Dreaming to Assist: Learning to Align with Human Objectives for Shared Control in High-Speed Racing

Anonymous Author(s)
Affiliation
Address
email

Abstract:
Tight coordination is required for effective human-robot teams in domains involving fast dynamics and tactical decisions, such as multi-car racing. In such settings, robot teammates must react to cues of a human teammate's tactical objective to assist in a way that is consistent with the objective (e.g., navigating left or right around an obstacle). To address this challenge, we present DREAM2ASSIST, a framework that combines a rich world model able to infer human objectives and value functions, and an assistive agent that provides appropriate expert assistance to a given human teammate. Our approach builds on a recurrent state space model to explicitly infer human intents, enabling the assistive agent to select actions that align with the human and enabling a fluid teaming interaction. We demonstrate our approach in a high-speed racing domain with a population of synthetic human drivers pursuing mutually exclusive objectives, such as "stay-behind" and "overtake". We show that the combined human-robot team, when blending its actions with those of the human, outperforms the synthetic humans alone as well as several baseline assistance strategies, and that intent-conditioning enables adherence to human preferences during task execution, leading to improved performance while satisfying the human's objective.

Keywords: Recurrent State-Space Models, Human-Robot Interactions, Shared-Control

1 Introduction

In high-stakes situations where members of a team must coordinate their physical actions in the world for the team to succeed, early coordination on tactical objectives is crucial. In a rapidly-evolving task, such as in high-speed competitive sports, agents must find a way to attain such coordination without explicit communication. A robotic assistive agent equipped with an ability to reason using theory of mind [1] has been shown to be critical to successful collaboration without the burden and latency of explicit communication. In such settings, agents must maintain a rich-enough model of the world to cover a common set of concepts that each agent needs to plan. This includes dynamics of the physical world, the objectives of the team, and the current intent of the other members of the team. Such considerations are prevalent in sports [2, 3], manufacturing [4], and healthcare settings [5], among others.

High-speed performance driving presents a domain where accurate and expressive models are required to have effective human-robot teams. The dynamics in racing evolve quickly, preventing team members from communicating their goals or objectives before taking an action [6]. Because of this constraint, existing approaches in shared control or advanced driver assistance often split authority on predefined boundaries (e.g. steering vs. throttle and brake) [7, 8]. The driving domain also requires us to tackle the multimodal nature of the problem (e.g. rules or maneuvers), as shown by prior work [9, 10, 11, 12]. We hypothesize that a theory-of-mind-inspired model, coupled with

Submitted to the 8th Conference on Robot Learning (CoRL 2024). Do not distribute.

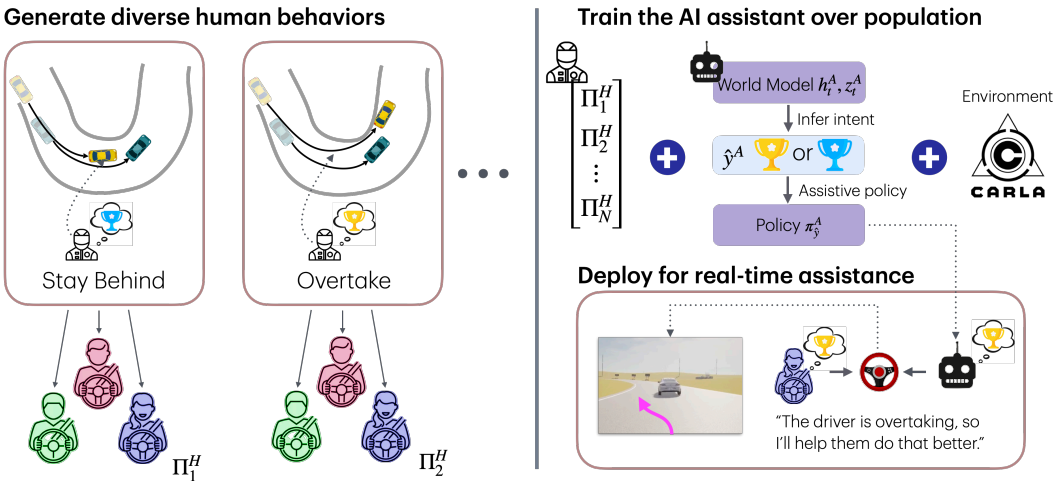

Figure 1: We mimic human preferences on discrete decisions via population clusters during world model formation. Our assistive models then learn to decisively help on the overall discrete-continuous control task while taking into account multiple possible human preferences.

recent advances in flexible planning and reinforcement learning, can enable new ways for machines to help humans drive more efficiently and effectively.

In this paper, we present a modeling paradigm to explicitly infer and support a human teammate's intent, enabling our agent to learn how to interact with diverse human strategies in a high-speed continuous control task. We consider the problem of a robotic assistive agent in a race car that helps a human driver perform tactical overtaking maneuvers more safely and more optimally in a high-speed competitive domain where explicit communication is difficult. While the driver remains in control of the vehicle at all times, the assistive agent can augment the driver's lateral (steering) and longitudinal (throttle and brake) actions. The assistant must infer the driver's intention (e.g., whether or not the driver is going to attempt an overtake) in order to provide optimal control augmentations and modifications that help the driver execute the inferred intention.

In this work, we demonstrate a novel model-based reinforcement learning (MBRL) paradigm that jointly learns the dynamics of the physical world and the human's intentions, enabling fluid shared control with continuous assistance during a race. We infuse a world model with a richer world-representation by using fictitious humans with multiple, mutually exclusive, objectives with the aim of encoding diverse human intents into the model's world representation.

Specifically, we contribute:

1. A novel approach to MBRL for shared control in a tactical setting, leveraging driver intent modeling using fictitious co-play, conditioned on a fixed set of human objectives.
2. A means for expert human value alignment, which conditions an assistive agent's rewards based on inferred intent to reason jointly over the physical world and the human's behavior.
3. An evaluation of our approach on a shared control racing domain, demonstrating our model's utility over a diverse set of imperfect fictitious humans with different personal objectives.

1.1 Related Works

Significant work has taken place in sharing control [7] and human-robot interaction [13, 14], exploring topics from ergonomics and physiology to language-based strategic teaming. Prior research has presented learning-based approaches to human-robot interactions that target joint-representation learning [15, 16, 17, 18, 19, 20, 21, 22]. These works focus on a range of topics, including data-efficient representation learning [23], intent or plan modeling [24, 25, 26, 27], or other cognitive and social motivations [28, 29, 30]. We go beyond prior work by inferring the human's discrete intent and modifying our robotic assistant's objective to encourage support of the inferred intent.

In the context of driving, shared control has been explored for planning approaches [31, 32, 33, 34, 35, 36] and learning-based approaches [37, 38, 39, 40]. Prior work has also considered incorporating game-theory into shared control [40, 41], as well as explicit human-centric design considerations [42, 43]. We refer readers to [44, 45] for comprehensive reviews. Beyond shared control, significant literature explored intent prediction in driving, see [46, 47], and references therein.

Recent work has proposed a shared-control model using model-predictive control that considers predicted trajectory information [33], therefore implicitly capturing driver intent. Additional recent work has augmented a model-predictive controller with the ability to explicitly infer driver intent [8], enabling the controller to share steering and control actions with a human. In our approach, we capture discrete driver intent in a way that is conducive to semantically-meaningful, multi-modal continuous behaviors (e.g., going left *or* right around an obstacle). Further, by framing our task as a multi-agent reinforcement-learning problem, our proposed solution extends to multiple agents much more naturally than prior work. Finally, we frame our decision making approach within a recurrent state-space model [48] which is extended to infer the objective of the human, building on recent work that has explored hierarchical or hybrid state abstractions [49, 50].

2 Background and Problem Statement

We target the problem of shared control in the highly-dynamic setting of high-performance racing against other racing opponents. We aim to build an assistive agent that is capable of reasoning over well-defined *task objectives*, as well as more general, harder-to-define, and harder-to-observe *human objectives*. The assistive agent is given a map of the track, states of the ego vehicle and opponent vehicle, the ego driver's steering and throttle controls. The agent's task is to assist the ego driver through modifications to the steering and throttle of the ego car, as depicted in Fig. 1 (lower right).

To achieve optimal performance, the agent must provide continuous control adjustments as the driver progresses along the track, helping the driver to stay on course, maintain proper speeds, and avoid collisions. Further, the agent must accurately infer the intentions of the driver (such as "stay to the opponent's left") and provide control augmentations that help the driver to accomplish their immediate task objective more optimally. Note that this problem is different from a conventional autonomous driving problem, as the agent's actions are conditioned explicitly on the human driver's control input, and the control signal that actuates the car is a linear blend of the agent and the driver's control signals. This problem also differs from conventional human-robot teaming, in that the agent must infer a human's intent (i.e., the human objective) early and as consistently as possible, as the efficacy on its assistance (in terms of safety and performance) depends strongly to how early, accurately, and reliably the human objective is captured.

3 Approach

Our approach learns a common latent representation in the structure of a recurrent state space model (RSSM) [51, 52] to build a hybrid discrete / continuous state of the world, partially encompassing the behaviors and values of the human partner. RSSMs have been shown to be suitable for many domains (e.g. locomotion, Atari, Minecraft). To accommodate both a *human* and *assistive* agent operating in a collaborative setting, we train the assistive agent alongside human agents, with both agents sharing actions taken in an environment. Each agent receives a reward for each action taken, and each are trained to (1) build an accurate RSSM and (2) learn how to act in order to maximize its own expected returns [53]. One challenge in collaboration is that part of the assistive agent's world model includes the preferences, desires, and goals of the human (i.e., the human objective), which are often unknown and only weakly observed by their behavior. We extend the RSSM formulation to allow it to be supervised on a diverse set of humans, whose objectives (intents) are known and labeled, to force the representation to be jointly aware of the behavior and intent across a variety of human types and physical environments. We then devise a scheme that feeds in inference of the human's objective into the assistant's reward function.

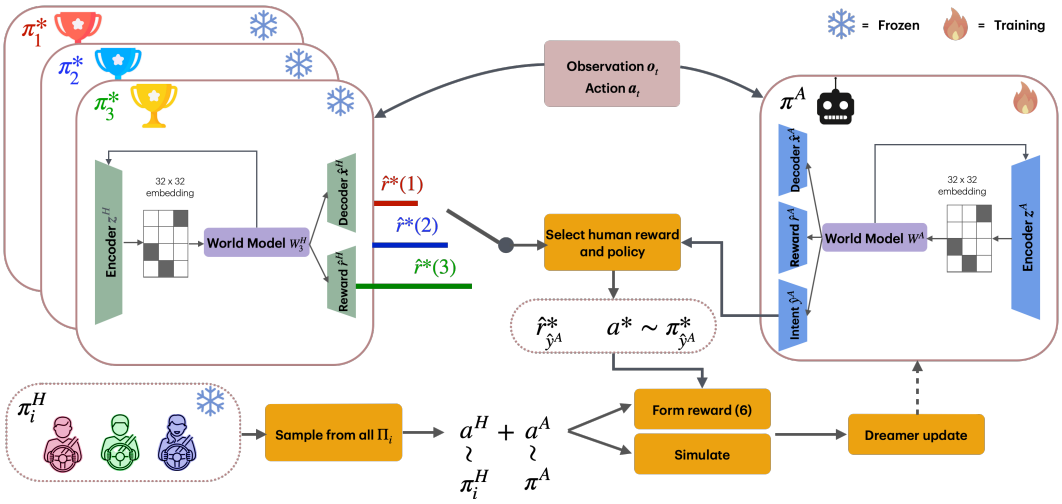

Figure 2: Overview of value alignment for the assistive agent. We start with a set of frozen human policies whose values are annotated with predetermined outcomes, then blend these human actions linearly with the actions from an unfrozen assistive agent. For both human and assistive agents, a multi-head RSSM architecture is then used to predict the observations, reward, and human intent (assistive only), trained to maximize their log-likelihoods against samples taken from the environment. The intent head is trained to match frozen human intents, which are fixed a priori. The assistive agent's rewards are shaped based on the *optimal policy* $\pi^*_{\hat{y}^A}$ and optimal predicted reward $r^*_{\hat{y}^A}$ for inferred intent $\hat{y}^A$.

3.1 Building a World Model over the Human and Physical Environment

In order to allow the system to maintain an ability to reason over the preferences of the human independent of that of the joint human-robot system, our approach considers both the human driver and the assistive agent planner as separate models. The training process is outlined in Fig. 2.

In the cooperative setting, the model of the environment follows a structure in which there are certain task-specific rewards which are available to both human and assistive agents, with the key distinctions being human-objective (specific to the human), and intervention penalties (specific to the assistant).

Task-Specific Rewards We assume standard task-specific rewards for high-performance driving derived from prior work [54, 55], including an out-of-bounds penalty, passing reward, and collision penalty.

$$r_t^{task} = r_t^{collision} + r_t^{bounds} + r_t^{finish} \tag{1}$$

where $r_t^{collision}$ is a negative reward for collisions, r_t^{bounds} is a negative reward for driving too far off the track, and r_t^{finish} is a positive reward for reaching the finish line.

Intent-Aware World Model The objective of the world model is to provide a representation that the agent can use to interact with the driver and the world, and we posit that this representation can support the agent to reason jointly about both the task and human objectives. We build off of the recurrent state-space model (RSSM) of DreamerV2/V3 [56, 51], according to the architecture in Figure 2. For agent $\kappa \in \{H, A\}$ (respectively, the **H**uman and **A**ssistive agent), the RSSM model, parameterized by ϕ and denoted W_ϕ^κ, includes:

$$\begin{aligned}
&\text{Encoder for discrete representation } \boldsymbol{z}_t^\kappa: && \boldsymbol{z}_t^\kappa \sim q_\phi^\kappa(\boldsymbol{z}_t^\kappa | \boldsymbol{h}_t^\kappa, \boldsymbol{x}_t^\kappa) \\
&\text{Sequence model for recurrent state } \boldsymbol{h}_t^\kappa: && \boldsymbol{h}_t^\kappa = f_\phi^\kappa(\boldsymbol{h}_{t-1}^\kappa, \boldsymbol{z}_{t-1}^\kappa, \boldsymbol{a}_{t-1}^\kappa) \\
&\text{Dynamics predictor:} && \hat{\boldsymbol{z}}_t^\kappa \sim p_\phi^\kappa(\hat{\boldsymbol{z}}_t^\kappa | \boldsymbol{h}_t^\kappa)
\end{aligned} \tag{2}$$

The output heads are similar to the DreamerV2/V3 architecture, with the addition of an intent predictor for the assistant agent (in red), and are all bottlenecked on the hidden states $\boldsymbol{s}_t^\kappa = \{\boldsymbol{h}_t^\kappa, \boldsymbol{z}_t^\kappa\}$,

$$\begin{aligned}
&\text{Decoder:} && \hat{\boldsymbol{x}}_t^\kappa \sim p_\phi^\kappa(\hat{\boldsymbol{x}}_t^\kappa | \boldsymbol{s}_t^\kappa) && \text{Continue predictor:} && \hat{c}_t^\kappa \sim p_\phi^\kappa(\hat{c}_t^\kappa | \boldsymbol{s}_t^\kappa) \\
&\text{Reward:} && \hat{r}_t^\kappa \sim p_\phi^\kappa(\hat{r}_t^\kappa | \boldsymbol{s}_t^\kappa) && \text{Intent predictor:} && \hat{y}_t^A \sim p_\phi^A(\hat{y}_t^A | \boldsymbol{s}_t^A)
\end{aligned} \tag{3}$$

To train the intent predictor, we assume at least partial access to the intent labels for a given episode, which takes the form of an integer-valued target function. We normalize the output by predicting the symlog ($\text{sign}(y)\ln(|y|+1)$) of the output, and use a discrete distribution to predict y using the two-hot encoding of [51]. The parameters ϕ of the world model W_ϕ^κ are trained to minimize the loss

$$\mathcal{L}^\kappa(\phi) = \mathbb{E}_{q_\phi^\kappa}\left[\beta_{pred}\mathcal{L}_{pred}^\kappa + \beta_{KL}\mathcal{L}_{KL}^\kappa\right]$$

The prediction loss $\mathcal{L}_{pred}^\kappa$ minimizes the likelihood under the predictor distributions in (3), $\mathcal{L}_{KL}^\kappa$ minimizes the KL divergence between the prior p_ϕ^κ and the approximate posterior q_ϕ^κ, and β_{pred} and β_{KL} are scalar weighting values. Given an intent label y_t, the portion of loss $\mathcal{L}_{pred}^A$ for the intent predictor is taken as the negative log-likelihood of the label under p_ϕ^A.

The world model training is alternated with training for the behavior model governing actions $\boldsymbol{a}_t^\kappa$, with the behavior model learned via an actor-critic policy training over the estimates $\boldsymbol{s}_t$, $\hat{r}_t$,

$$\text{Actor:} \qquad \boldsymbol{a}_t^\kappa \sim \pi_\theta^\kappa(\boldsymbol{a}_t^\kappa \mid \boldsymbol{s}_t^\kappa) \qquad \text{Critic:} \qquad v_\psi^\kappa(\boldsymbol{s}_t) \approx \mathbb{E}_{p_\phi^\kappa,\pi_\theta^\kappa}[\hat{r}_t^\kappa] \tag{4}$$

The critic is trained to minimize the temporal-difference loss on the value function v_ψ^κ. The actor attempts to maximize the critic-predicted value. For further details, we refer the reader to [56].

3.2 Alignment with the Human

To learn alignment with human drivers, we expose the assistive agent to different humans at training time. Training with human partners in a fictitious co-play setting [23] allows the assistant to become robust to different possible human behaviors, but does not teach the assistant to distinguish between the discrete modes of human behavior, which is necessary for assistance in our driving task.

Similar to recent work describing human preferences for a task using reward shaping [57, 58], we generate a population of humans, but different from these works, we group labeled sub-populations according to certain inherent objectives or preferences. We train several sub-populations using a separate human reward $r_t^H = r_t^{task} + r_t^y$, which is composed of the base task rewards in (1), with the addition of several distinct *human objectives*, spanning different multimodal behaviors of the human. These are then formed into a collection of humans $\{\pi_1^H, \pi_2^H, \ldots, \pi_N^H\} \in \{\Pi_1^H, \Pi_2^H, \ldots, \Pi_N^H\}$ for N different objectives. Examples of different settings for r_t^y are given in the supplementary material.

If trained to optimality, such rewards induce optimal behaviors for each human objective $\{\pi_1^*, \pi_2^*, \ldots, \pi_N^*\}$, which we can consider as the collection of *expert policies* of each enumerated objective. We extend this notion to that of *intent* at runtime. At any given moment, the human may adopt an intent y to abide by policy π_y^H. To simplify training, we consider drivers and rollouts with fixed intents, y, though our problem setup generalizes to variable-intent scenarios (e.g., a human switching from "following" to "passing" during a race).

Humans are error-prone and pursue their objectives irrationally [59]. To capture this, we partially train human partners to pair with the assistive agent in a process known as fictitious co-play (FCP) [23], yielding a population of humans $\{\pi_{y,k}^H\}_k \cup \{\pi_y^*\}$ for each intent sub-population Π_y^H, where k represents the policy checkpoint for some amount of completed training. Using the FCP framework, each checkpoint $k \in [1, K]$ is an agent trained to $k\%$ completion and the final checkpoint $k = K$ is an optimal policy. Each (sub)optimal human is endowed with a (sub)optimal RSSM world model. While we apply FCP in this work, we note that our approach is agnostic to how human behaviors are generated and future work may consider other approaches to synthesizing human policies [60, 61].

3.3 Assistive Agent Objectives

For continuous actions $\boldsymbol{a}_t^H, \boldsymbol{a}_t^A \in \mathbb{R}^m$ for the human and assistant, respectively, we encourage the assistive agent to minimize its action intervention by adopting a reward of the form

$$r_t^{interv} = -\|\boldsymbol{a}_t^A\|_2$$

which penalizes the magnitude of the intervention according to an L_2-norm. We note that aggregated over time, the intervention cost forms a mixed-norm $l_1 - l_2$ regularization [62], encouraging the intervention to be sparse over time.

In addition to the intervention norm penalty, the agent is rewarded for task completion and penalized for driving out of bounds or for collisions. These sparse negative rewards serve to encourage the assistant to act as a "racing guardian", keeping the driver on the track and avoiding collisions while making progress towards the finish line. Note that these rewards align with the *task* objective (Sec 3.1), but do not favor any particular *human* objective or intent.

Finally, the assistive agent receives reward for aligning with the inferred objective of the *human driver*. First, the agent receives a reward equal to the inferred reward under the optimal human's world model. In other words, at every time step t, the agent predicts the human's intent, $\hat{y}_t^A$ and then queries the world model corresponding to the optimal human policy $W^*_{\phi(\hat{y}_t^A)}$ for its reward prediction $\hat{r}^*(\hat{y}_t^A)$, for the current state-action pair. Second, the agent receives reward for maximizing the likelihood of the human-robot's combined action under the optimal human model for the predicted human intent:

$$\pi^*_{\hat{y}_t^A}(\boldsymbol{a}_t^A + \boldsymbol{a}_t^H | \boldsymbol{x}_t) := \pi^*(\boldsymbol{a}_t^A + \boldsymbol{a}_t^H | \boldsymbol{x}_t, \hat{y}_t^A) \tag{5}$$

Note that $\boldsymbol{a}_t^H$ is the current sub-optimal human action. Intuitively, these rewards encourage the assistant to: (1) select actions for which the optimal human assigns high value and (2) select actions that bring the sub-optimal human closer to the optimal human. Under this setup, although we infer $\hat{y}_t^A$ from the joint human-robot trajectory, we consider it to be more aligned with the *human's* intent, due to: (1) the presence of the intervention sparsity term r_t^{interv}, and (2) the fact that the inferred expert alignment term tilts the joint behavior toward the inferred expert human's behavior.

The complete reward for the assistant is a combination of rewards that promote successful task execution, minimize the magnitude of intervention, and account for human preferences:

$$r_t^A = \underbrace{r^{task} + r_t^{interv}}_{\text{task performance and intervention sparsity}} + \underbrace{\alpha_r \hat{r}_t^*(\hat{y}_t^A) + \alpha_a \|\boldsymbol{a}_t^H + \boldsymbol{a}_t^A - \boldsymbol{a}_t^*(\hat{y}_t^A)\|_2}_{\text{expert alignment}} \tag{6}$$

Here, scalars α_r and α_a weight the contributions of the human terms. We capture both the inferred optimal reward value, $\hat{r}^*(\hat{y}_t^A)$, as well as the similarity between the combined human-robot actions and an action from the inferred optimal policy $\hat{\boldsymbol{a}}^*(\hat{y}_t^A) \sim \pi^*_{\hat{y}_t^A}(\boldsymbol{x}_t)$. Note that, though we empirically compare different values of α_r and α_a in the supplementary material, in the next section, we only examine results for the reward-only case ($\alpha_r = 1$, $\alpha_a = 0$).

4 Experiments and Results

We examine the performance of our approach, dubbed DREAM2ASSIST, on different racing tasks with a fictitious human driver. We examine two racing settings, each derived from portions of a two-mile race track, and implemented in the CARLA Simulator [63] using a rear-wheel race vehicle physics model. For each task, opponents are randomly instantiated as replays of real human trajectories from the track, meaning that they do not react to the ego vehicle. In each setting, we run two sets of experiments—one with the *pass vs. no-pass* humans, and one with the *left vs. right* humans.

In each experiment, we train a population of humans for each human objective (e.g., "left" or "right"), and then train a DREAM2ASSIST agent over the combined population. We then evaluate the degree to which the assistive agent can improve fictitious human performance on the track, where high performance means high track progress, high average speed, and no collisions. To measure the contributions of each assistive agent for the fictitious human population, we sample checkpoints at every 20% performance increment for agents up to at least 75% of maximum, or from the bottom five performers if none are under this threshold. We report the mean change in performance (track **progress** and **collisions**) when an assistive agent is deployed to the car, as well as the **return** under each human objective in Table 1. Means and standard deviations are computed over all five sub-optimal fictitious drivers.

The two settings we consider include **Straightaway Driving** and **Hairpin Driving**. The straightaway is a flat 370-meter portion of a track with a concrete barrier on the right-hand side. Because the section of road is straight and flat, drivers need only avoid colliding with the concrete wall or their opponent while racing to the finish. The hairpin section, however, is a 960-meter portion of a track involving several sharp turns and hills. Drivers must carefully manage their speed to avoid spin-outs or collisions, and passing is much more challenging in this section of track.

Table 1: Results on straightaway and hairpin experiments. Blue indicates improvement, **bold** is best.

	Pass (top) / Stay (bottom)						Left (top) / Right (bottom)					
	Hairpin			Straightaway			Hairpin			Straightaway		
	Progress ↑	Return ↑	Collisions ↓	Progress ↑	Return ↑	Collisions ↓	Progress ↑	Return ↑	Collisions ↓	Progress ↑	Return ↑	Collisions ↓
DREAMER	-11.3± 13.6	**0.5± 2.9**	-0.1± 0.2	-0.7± 4.6	**-0.1± 0.9**	0.0± 0.1	21.8± 28.7	-2.0± 2.2	0.1± 0.1	**10.8± 7.3**	-0.6± 0.3	0.0± 0.1
	-21.1± 68.6	0.3± 0.4	**0.0± 0.0**	-6.1± 17.5	**0.0± 0.1**	0.1± 0.2	10.0± 17.9	-1.3± 1.8	0.0± 0.1	**-1.1± 6.2**	**0.2± 1.2**	0.0± 0.1
DREAMER-AIL	-28.9± 46.2	-5.7± 7.3	**-0.3± 0.2**	-116.6± 58.7	-9.7± 4.2	**-0.4± 0.3**	-144.0± 89.9	-7.1± 6.6	**-0.4± 0.2**	-134.4± 13.0	-2.5± 0.9	**-0.5± 0.1**
	-216.6± 145.6	-1.7± 0.4	0.1± 0.2	-52.5± 45.9	-1.6± 0.2	**-0.1± 0.1**	-119.1± 79.9	-4.5± 17.3	**-0.4± 0.3**	-126.3± 53.2	-13.4± 7.0	**-0.5± 0.2**
DREAM2ASSIST	**71.8± 43.9**	-1.2± 3.3	-0.2± 0.2	**6.0± 9.1**	-0.4± 1.4	-0.1± 0.1	**54.8± 60.4**	**1.4± 5.2**	0.0± 0.1	5.0± 5.0	**0.1± 0.5**	0.0± 0.1
	60.5± 49.7	**0.8± 0.4**	0.1± 0.1	**57.8± 36.4**	-0.1± 0.2	0.1± 0.1	**27.2± 23.1**	**2.2± 2.3**	-0.2± 0.2	**-1.1± 31.91**	-2.8± 2.4	0.2± 0.2

Human Decision Characteristics Based on observed behaviors from a human study (see supplementary material Sec. B), we propose two different sets of human characteristics (ground truth intent): **pass vs. no-pass** and **left vs. right**. We train fictitious human agents to satisfy each of these objectives. For the *pass vs. no-pass* behaviors, the fictitious human agents are trained to either overtake or stay-behind the opponent vehicle. For the *left vs. right* behaviors, the fictitious humans are trained with a preference to stay on either the left or right side of the opponent for as much of the race as possible. Attempting to provide the wrong type of assistance with such distinct behaviors (e.g., offering "right" assistance to a "left" human) will result in fighting between the assistant and human, and will likely lead to collisions or spin-outs. For further details on the rewards for each agent, see supplementary material Sec. C.

Action and Observation Spaces The environment itself consists of the rewards in Sec. 3.1 and an observation and action space for each agent. The human and assistive agent's actions both adopt steering and acceleration values in the range of $[-1, 1]$. The assistant and human actions are summed together and clipped on $[-1, 1]$ before being passed to the vehicle. The observation space contains the ego vehicle state (position, velocity, tire slip angles, yaw rate, heading), current distance traveled, and an array of forward-looking track edge points. Finally, the human and assistive agents are each able to observe the other's actions.

Baselines The claim of our work is that an assistive agent will better support a human partner if the assistant can infer the human's objective and help to satisfy that objective. To test this claim, we compare **DREAM2ASSIST** against a baseline version of **DREAMER**, which makes no intent inference and is not rewarded according to a human objective. We also compare to a non-RSSM baseline, GAIL [64], obtained with the BeTAIL [65] framework. BeTAIL uses a behavior transformer as the human policy, coupled with an assistive agent trained via adversarial imitation learning to correct for distribution shift. Our **DREAMER-AIL** baseline challenges whether our RSSM approach is necessary at all, or whether a purely data-driven approach using behavior cloning and inverse reinforcement learning could satisfy the multimodal behavioral assistance task that we consider. For this baseline, a fictitious human using a DREAMER policy is paired with the AIL assistive agent from BeTAIL.

4.1 Straightaway Results

We observe a consistent trend for all straightaway results – DREAMER offers very low-magnitude intervention, leading to higher performance with expert drivers but poor performance with novice drivers (as the assistant is not helping). Conversely, DREAM2ASSIST offers higher-magnitude intervention, leading to higher performance with novice drivers, but lower performance gains with expert drivers (as the assistant is not needed). DREAMER-AIL suffers from mode collapse, trying to

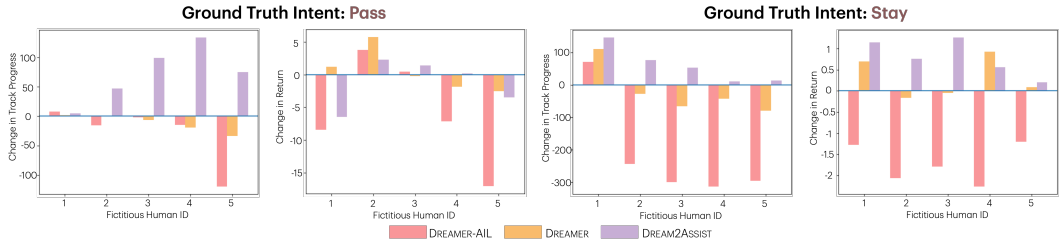

Figure 3: Change in track progress and return when adding assistance to five imperfect pass (left) and stay (right) fictitious humans in the hairpin problem setting. The addition of DREAM2ASSIST leads to higher gains in task performance and greater adherence to human objectives than baselines.

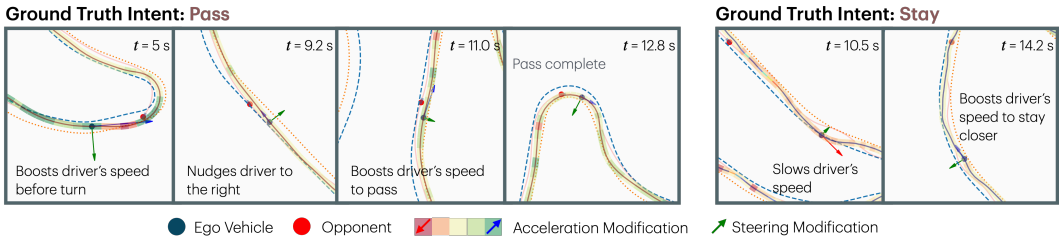

Figure 4: Examples of the DREAM2ASSIST agent's actions when paired with a human intending to pass and a human intending to stay. DREAM2ASSIST recognizes the driver's intent, making lateral corrections for a safer overtake (left) or throttle adjustments to stay behind the opponent while still progressing towards the finish (right), thereby helping to satisfy *task* and *human* objectives.

turn all drivers into either a "stay" driver or getting caught between "left" and "right" and therefore not moving. This behavior means that the DREAMER-AIL agent consistently underperforms a fictitious human with no assistance at all, as the AIL assistance keeps the driver far behind the opponent.

4.2 Hairpin Results

The hairpin domain is more challenging and yields fictitious humans that are not always consistently able to solve the task, thereby leaving greater scope for assistance from our trained agents. In our *pass vs. stay* experiment, DREAM2ASSIST significantly improves the drivers' abilities to solve the task while still satisfying the human objectives and not leading to an increase in collisions. Similarly, in the *left vs. right* experiment, DREAM2ASSIST leads to significant increases in track progress, reduction in collisions, and improvements in human-objective alignment. Baseline approaches fail to disentangle the "left" and "right" modes of driving. The DREAMER baseline instead opts to push the driver to stay behind the opponent vehicle in an effort to reduce collisions, thereby making it farther down the track but failing to overtake. The DREAMER-AIL baseline drives aggressively off the track, leading to a significant drop in track progress, collisions, and human-objective alignment. An illustrative example of DREAM2ASSIST is in Fig. 4, and videos are in the supplementary material.

Figure 3 provides an overview of the amount of assistance provided to drivers for both track progress and human-objective return for assistants deployed to the *pass vs. stay* problem in the hairpin section. Visualizing the assistance provided to each driver, we can more easily compare the scale of assistance provided by DREAM2ASSIST compared to baselines, showing that DREAM2ASSIST provides marked improvements relative to baseline assistive agents. Additional figures for each domain and driver-population are given in the supplementary material.

5 Conclusions and Limitations

We introduce an assistive paradigm, DREAM2ASSIST, that learns to interact with humans to help them perform more optimally while supporting their personal objectives for the task. We evaluate DREAM2ASSIST in a dynamic and challenging task of high-speed racing, and we show that our approach is able to disentangle and accommodate distinct human objectives more effectively than baseline methods. We show that DREAM2ASSIST results in higher human-robot team performance than baseline methods, suggesting that explicit intent-conditioning and reward-inference can provide crucial performance gains in settings with multimodal, mutually-exclusive, human objectives.

While DREAM2ASSIST represents a state-of-the-art improvement in human-robot teaming, there are limitations in our work that we hope to address in future work. First, our approach has been tested with fictitious humans, but we have not yet evaluated generalization to real human-robot teams. In future work, we intend to deploy DREAM2ASSIST in a human-subjects experiment to test how effectively our framework generalizes to real, sub-optimal human drivers. DREAM2ASSIST also relies on privileged access to the inferred reward values from the RSSM of an optimal policy; future work may consider how to estimate optimal policy rewards without such a model. Finally, future work may consider how to provide assistance via multiple modalities (e.g., providing sparse language guidance on when to overtake vs. dense control-level assistance during the overtake).

References

[1] M. Romeo, P. E. McKenna, D. A. Robb, G. Rajendran, B. Nesset, A. Cangelosi, and H. Hastie. Exploring theory of mind for human-robot collaboration. In *International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 461–468. IEEE, 2022.

[2] T. Reimer, E. S. Park, and V. B. Hinsz. Shared and coordinated cognition in competitive and dynamic task environments: An information-processing perspective for team sports. *International Journal of Sport and Exercise Psychology*, 4(4):376–400, 2006.

[3] R. Giske, J. Gjerde, R. Hoigaard, and B. T. Johansen. Importance, characteristics, and development. *Journal of Expertise*, 5(1):12–28, 2022.

[4] M. Demir, N. J. McNeese, and N. J. Cooke. Understanding human-robot teams in light of all-human teams: Aspects of team interaction and shared cognition. *Int. J. Hum. Comput. Stud.*, 140:102436, Aug. 2020.

[5] J. Braithwaite, R. A. Iedema, and C. Jorm. Trust, communication, theory of mind and the social brain hypothesis: deep explanations for what goes wrong in health care. *Journal of Health Organization and Management*, 21(4/5):353–367, 2007.

[6] H. Summala. Brake reaction times and driver behavior analysis. *Transportation Human Factors*, 2(3):217–226, 2000.

[7] D. A. Abbink, T. Carlson, M. Mulder, J. C. F. de Winter, F. Aminravan, T. L. Gibo, and E. R. Boer. A topology of shared control systems—finding common ground in diversity. *IEEE Transactions on Human-Machine Systems*, 48(5):509–525, Oct. 2018.

[8] E. Weiss and J. Christian Gerdes. Follow my lead: Designing an ADAS that shares decision making and control with the driver. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1482–1487. IEEE, Oct. 2023.

[9] N. Deo and M. M. Trivedi. Multi-modal trajectory prediction of surrounding vehicles with maneuver based lstms. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1179–1184. IEEE, 2018.

[10] A. Censi, K. Slutsky, T. Wongpiromsarn, D. Yershov, S. Pendleton, J. Fu, and E. Frazzoli. Liability, ethics, and culture-aware behavior specification using rulebooks. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8536–8542. IEEE, 2019.

[11] R. Karagulle, N. Aréchiga, J. DeCastro, and N. Ozay. Classification of driving behaviors using stl formulas: A comparative study. In *International Conference on Formal Modeling and Analysis of Timed Systems*, pages 153–162. Springer, 2022.

[12] X. Li, G. Rosman, I. Gilitschenski, C.-I. Vasile, J. A. DeCastro, S. Karaman, and D. Rus. Vehicle trajectory prediction using generative adversarial network with temporal logic syntax tree features. *IEEE Robotics and Automation Letters*, 6(2):3459–3466, 2021.

[13] T. B. Sheridan. Human–robot interaction: status and challenges. *Human Factors*, 58(4): 525–532, 2016.

[14] L. Onnasch and E. Roesler. A taxonomy to structure and analyze human–robot interaction. *International Journal of Social Robotics*, 13(4):833–849, July 2021.

[15] A. Xie, D. Losey, R. Tolsma, C. Finn, and D. Sadigh. Learning latent representations to influence multi-agent interaction. In J. Kober, F. Ramos, and C. Tomlin, editors, *Conference on Robot Learning*, pages 575–588. PMLR, 2021.

[16] S. Reddy, A. D. Dragan, and S. Levine. Shared autonomy via deep reinforcement learning. Feb. 2018.

[17] M. Tucker, Y. Zhou, and J. Shah. Latent space alignment using adversarially guided self-play. *International Journal of Human–Computer Interaction*, pages 1753–1771, Aug. 2022.

[18] S. Karamcheti, M. Srivastava, P. Liang, and D. Sadigh. LILA: Language-Informed latent actions. In A. Faust, D. Hsu, and G. Neumann, editors, *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 1379–1390. PMLR, 2022.

[19] A. Bobu, Y. Liu, R. Shah, D. S. Brown, and A. D. Dragan. SIRL: Similarity-based implicit representation learning. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 565–574, 2023.

[20] I. Sucholutsky, L. Muttenthaler, A. Weller, A. Peng, A. Bobu, B. Kim, B. C. Love, E. Grant, J. Achterberg, J. B. Tenenbaum, et al. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023.

[21] S. Parekh and D. P. Losey. Learning latent representations to co-adapt to humans. *Autonomous Robots*, 47(6):771–796, 2023.

[22] H. J. Jeon, D. P. Losey, and D. Sadigh. Shared autonomy with learned latent actions. *ArXiv*, abs/2005.03210, May 2020.

[23] D. Strouse, K. McKee, M. Botvinick, E. Hughes, and R. Everett. Collaborating with humans without human data. *Advances in Neural Information Processing Systems*, 34:14502–14515, 2021.

[24] R. C. Kelley. Models of intention for human-robot interaction. 2013.

[25] S. Thill and T. Ziemke. The role of intentions in human-robot interaction. In *International Conference on Human-Robot Interaction*, HRI '17, pages 427–428, New York, NY, USA, Mar. 2017. Association for Computing Machinery.

[26] D. P. Losey, C. G. McDonald, E. Battaglia, and M. K. O'Malley. A review of intent detection, arbitration, and communication aspects of shared control for physical human–robot interaction. *Applied Mechanics Reviews*, 70(1):010804, 2018.

[27] Y. Cheng, L. Sun, C. Liu, and M. Tomizuka. Towards efficient human-robotcollaboration with robust plan recognition and trajectory prediction. *IEEE Robotics Automation and Letters*, 5(2): 2602–2609, Apr. 2020.

[28] C. Lebiere, F. Jentsch, and S. Ososky. Cognitive models of decision making processes for human-robot interaction. In *Virtual Augmented and Mixed Reality. Designing and Developing Augmented and Virtual Environments: 5th International Conference, VAMR 2013*, pages 285–294. Springer, 2013.

[29] B. Mutlu, N. Roy, and S. Šabanović. Cognitive human–robot interaction. *Springer Handbook of Robotics*, pages 1907–1934, 2016.

[30] H. Admoni and B. Scassellati. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63, 2017.

[31] C. G. Bobier, S. Joe, and J. C. Gerdes. Sliding surface envelope control: Keeping the vehicle within a safe state-space boundary. In *Dynamic Systems and Control Conference*, volume 44175, pages 149–156, 2010.

[32] C. G. Bobier and J. C. Gerdes. Staying within the nullcline boundary for vehicle envelope control using a sliding surface. *Veh. Syst. Dyn.*, 51(2):199–217, Feb. 2013.

[33] J. Talbot, M. Brown, and J. C. Gerdes. Shared control up to the limits of vehicle handling. *IEEE Transactions on Intelligent Vehicles*, pages 2977 – 2987, 2023. doi:10.1109/TIV.2023.3300989.

[34] W. Schwarting, J. Alonso-Mora, L. Paull, S. Karaman, and D. Rus. Parallel autonomy in automated vehicles: Safe motion generation with minimal intervention. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1928–1935, May 2017.

[35] A. Benloucif, A.-T. Nguyen, C. Sentouh, and J.-C. Popieul. Cooperative trajectory planning for haptic shared control between driver and automation in highway driving. *IEEE Transactions on Industrial Electronics*, 66(12):9846–9857, 2019.

[36] Y. Liang, Z. Yin, and L. Nie. Shared steering control for lane keeping and obstacle avoidance based on multi-objective mpc. *Sensors*, 21(14):4671, 2021.

[37] P. G. Griffiths and R. B. Gillespie. Sharing control between humans and automation using haptic interface: Primary and secondary task performance benefits. *Human Factors: The Journal of Human Factors and Ergonomics Society*, 47:574–590, Sept. 2005.

[38] M. Mulder, D. A. Abbink, and E. R. Boer. Sharing control with haptics: seamless driver support from manual to automatic control. *Human Factors*, 54(5):786–798, Oct. 2012.

[39] A. Balachandran, M. Brown, S. M. Erlien, and J. Christian Gerdes. Predictive haptic feedback for obstacle avoidance based on model predictive control. *IEEE Trans. Autom. Sci. Eng.*, 13(1): 26–31, Jan. 2016.

[40] A. H. Ghasemi. Game theoretic modeling of a steering operation in a haptic shared control framework. *ASME 2018 Dynamic Systems and Control Conference*, Nov. 2018.

[41] S. C. Jugade, A. C. Victorino, and V. B. Cherfaoui. Shared driving control between human and autonomous driving system via conflict resolution using non-cooperative game theory. In *ITSC*, pages 2141–2147. IEEE, Oct. 2019.

[42] H. Deng, Y. Zhao, S. Feng, Q. Wang, and F. Lin. Shared control for intelligent vehicle based on handling inverse dynamics and driving intention. *IEEE Trans. Veh. Technol.*, 71(3):2706–2720, Mar. 2022.

[43] J. Xie, X. Xu, F. Wang, Z. Liu, and L. Chen. Coordination control strategy for Human-Machine cooperative steering of intelligent vehicles: A reinforcement learning approach. *IEEE Trans. Intell. Transp. Syst.*, 23(11):21163–21177, Nov. 2022.

[44] M. Marcano, S. Díaz, J. Pérez, and E. Irigoyen. A review of shared control for automated vehicles: Theory and applications. *IEEE Transactions on Human-Machine Systems*, 50(6): 475–491, Dec. 2020.

[45] J. Sarabia, M. Marcano, J. Pérez, A. Zubizarreta, and S. Diaz. A review of shared control in automated vehicles: System evaluation. *Frontiers in Control Engineering*, 3, 2023.

[46] S. Lefèvre, D. Vasquez, and C. Laugier. A survey on motion prediction and risk assessment for intelligent vehicles. *ROBOMECH Journal*, 1(1):1–14, July 2014.

[47] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras. Human motion trajectory prediction: a survey. *Int. J. Rob. Res.*, 39(8):895–935, July 2020.

[48] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2019.

[49] A. Zholus, Y. Ivchenkov, and A. Panov. Factorized world models for learning causal relationships. In *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality*, Mar. 2022.

[50] C. Gumbsch, N. Sajid, G. Martius, and M. V. Butz. Learning hierarchical world models with adaptive temporal abstractions from discrete latent dynamics. In *The Twelfth International Conference on Learning Representations*, 2023.

[51] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap. Mastering diverse domains through world models. Jan. 2023.

[52] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, pages 2555–2565. PMLR, 2019.

[53] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[54] L. Chen, S. M. J. D. J. Subosits, and P. Tylkin. Learn thy enemy: Online, task-aware opponent modeling in autonomous racing. 2023.

[55] P. R. Wurman, S. Barrett, K. Kawamoto, J. MacGlashan, K. Subramanian, T. J. Walsh, R. Capobianco, A. Devlic, F. Eckert, F. Fuchs, L. Gilpin, P. Khandelwal, V. Kompella, H. Lin, P. MacAlpine, D. Oller, T. Seno, C. Sherstan, M. D. Thomure, H. Aghabozorgi, L. Barrett, R. Douglas, D. Whitehead, P. Dürr, P. Stone, M. Spranger, and H. Kitano. Outracing champion Gran Turismo drivers with deep reinforcement learning. *Nature*, 602(7896):223–228, Feb. 2022.

[56] D. Hafner, T. Lillicrap, M. Norouzi, and J. Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

[57] Z. Tang, C. Yu, B. Chen, H. Xu, X. Wang, F. Fang, S. Du, Y. Wang, and Y. Wu. Discovering diverse multi-agent strategic behavior via reward randomization. *arXiv preprint arXiv:2103.04564*, 2021.

[58] C. Yu, J. Gao, W. Liu, B. Xu, H. Tang, J. Yang, Y. Wang, and Y. Wu. Learning zero-shot cooperation with humans, assuming humans are biased. *arXiv preprint arXiv:2302.01605*, 2023.

[59] S. E. Chipman. *The Oxford Handbook of Cognitive Science*. Oxford University Press, 2016.

[60] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023.

[61] D. Venuto, S. N. Islam, M. Klissarov, D. Precup, S. Yang, and A. Anand. Code as reward: Empowering reinforcement learning with VLMs. *arXiv preprint arXiv:2402.04764*, 2024.

[62] M. Elad. *Sparse and Redundant Representations*. Springer New York.

[63] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. CARLA: An open urban driving simulator. In *Conference on Lobot learning*, pages 1–16. PMLR, 2017.

[64] J. Ho and S. Ermon. Generative adversarial imitation learning. *Advances in Neural Information Processing Systems*, 29, 2016.

[65] C. Weaver, C. Tang, C. Hao, K. Kawamoto, M. Tomizuka, and W. Zhan. BeTAIL: Behavior transformer adversarial imitation learning from human racing gameplay. *arXiv preprint arXiv:2402.14194*, 2024.