De Haan, P., Jayaraman, D., and Levine, S. (2019). Causal confusion in imitation learning. *Advances in Neural Information Processing Systems*, 32.
- Ghasemipour, S.K.S., Zemel, R., and Gu, S. (2020). A divergence minimization perspective on imitation learning methods. In L.P. Kaelbling, D. Kragic, and K. Sugiura (eds.), *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, 1259–1277. PMLR.
- Gillis, J., Vandewal, B., Pipeleers, G., and Swevers, J. (2020). Effortless modeling of optimal control problems with rokit. In *39th Benelux Meeting on Systems and Control*.
- Gros, S. and Zanon, M. (2020). Data-driven economic NMPC using reinforcement learning. *IEEE Transactions on Automatic Control*, 65(2), 636–648.
- Hewing, L., Wabersich, K.P., Menner, M., and Zeilinger, M.N. (2020). Learning-based model predictive control: Toward safe learning in control. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(1), 269–296.
- Hu, A., Corrado, G., Griffiths, N., Murez, Z., Gurau, C., Yeo, H., Kendall, A., Cipolla, R., and Shotton, J. (2022). Model-based imitation learning for urban driving. *arXiv preprint arXiv:2210.07729*.
- Markkula, G. and Engström, J. (2006). A steering wheel reversal rate metric for assessing effects of visual and cognitive secondary task load. In *Proceedings of the 13th ITS World Congress*. Leeds.
- Milliken, W.F., Milliken, D.L., et al. (1995). *Race car vehicle dynamics*, volume 400. Society of Automotive Engineers Warrendale, PA.
- Qian, X., de La Fortelle, A., and Moutarde, F. (2016). A hierarchical model predictive control framework for on-road formation control of autonomous vehicles. In *2016 IEEE Intelligent Vehicles Symposium (IV)*. IEEE.
- Scheel, O., Bergamini, L., Wolczyk, M., Osinski, B., and Ondruska, P. (2021). Urban driver: Learning to drive from real-world demonstrations using policy gradients. In *Conference on Robot Learning (CoRL)*.
- Suo, S., Regalado, S., Casas, S., and Urtasun, R. (2021). Trafficsim: Learning to simulate realistic multi-agent behaviors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10400–10409.