

Appendix:

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

Anonymous Author(s)

Affiliation

Address

email

1 A MBRL Preliminaries

2 We frame the model-based reinforcement learning (MBRL) problem as a two-player (human ego
3 driver, assistive agent ego driver) partially-observable Markov decision process (POMDP) defined
4 by the tuple $M = \langle \mathcal{X}^\kappa, \mathcal{A}^\kappa, \mathcal{T}^\kappa, \mathcal{R}^\kappa, \gamma \rangle^{\kappa=H,A}$, where, for agent $\kappa \in \{H, A\}$ (for the human, and AI
5 agent, respectively), \mathcal{X}^κ denotes the imagined states of the world, \mathcal{A}^κ denotes the agent’s (continuous
6 or discrete) actions, $\mathcal{T}^\kappa : \mathcal{X}^\kappa \times \mathcal{A}^\kappa \mapsto [0, 1]$ is the transition probability, $\mathcal{R}^\kappa : \mathcal{X}^\kappa \times \mathcal{A}^\kappa \times \mathcal{X}^\kappa \mapsto \mathbb{R}$
7 is a reward function, and $\gamma \in [0, 1]$ is a discount factor. We aim to train both agents such that they
8 maximize their expected returns $R^\kappa = \mathbb{E} \left[\sum_{t=1}^T r_t^\kappa \right]$.

9 Crucially, in the semi-cooperative shared control setting, each reward r_t^κ is factored into sub-
10 components, with both sharing the same task (driving) rewards, but where r_t^H contains an additional
11 term for a human’s objective, and r_t^A contains additional terms to weaken its contribution in relation
12 to the human’s and enforces alignment to the human.

13 B Human Subject Data Collection

14 We briefly discuss a study conducted for gathering human subject behavior data in the racing domain
15 we use in the paper. The purpose of the study was to gather qualitative and statistical data on
16 individuals’ behavior and objectives in a racing context, and to use that to inform what criteria
17 are important for building models of human objectives. We recruited 48 participants to drive a
18 simulator with the hairpin and straightaway segments of the two-mile track, the same domains for the
19 computational results in this paper. The scenarios were chosen so as to present overtake opportunities
20 in portions of the track of varying levels of difficulty, while keeping the overall task short enough to
21 ensure there is a rich interaction between the ego and opponent. Participants completed a series of
22 warm-up trials in each domain, with three trials devoted to the straightaway segment and eight trials
23 in the hairpin segment, each featuring different opponents of varying difficulty (fixed trajectories) to
24 race against. Again, these were the same trajectories used in our domains.

25 At the conclusion of each trial, participants answered the question: “Did you attempt to pass the other
26 vehicle?” on an iPad. We also gathered, from trajectory data, whether or not the participant actually
27 completed an overtake without collisions or spin-outs. These results are reported in Table B.1. We
28 conclude that even in a simulated setting, there were a lower number of actual overtakes that occurred
29 than were attempted. This suggests that there is room to assist those wishing to overtake, but unable
30 to do so.

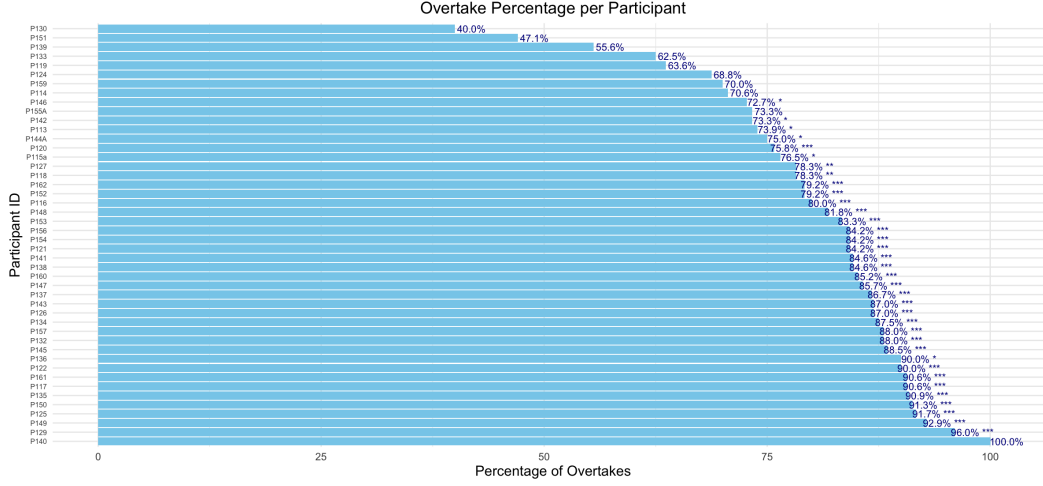


Figure B.1: Consistency of overtake versus non-overtakes.

Table B.1: Number and percentage of overtakes occurred and attempted. Note the diversity in intent and in overtaking-difficulty for the subjects, motivating the need for assistive shared autonomy.

Overtake occurred?	Frequency	Percentage	Overtake attempted?	Frequency	Percentage
“No”	178	30.07	“No”	65	11.02
“Yes”	414	69.93	“Yes”	525	88.98

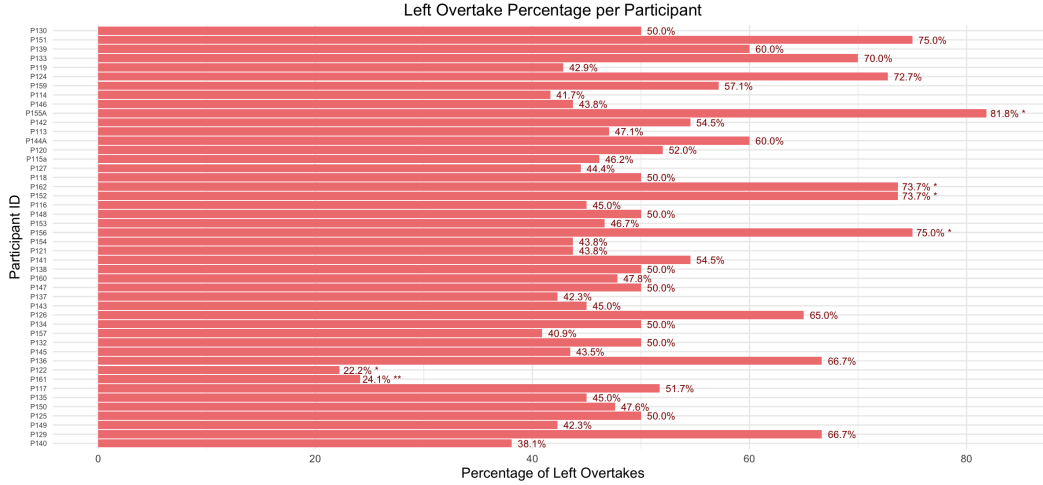


Figure B.2: Consistency of left-handed versus right-handed overtakes.

31 We consider additional statistics, including statistics on left- and right-hand passing, as well as
32 collisions with the other vehicle or objects, and spin-outs. We include these results in Table B.2. We
33 note that there is a nearly-equal number of overtakes on the right versus left. On an individual level,
34 we ran chi-square tests of test for given probabilities to look for side preferences. We found that
35 only 6 of the 48 participants showed a statistically significant ($p < .05$) passing side bias, with two
36 participants having a bias for the right side and four participants having a bias for the left side B.2.
37 We also note that participants were, in general, imperfect in their driving, with nearly 50% of trials
38 having a collision and 8% having at least one spin-out.

39 C Human Objectives

40 In this section, we discuss the reward terms used to generate the explicit decision-making tendencies
41 of the fictitious human drivers, via $r_t^H = r_t^{task} + r_t^y$. Each use the task-specific reward terms outlined

Table B.2: Other effects. Percentages are percent of trials with the listed event.

Observed event	Percentage
Left-handed overtakes	50.82%
Right-handed overtakes	49.18%
Collisions	48.66%
Spin-outs	8.36%

in (1), in combination with objective-specific rewards. Many of the task rewards are borrowed from [1]. We focus here on the human objective term r_t^y .

Pass We adopt a dense reward that provides a penalty when the vehicle is behind the opponent vehicle, and a reward bonus when in front of that vehicle, up to a threshold, to incentivize passing. That is,

$$r_t^y = c_{pass} (\Delta s_t - \Delta s_{t-1}) \mathbb{I}((s_{low} \leq \Delta s_t \leq s_{high}) \vee (s_{low} \leq \Delta s_{t-1} \leq s_{high}))$$

Where $\mathbb{I}(\cdot)$ is the indicator function, and Δs_t is the difference in longitudinal positions, relative to track coordinates, between the ego and opponent vehicles, $\Delta s_t = s_t^{ego} - s_t^{opp}$. We take the scalar $c_{pass} = 10$. In other words, if the difference between s_t^{ego} (the ego position) and s_t^{opp} (the opponent position) is between s_{low} and s_{high} , the passing reward is equal to $10 * (s_t^{ego} - s_t^{opp})$. This means there is a high positive reward for getting far ahead of the opponent, and a high negative reward for falling behind the opponent. For both the *pass* reward, we set $s_{high} = -s_{low} = 800$, which ensures that the pass reward is active for the entire trial. Note that we do not impose a progress reward with the pass objective.

Stay-Behind Due to the non-symmetry of the problem, the stay-behind reward cannot be the complement of the pass reward (otherwise, the stay-behind agent would drive backwards to get away from the opponent). Because the task reward does not consider making progress, we add that here to the human-specific reward. The stay-behind reward is then:

$$r_t^y = r_t^{prog} + c_{stay} (\Delta s_t - \Delta s_{t-1}) \mathbb{I}((s_{low} \leq \Delta s_t \leq s_{high}) \vee (s_{low} \leq \Delta s_{t-1} \leq s_{high}))$$

Where $c_{stay} = -2$, and we impose a progress reward similar to [1], i.e., $r_t^{prog} = s_t^{ego} - s_{t-1}^{ego}$, with s_t^{ego} being the ego's longitudinal position in track coordinates. For the *stay-behind* reward, we set $s_{high} = -s_{low} = 50$. In practice, this means that the stay-behind agent is encouraged to make progress along the track (r^{prog}), but to stay at least 50 meters behind the opponent.

Both the left- and right-biased passing agents are passing agents with an additional reward term that encourages a bias to the left or right. Note that these additional treatments do not guarantee passing on one side or the other.

Left-Biased We adopt a reward bonus for driving on the opponent's left; i.e.

$$r_t^y = (\Delta s_t - \Delta s_{t-1}) \mathbb{I}((s_{low} \leq \Delta s_t \leq s_{high}) \vee (s_{low} \leq \Delta s_{t-1} \leq s_{high})) + (\Delta e_t + c_{margin})$$

where Δe_t is the difference in lateral positions of the two vehicles, in the track coordinate frame; i.e. $\Delta e_t = e_t^{ego} - e_t^{opp}$, and c_{margin} is a margin (which we set to $c_{margin} = 0.3$).

Right-Biased Right-biased reward is the complement of the left-biased reward:

$$r_t^y = (\Delta s_t - \Delta s_{t-1}) \mathbb{I}((s_{low} \leq \Delta s_t \leq s_{high}) \vee (s_{low} \leq \Delta s_{t-1} \leq s_{high})) - (\Delta e_t + c_{margin})$$

D Additional Model Details

We provide a summary of the DREAM2ASSIST training procedure in Algorithm 1. The procedure is split into two phases: the first is a human population generation phase in which we use the rewards in Sec. C to generate a population of humans included in the tuple $\langle \mathcal{W}_i^H, \Pi_i^H \rangle$ of world models and policies, respectively, and the expert human models denoted by the tuple $\langle W_i^*, \pi_i^* \rangle$ for each human objective y_i . The second phase entails drawing from samples of $\langle \{y_j\}_{j=0}^N, \{\mathcal{W}_j^H\}_{j=0}^N, \{\Pi_j^H\}_{j=0}^N \rangle$ using fictitious co-play (FCP) [2] in order to train the assistant's world model W^A and policy π^A . At runtime, both the trained policy π^A and world model W^A are executed, with W^A being additionally useful as a means to interpret the decisions made by π^A ; e.g. the intent estimate \hat{y}_t^A , the estimated reward \hat{r}_t^A or the latent variables \hat{z}_t .

Algorithm 1 DREAM2ASSIST using FCP

Given: diverse intents, $y_i \in \{1, 2, \dots, M\}$
Given: reward functions for each intent y
for $y_i, i \in \{1, 2, \dots, M\}$ **do** ▷ Generate human population
 Initialize π_i^H, W_i^H
 while not converged **do**
 Sample an opponent policy π_{opp} from Π_{opp}
 Initialize \mathbf{x}_0 , with $t = 0$
 while not done **do**
 Perform gradient step; update π_i^H, W_i^H
 Sample action $\mathbf{a}_{i,t}^H$ from π_i^H
 Step the environment with action $\mathbf{a}_{i,t}^H$ and $\mathbf{a}_t^{opp} \sim \pi^{opp}$
 Shape rewards according to i th agent reward
 $t \leftarrow t + 1$
 end while
 Append checkpoint $\langle W_i^H, \pi_i^H \rangle$ to $\langle \mathcal{W}_i^H, \Pi_i^H \rangle$
 end while
 Append final $\langle W_i^*, \pi_i^* \rangle$ to $\langle \mathcal{W}_i^H, \Pi_i^H \rangle$
end for
Freeze agents and world models $\{\mathcal{W}_j^H\}_{j=0}^N, \{\Pi_j^H\}_{j=0}^N$
Initialize π^A, W^A
while not converged **do** ▷ Train assistant agent
 Sample intent i from $\langle \{y_i\}_{i=0}^M \rangle$
 Sample checkpoint j for intent y_i from $\langle \{\mathcal{W}_j^H\}_{j=0}^N, \{\Pi_j^H\}_{j=0}^N \rangle$
 Sample an opponent policy π_{opp} from Π_{opp}
 Initialize \mathbf{x}_0 , with $t = 0$
 while not done **do**
 Perform gradient step; update π^A, W^A using ground truth label y_i
 Sample action \mathbf{a}_t^A from $\pi^A(\mathbf{x})$, \mathbf{a}_t^H from $\pi_j^H(\mathbf{x}_t)$
 Step the environment using shared action $\mathbf{a}_t^H + \mathbf{a}_t^A$ and $\mathbf{a}_t^{opp} \sim \pi^{opp}$
 Evaluate intent \hat{y}_t^A using W^A
 Shape rewards as per (6), using $r^*(\hat{y}_t^A)$ from $W_j^*, \mathbf{a}_t^*(\hat{y}_t^A) \sim \pi_{\hat{y}_t^A}^*(\mathbf{x}_t)$
 $t \leftarrow t + 1$
 end while
end while

Table C.1: Training Hyperparameters.

Hyperparameter	Value	Hyperparameter	Value
Encoder / decoder MLP layers	2	Steps	2e6
Encoder / decoder MLP units	512	Batch size	16
Predictor head layers	2	Batch length	64
Predictor head units	512	Training ratio	512
Discount factor	0.997	Model learning rate	1e-4
Discount λ	0.95	Value learning rate	3e-5
Imagined horizon	15	Actor learning rate	3e-5
Actor entropy	3e-4	Dataset max size	1e6
Dynamics hidden units	512	# Steps between evaluations	1e4
Dynamics discrete dimension	32	# Episodes to evaluate	10

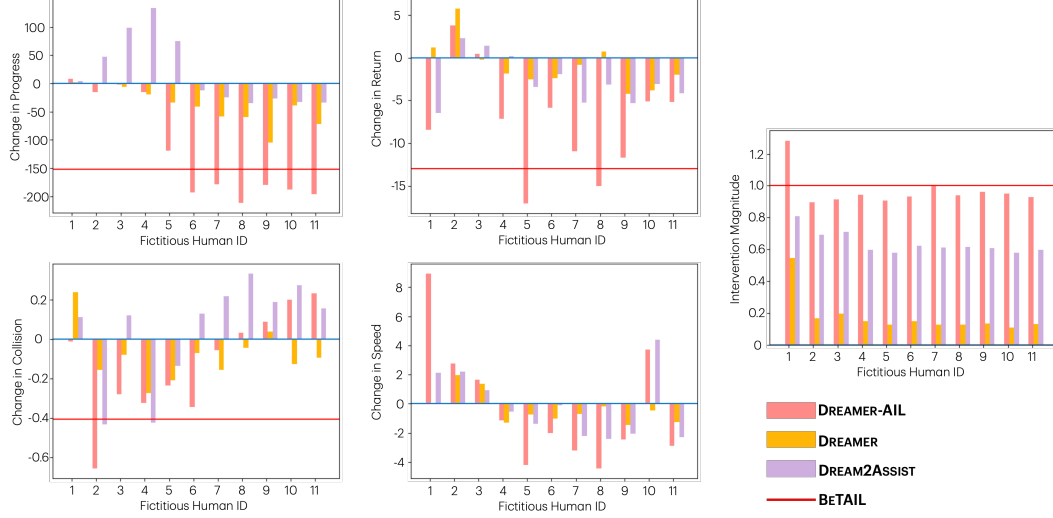


Figure D.1: Changes in various metrics in the hairpin scenario when adding assistance to various imperfect (1–5) and near-perfect (6–11) humans tending to *pass*, averaged over four random seeds. Due to the fact that BeTAIL uses its own internal human model, we compare only one instance / human, as denoted by the red line.

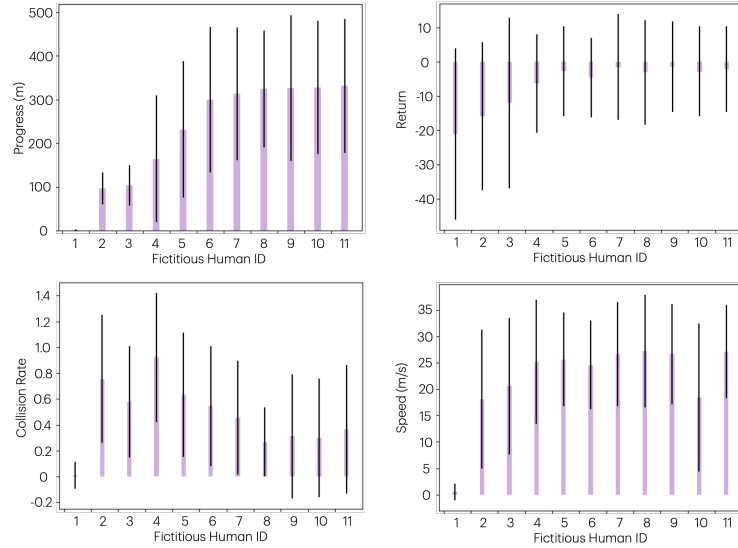


Figure D.2: Absolute metrics in the hairpin scenario evaluated for the *unassisted* imperfect (1–5) and near-perfect (6–11) humans tending to *pass*, averaged over four random seeds. 1- σ error bars are shown.

80 D.1 Training Hyperparameters and Environment Specifics

81 We provide the DREAM2ASSIST hyperparameters in Table C.1. We train using the Adam optimizer
 82 for 2×10^6 steps.

83 The CARLA simulator is used for our environment, and is executed with step size of 0.1 sec. We
 84 terminate episodes if: (a) the ego collides with the opponent or other collidable objects (e.g. static
 85 barriers), (b) the ego vehicle veers too far off course, or (c) a predefined finish line is reached. The
 86 map used is a geospatially-calibrated representation of the Thunderhill Raceway in Willows, CA.

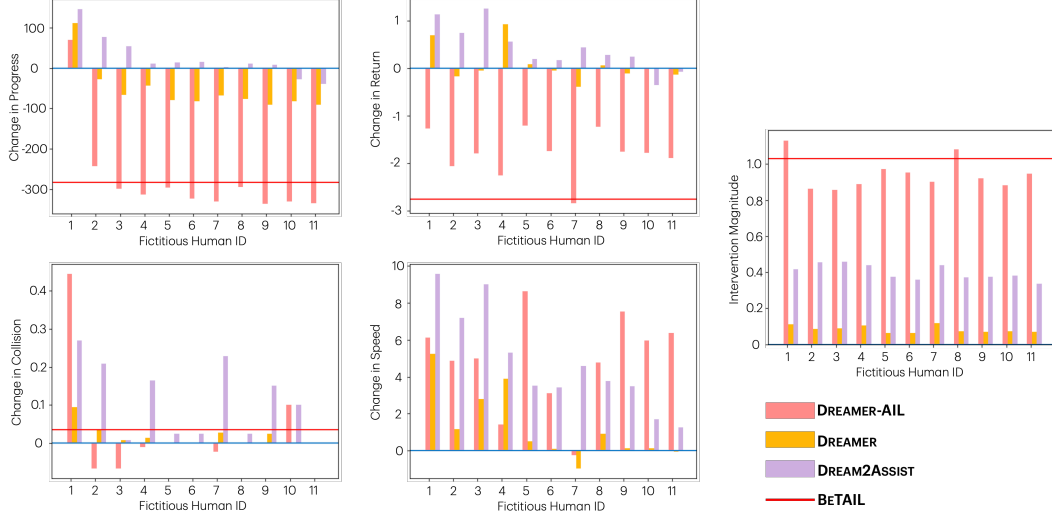


Figure D.3: Changes in various metrics in the hairpin scenario when adding assistance to various imperfect (1–5) and near-perfect (6–11) humans tending to *stay*, averaged over four random seeds. Due to the fact that BeTAIL uses its own internal human model, we compare only one instance / human, as denoted by the red line.

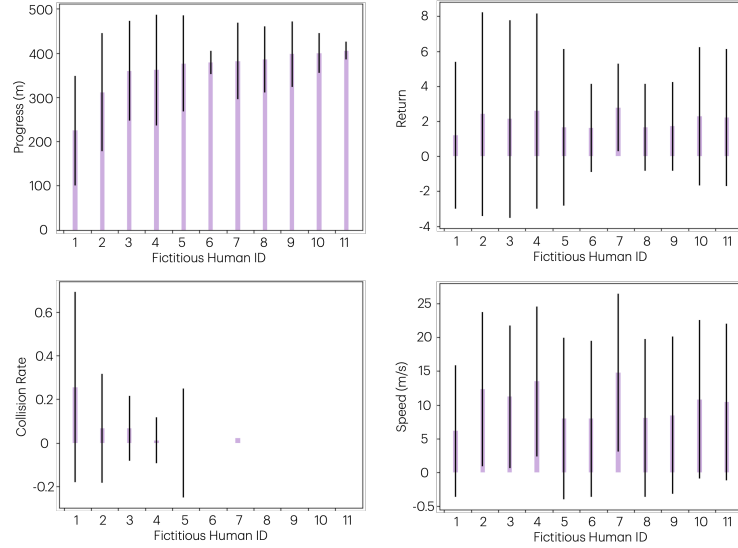


Figure D.4: Absolute metrics in the hairpin scenario evaluated for the *unassisted* imperfect (1–5) and near-perfect (6–11) humans tending to *stay*, averaged over four random seeds. 1- σ error bars are shown.

87 E Additional Experimental Results

88 E.1 Performance across Different Humans

89 We provide a more complete comparison of the results, showing additional metrics in an evaluation
 90 that extends Fig. 3 from imperfect to near-perfect humans (1–11, ordered according to unassisted
 91 track progress performance) for **DREAMER**, **DREAMER-AIL**, **BETAIL**, and **DREAM2ASSIST**.
 92 We show these results in Figs. D.1–D.4. Note that the additional fictitious humans (6–11) achieve
 93 nearly-identical baseline performance across all metrics, as indicated by Figs. D.2 and D.4. Hence,
 94 the changes across humans 6–11 in Figs. D.1 and D.3 are likewise similar.

95 From Fig. D.1, we observe that DREAM2ASSIST, when applied to the imperfect *pass* humans (1–5)
 96 generally yield improvements in progress, collisions, and speed, with a slight overall decrease in
 97 reward, and an overall moderate intervention level compared to the baselines. For humans 6–10,

Table E.1: Improvement over unassisted humans for the *pass-stay* humans on straightaway and hairpin experiments, with statistics aggregated across four random seeds. **Blue** indicates improvement over unassisted humans, **bold** is best.

	Pass (top) / Stay (bottom)									
	Progress (m) ↑	Return ↑	Hairpin Collisions ↓	Interventions ↓	Speed (m/s) ↑	Progress (m) ↑	Return ↑	Straightaway Collisions ↓	Interventions ↓	Speed (m/s) ↑
DREAMER	-11.3 ± 13.6	0.5 ± 2.9	-0.1 ± 0.2	0.2 ± 0.2	0.1 ± 1.1	-0.7 ± 4.6	-0.1 ± 0.9	0.0 ± 0.1	0.1 ± 0.0	-0.4 ± 1.9
DREAMER-AIL	-21.1 ± 68.6	0.3 ± 0.4	0.0 ± 0.0	0.1 ± 0.0	2.9 ± 1.6	-6.1 ± 17.5	0.0 ± 0.1	0.1 ± 0.2	0.1 ± 0.0	0.3 ± 0.5
DREAM2ASSIST	-28.9 ± 46.2	-5.7 ± 7.3	-0.3 ± 0.2	1.0 ± 0.2	1.4 ± 4.4	-116.6 ± 58.7	-9.7 ± 4.2	-0.4 ± 0.3	1.4 ± 0.0	-21.2 ± 10.6
DREAM2ASSIST-a	-216.6 ± 145.6	-1.7 ± 0.4	0.1 ± 0.2	0.9 ± 0.1	5.4 ± 2.4	-52.5 ± 45.9	-1.6 ± 0.2	-0.1 ± 0.1	1.4 ± 0.0	-1.2 ± 1.1
DREAM2ASSIST+a	71.8 ± 43.9	-1.2 ± 3.3	-0.2 ± 0.2	0.7 ± 0.1	0.5 ± 1.4	6.0 ± 9.1	-0.4 ± 1.4	-0.1 ± 0.1	0.3 ± 0.0	-1.2 ± 1.9
DREAM2ASSIST+a-r	60.5 ± 49.7	0.8 ± 0.4	0.1 ± 0.1	0.4 ± 0.0	7.1 ± 2.4	57.8 ± 36.4	-0.1 ± 0.2	0.1 ± 0.1	0.4 ± 0.1	1.7 ± 1.0
DREAM2ASSIST+a-r	25.2 ± 11.0	-2.1 ± 7.4	-0.3 ± 0.2	0.7 ± 0.2	0.8 ± 4.9	6.9 ± 9.9	-5.6 ± 2.4	-0.1 ± 0.2	0.8 ± 0.2	-14.8 ± 7.7
DREAM2ASSIST+a-r	24.3 ± 66.6	0.2 ± 0.6	0.0 ± 0.4	0.7 ± 0.0	5.1 ± 3.0	9.2 ± 7.8	-6.6 ± 0.7	0.1 ± 0.1	0.9 ± 0.1	0.4 ± 0.2
DREAM2ASSIST+a-r	-79.0 ± 53.2	-13.5 ± 3.9	-0.5 ± 0.6	0.6 ± 0.0	-6.1 ± 5.6	-9.0 ± 20.2	-2.9 ± 1.0	0.0 ± 0.1	0.3 ± 0.3	-0.7 ± 1.2
DREAM2ASSIST+a-r	-281.9 ± 68.7	-2.5 ± 0.5	-0.1 ± 0.1	0.7 ± 0.0	-1.4 ± 2.3	10.4 ± 38.3	-1.0 ± 0.1	0.1 ± 0.3	0.8 ± 0.2	1.8 ± 0.5

progress, reward, collisions, and speed are all negatively impacted, hinting at room for improvement in handling near-perfect humans. Similar trends for imperfect humans (1–5) tending to *stay* can be seen in Fig. D.3, where all metrics except collision see improvement. For near-perfect humans (6–11), the results generally indicate marginal improvement across all metrics, except collision.

We also complement Table 1 with additional metrics, including the magnitude of intervention, and speed, both averaged over time. In Tables E.1 and E.2, we compare **Dreamer**, **DREAM2ASSIST**, and **DREAM2ASSIST-AIL**, and further include results for a variant of DREAM2ASSIST with the action-based reward term in (6) using $\alpha_r = \alpha_a = 1$, which we call **DREAM2ASSIST+a**, as well as a variant of DREAM2ASSIST with the action-based reward term and no reward, i.e. $\alpha_a = 1, \alpha_r = 0$, which we call **DREAM2ASSIST+a-r**. For the pass vs. stay case in Table E.1, DREAM2ASSIST achieves best performance in 5 categories, while improving over unassisted humans in 10 categories. The performance of DREAM2ASSIST+a and DREAM2ASSIST+a-r were mixed. DREAM2ASSIST+a was able to improve over unassisted humans, but with generally lower progress than DREAM2ASSIST, while DREAM2ASSIST+a-r almost completely hindered the human’s progress, due to the fact that the reward term no longer explicitly captures the dense progress sub-reward, and are not implicitly reflected in the actions of the optimal human. In E.2, we see similar trends, with DREAM2ASSIST outperforming baseline approaches in 5 categories, and performing better than unassisted humans in 8 categories. DREAM2ASSIST+a-r is unable to make progress, and DREAM2ASSIST+a also reveals lower progress than the DREAM2ASSIST in the hairpin and straightaway domains.

E.2 Intent Classification Performance

We next probe the performance of the intent classification. F_1 scores achieved on training data yields high performance, as shown in Table E.3.

We provide two example time traces to illustrate stability of inferring the human’s intent by the assistant’s world model in Figs. E.2 and E.3.

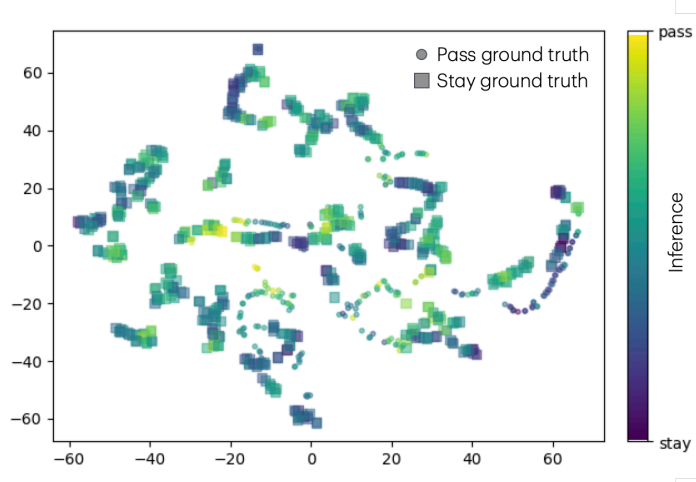
To uncover whether intent inference is due to world model training, we evaluate the t-SNE embeddings of the logits of the assistant’s discrete latent state \hat{z}_t^A . We see that in the Dreamer case, t-SNE is unable to find strong separations between the ground truth intent classes without intent inference in the latents, while in DREAM2ASSIST, there is a stronger separation, and the ground truth intent classes are more strongly clustered in the embedding space, allowing intent to be inferred with much higher accuracy.

Table E.2: Improvement over unassisted humans for the *left-right* humans on straightaway and hairpin experiments, with statistics aggregated across four random seeds. **Blue** indicates improvement over unassisted humans, **bold** is best.

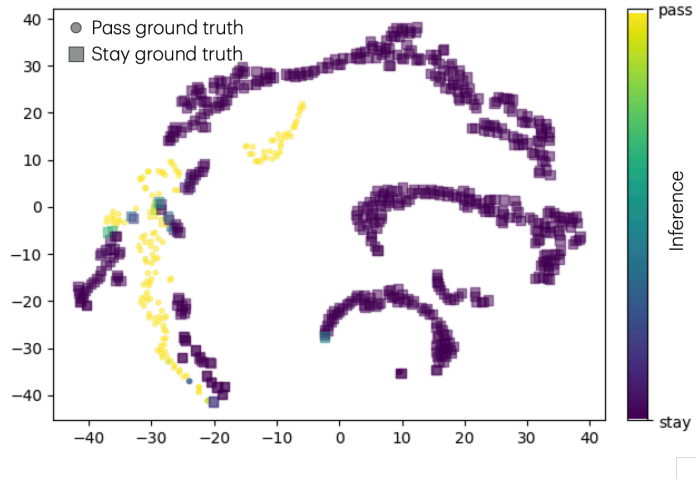
	Left (top) / Right (bottom)									
	Progress (m) ↑	Return ↑	Hairpin Collisions ↓	Interventions ↓	Speed (m/s) ↑	Progress (m) ↑	Return ↑	Straightaway Collisions ↓	Interventions ↓	Speed (m/s) ↑
DREAMER	21.8 ± 28.7	-2.0 ± 2.2	0.1 ± 0.1	0.14 ± 0.1	0.4 ± 1.2	10.8 ± 7.3	-0.6 ± 0.3	0.0 ± 0.1	0.2 ± 0.0	6.0 ± 7.0
DREAMER-AIL	10.0 ± 17.9	-1.3 ± 1.8	0.0 ± 0.1	0.2 ± 0.1	0.1 ± 1.5	-1.1 ± 6.2	0.2 ± 1.2	0.0 ± 0.1	0.1 ± 0.0	0.9 ± 1.0
DREAM2ASSIST	-144.0 ± 89.9	-7.1 ± 6.6	-0.4 ± 0.2	1.2 ± 0.1	-14.4 ± 9.5	-134.4 ± 13.0	-2.5 ± 0.9	-0.5 ± 0.1	1.4 ± 0.0	-11.4 ± 7.8
DREAM2ASSIST-a	-119.1 ± 79.9	-4.5 ± 17.3	-0.4 ± 0.3	1.3 ± 0.0	-15.2 ± 5.6	-126.3 ± 53.2	-13.4 ± 7.0	-0.5 ± 0.2	1.4 ± 0.0	-23.1 ± 5.1
DREAM2ASSIST+a	54.8 ± 60.4	1.4 ± 5.2	0.0 ± 0.1	0.8 ± 0.1	-0.6 ± 6.0	5.0 ± 5.0	0.1 ± 0.5	0.0 ± 0.1	0.6 ± 0.1	-0.1 ± 1.2
DREAM2ASSIST+a-r	27.2 ± 23.1	2.2 ± 2.3	-0.2 ± 0.2	0.8 ± 0.1	3.2 ± 4.7	-1.1 ± 31.9	-2.8 ± 2.4	0.2 ± 0.2	0.8 ± 0.1	-2.1 ± 4.3
DREAM2ASSIST+a-r	9.3 ± 32.0	3.2 ± 4.3	-0.1 ± 0.1	0.9 ± 0.1	2.9 ± 5.8	-18.3 ± 12.9	0.2 ± 1.7	-0.2 ± 0.2	0.9 ± 0.1	-4.9 ± 5.3
DREAM2ASSIST+a-r	-61.2 ± 40.5	-0.2 ± 2.1	-0.2 ± 0.3	0.9 ± 0.1	-1.4 ± 3.1	-8.2 ± 35.6	-5.2 ± 5.4	-0.1 ± 0.2	1.0 ± 0.0	-14.0 ± 6.0
DREAM2ASSIST+a-r	-147.3 ± 93.0	-24.2 ± 13.1	-0.4 ± 0.2	0.5 ± 0.1	-9.6 ± 7.7	6.9 ± 4.1	0.3 ± 1.5	0.0 ± 0.1	0.4 ± 0.1	5.2 ± 3.9
DREAM2ASSIST+a-r	-179.9 ± 104.2	-22.2 ± 16.3	-0.4 ± 0.4	0.6 ± 0.1	-10.7 ± 6.8	-3.2 ± 12.6	-0.0 ± 3.0	0.1 ± 0.2	0.1 ± 0.0	-0.1 ± 1.7

Table E.3: F_1 -Scores over the training set.

Pass vs. Stay Hairpin	Pass vs. Stay Straightaway	Left vs. Right Hairpin	Left vs. Right Straightaway
0.99 ± 0.006	1.00 ± 0.000	0.95 ± 0.00	0.98 ± 0.00



(a) Dreamer



(b) DREAM2ASSIST

Figure E.1: t-SNE embeddings of \hat{z}_t^A for: (a) the non-intent-aware world model of Dreamer versus (b) the intent-supervised world model of DREAM2ASSIST. The consistency of the clusterings present in the DREAM2ASSIST world model states indicates that the world model has learned to identify the human’s intent.

128 E.3 Learning Curves

129 We provide learning curves for the **DREAMER**, **DREAM2ASSIST** and **DREAM2ASSIST+a** assistance
 130 schemes in the two domains across all the human objectives in Fig. E.4. We compare these across
 131 track progress, and observe a general trend of stability in training.

132 E.4 Visuals of Assistance

133 We show, in Fig. E.6, CARLA and bird’s-eye-view snapshots of driving with assistance, and compare
 134 that to a human without assistance in Fig. E.5.

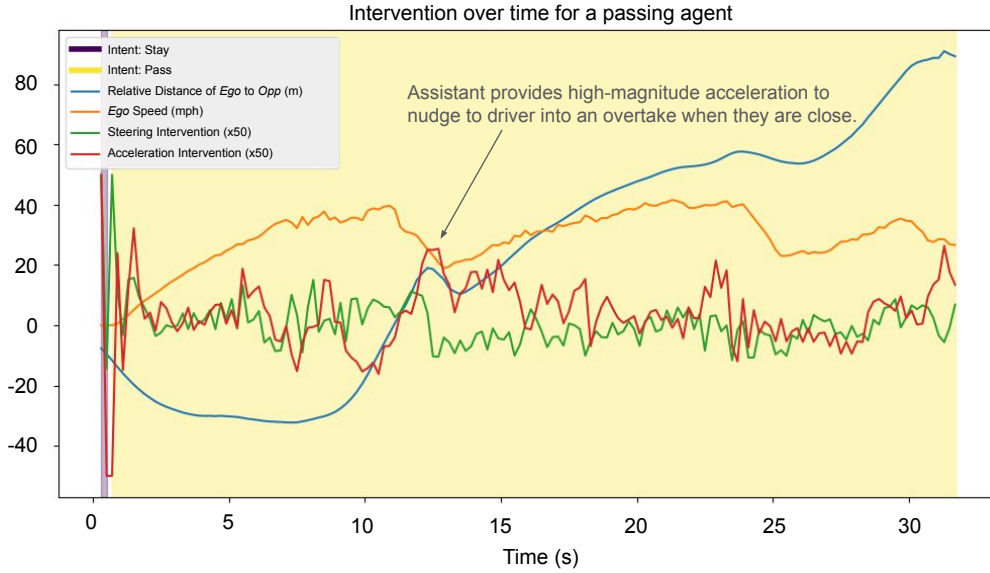


Figure E.2: Time traces in the hairpin scenario showing the intent inference (denoted by background color), along with ego-opponent distance, speed, and DREAM2ASSIST steering and acceleration modifications for a human tending to *pass*. Notice that DREAM2ASSIST maintains an accurate estimate of the driver's intent, and provides a high-magnitude acceleration intervention to assist as the ego begins to overtake the opponent.

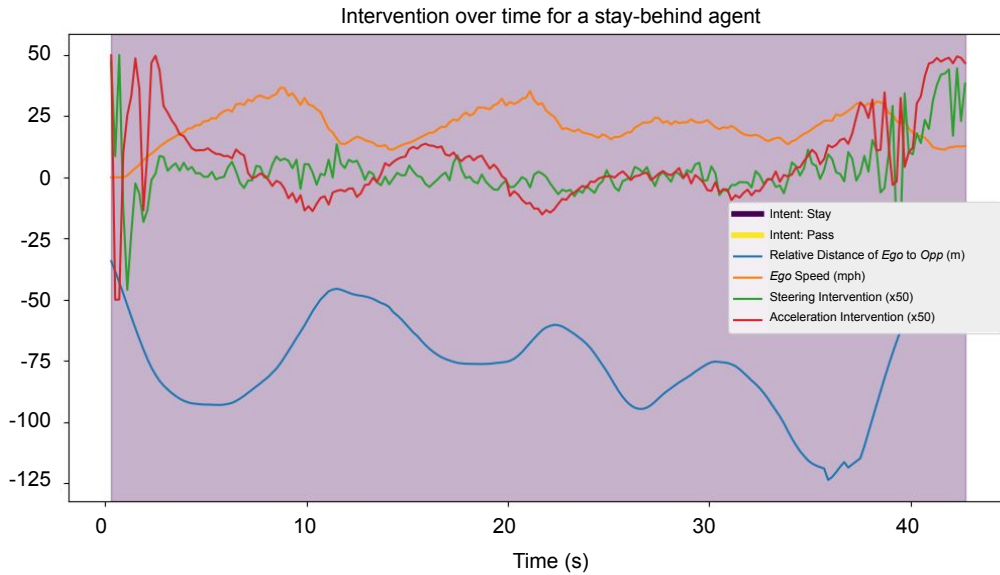


Figure E.3: Time traces in the hairpin scenario showing the intent inference (denoted by background color), along with ego-opponent distance, speed, and DREAM2ASSIST steering and acceleration modifications for a human tending to *stay*.

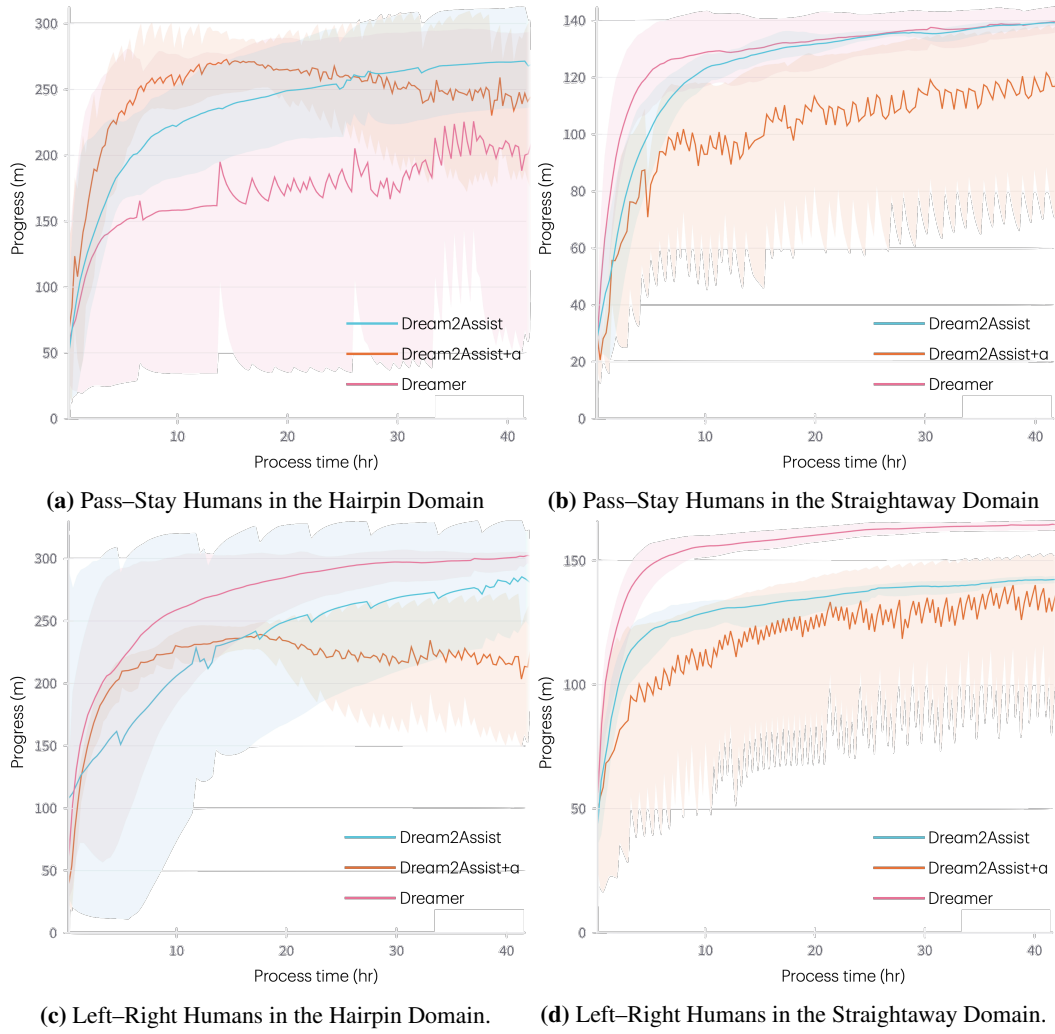


Figure E.4: Learning curves for each domain, averaged over four random seeds.

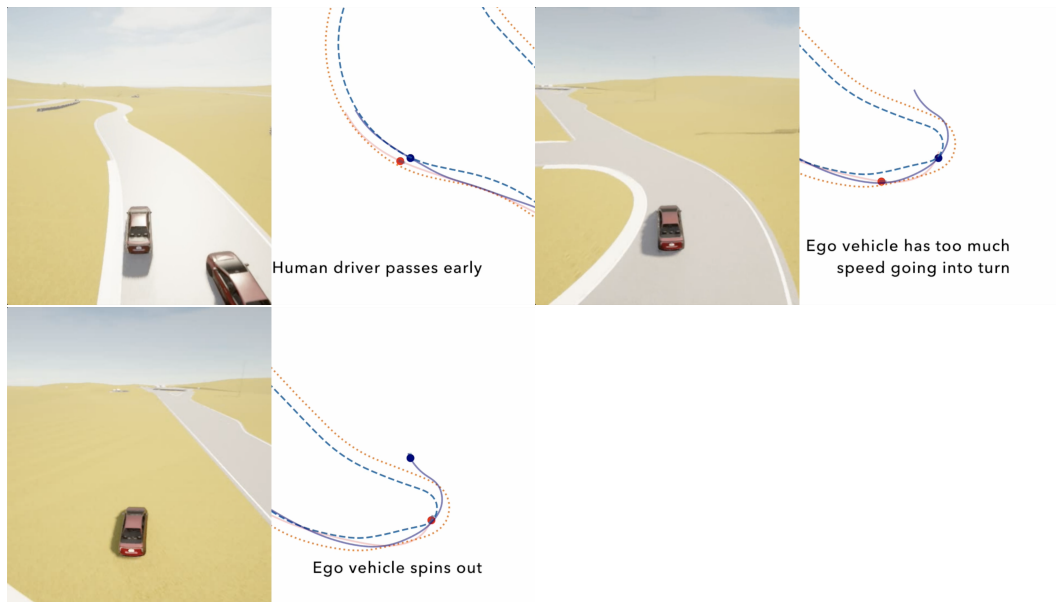


Figure E.5: Example of a time sequence of an imperfect passing human driving in the hairpin domain.

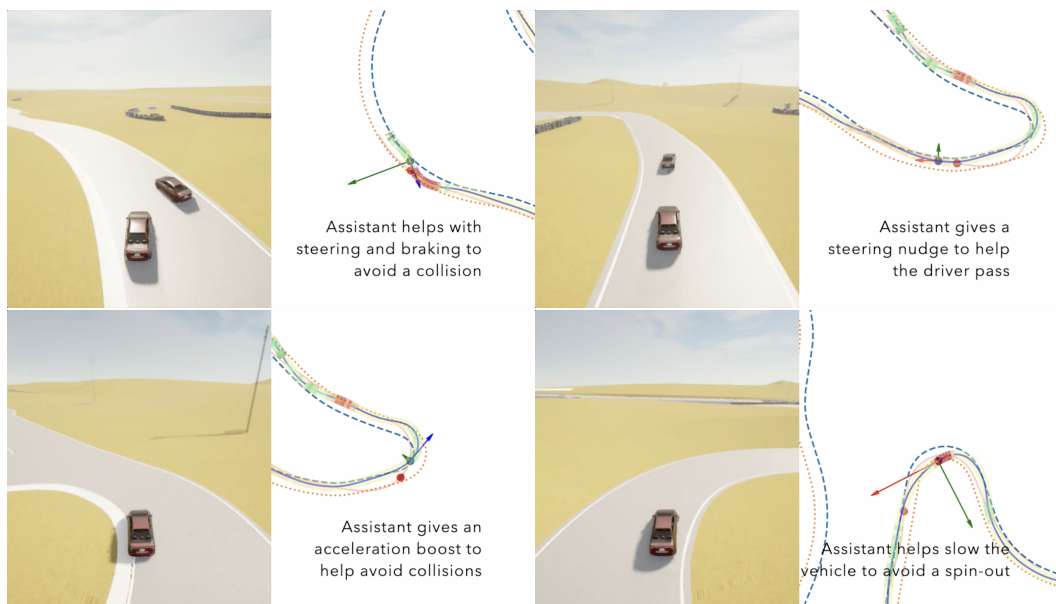


Figure E.6: Example of a time sequence of DREAM2ASSIST assistance to help an imperfect passing human in the hairpin domain.

135 **References**

- 136 [1] 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,
137 P. MacAlpine, D. Oller, T. Seno, C. Sherstan, M. D. Thomure, H. Aghabozorgi, L. Barrett,
138 R. Douglas, D. Whitehead, P. Dürr, P. Stone, M. Spranger, and H. Kitano. Outracing champion
139 Gran Turismo drivers with deep reinforcement learning. *Nature*, 602(7896):223–228, Feb. 2022.
140
- 141 [2] D. Strouse, K. McKee, M. Botvinick, E. Hughes, and R. Everett. Collaborating with humans
142 without human data. *Advances in Neural Information Processing Systems*, 34:14502–14515,
143 2021.