Adaptive Robot Navigation: A Human-Centered and Multi-Objective Approach with Demonstrations

Jorge de Heuvel Tharun Sethuraman Maren Bennewitz

***Abstract*— Preference-aligned robot navigation in human environments is typically achieved through learning-based approaches, utilizing demonstrations and user feedback for personalization. However, personal preferences are subject to change and might even be context-dependent. Yet traditional reinforcement learning (RL) approaches with a static reward function often fall short in adapting to these varying user preferences. This paper introduces a compact framework for mobile robot navigation that combines multi-objective reinforcement learning (MORL) with demonstration-based learning. Our approach allows for dynamic adaptation to changing user preferences and environmental contexts without retraining. Through rigorous evaluations we demonstrate our framework's capability to reflect user preferences accurately while achieving high navigational performance.**

I. Introduction

In recent years, the field of robotics has witnessed remarkable advancements, particularly in the domain of autonomous mobile robot navigation. Central to these advancements is the application of deep reinforcement learning (RL), offering approaches to end-to-end learning of nuanced navigation policies. But whenever robots coexist and navigate in human environments, a central aspect for the best human-robot interaction is the alignment of embodied AI systems to user preferences and values [1], [2].

While single-objective RL frameworks have laid a strong foundation, they fall short in addressing the complexities of social navigation, where multiple, often conflicting objectives must be balanced. For instance, navigating in human-populated environments requires not just navigational efficiency and safety but also adherence to social norms, personal space, and user preferences regarding proxemics, approaching behavior and navigational efficiency. Furthermore, existing research on navigation often treats user preferences as static, and the training process bakes specific reward shapes and behaviors into the system. Ultimately, the pre-configured reward functions of traditional RL policies cannot be adapted to changing preferences or complex social dynamics without retraining, highlighting a significant gap in the current methodology.

As a solution, multi-objective reinforcement learning (MORL) emerges [4], enabling instant post-training adaptation of behavior to accommodate changing preferences without the need for retraining. This flexibility offers on-the-fly adaption of various criteria including efficiency, safety, and social appropriateness but also ensures the maintenance of baseline objectives, including goal pursuance and collision avoidance. While those criteria can be expressed in analytical reward functions, this is inherently difficult for diverse human-anchored navigation styles. Simultaneously, learning-based policies have the potential to outperform traditional navigation approaches on nuanced and foresighted navigation styles, when given information-rich feedback from the user.

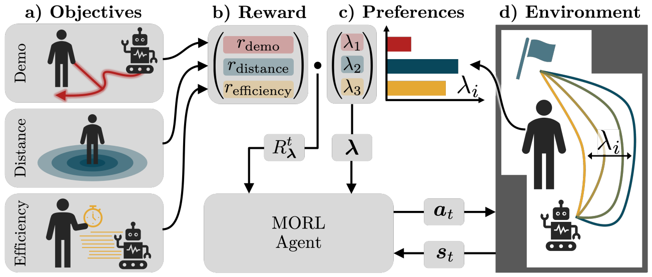

Fig. 1: Our multi-objective reinforcement learning (MORL) navigation agent learns to adapt its behavior to varying preferences without re-training. **a)** The navigation style can fluently shift between demonstration-induced, distance keeping, and efficiency objectives. **b)** A MORL reward vector $\boldsymbol{r}_t$ is modulated **c)** with a varying preference $\boldsymbol{\lambda}$, while providing $\boldsymbol{\lambda}$ as input to the agent. **d)** The resulting human-centered policy balances the baseline objectives with preference-reflection. A context predictor can map navigational context such as room types to certain preferences.

Beside query-based preference alignment, an essential feedback modality are demonstrations. De Heuvel *et al.* [2], [3] have employed an additional behavior cloning loss to shape RL-based navigation behavior around the human and room. Yet, the application of personalized navigation strategies raises critical questions regarding the balance of demonstration-based preferences with respect to baseline objectives and contextual adaptability. It becomes essential to devise mechanisms that can modulate the influence of demonstrations and preferences [5], ensuring flexibility and contextual appropriateness.

This paper introduces a novel framework that unites the flexibility of multi-objective reinforcement learning with the personalization capabilities of demonstration-based learning, tailored to the unique challenges of social robot navigation, see Fig.1. Our study case resembles the navigation in the vicinity of a human and unknown obstacles. With changing preferences of the user, the navigation behavior can be adjusted on-the-fly without retraining. Our reward design achieves distinct preference-adaption while simultaneously meeting essential navigational baseline objectives such as goal pursuance or collision avoidance. In our extensive evaluation, we demonstrate the navigational performance and quality of preference-reflection.

In summary, the main contributions of our work are:

- A MORL social robot navigation framework that enables policy adaption post training.
- The incorporation of demonstration data as a tuneable objective.

- Extensive qualitative and quantitative analyses of the navigation behavior under various preferences.

II. Related Work

In the context of learning-based mobile robot navigation, this section spans multi-objective reinforcement learning, and learning from demonstrations.

A. Multi-Objective Reinforcement Learning in Robotics

Multi-objective reinforcement learning (MORL) extends traditional RL frameworks to optimize several objectives simultaneously, offering a more versatile approach to decision-making in complex scenarios. While MORL frameworks for discrete [6] and continuous action spaces [7], [8] have been presented, the research field around unified benchmarking libraries [9] is just gaining traction. Yet, MORL has been applied to the fields of autonomous driving [10], and robotic tasks [11], [5].

In the scope of mobile robot navigation, the work of Marta *et al.* [5] demonstrates the fluent interpolation between navigational preferences obtained from a human feedback reward model and baseline objectives. Specifically, a social force model is modulated that guides the robot to the goal among surrounding pedestrians. While their reward model is based on pairwise trajectory queries, we modulate the influence of demonstration data on a policy that directly drives the robot. Picking up the idea of tuneable preference reflection, our approach therefore allows behavior diversity beyond the query space by integrating specific demonstration data.

To adjust navigation behavior among humans, Ballou *et al.* [12] have employed a meta reinforcement learning setup. The policy can be finetuned to changes in the reward function efficiently by leveraging the learned meta knowledge, e.g. for goal pursuance or distance keeping. However, the adaptation to shifting objectives like goal pursuance or distance keeping is not instantaneous. In contrast, our MORL policy will adapt to changes in preference weights immediately, without requiring any retraining.

The closest to our study, yet not specifically MORL by definition, is the work by Choi *et al.* [13], who use multi-agent training with parameterized rewards [14] for an adaptable navigation policy. Each agent is initialized with random reward weights at the beginning of each episode. Finally, the authors estimate preferences in a human feedback loop using Bayesian inference. Our study focuses on navigational analysis rather than preference inference, and the MORL agent estimates different objective's Q-values separately.

B. Personalized Navigation from Preferences

The concept of personalized navigation has gained traction, with researchers exploring how robots can adapt their navigation strategies based on individual preferences. Users can express navigational preferences either by ranking pairwise trajectory queries [13], [15], [5] or providing demonstrations [2], [3]. From both feedback modalities, a preference-aligned navigation policy can be distilled. In this work, we lay the groundwork for a demonstration-data infused policy that can be further aligned on-thy-fly without retraining, e.g., via Bayesian preference inference. The reader should note that we do not focus on the preference inference process from users, which has extensively been studied [13], [5], but on the policy's behavior under chaining preferences.

III. Our Approach

A. Problem Statement

We consider a differential-wheeled robot navigating in the vicinity of a single human and unknown obstacles towards a local goal with a learning-based policy. The human has certain preferences about the navigation style of the robot that may change depending on navigational context. Furthermore, these navigation preferences can be expressed in the form of demonstrations, e.g., through a VR interface [2]. The robot should pursue a local goal location in the environment while performing collision avoidance, rendering this work into the scope of point navigation. We assume the position of the agent and human in the environment to be known. The robot observes its environment through a 2D lidar sensor. The sensor and goal information is processed by the navigation policy together with a context-dependent preference vector, based on which the policy generates style-adapted navigation behavior through velocity commands for the robot.

B. Multi-Objective Reinforcement Learning

Multi-objective reinforcement learning (MORL) broadens the scope of traditional RL by integrating multiple, often conflicting, objectives [4]. In MORL, the agent is trained to learn policies that strike a balance among these diverse objectives, as opposed to a singular reward function in classical RL.

The MORL problem is formulated within the framework of a Markov Decision Process (MDP), defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$. Here, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the state transition probability, and γ is the discount factor. A distinctive feature of MORL is the multi-dimensional reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^n$, which outputs a vector of rewards $\boldsymbol{r}_t$ for n different objectives.

A single policy optimally adheres to a given set of preferences, represented by the convex preference weight vector $\boldsymbol{\lambda} \in \mathbb{R}^n$, thus $\sum_i \lambda_i = 1$. This policy $\pi(s, \boldsymbol{\lambda})$ optimizes a scalarized reward function $R_{\boldsymbol{\lambda}}(s, a) = \boldsymbol{\lambda}^\top \boldsymbol{r}(s, a)$, itemizing the different objectives.

We employed the MORL-TD3 implementation of Basaklar *et al.* [7], which introduces several modification with respect to the policy loss and preference-space exploration. Together, they enhance policy convergence in terms of training efficiency, robustness and preference-reflection.

An integral component of this framework is the preference interpolator $I(\boldsymbol{\lambda}) = \boldsymbol{\lambda}_p$, which projects the original preference vectors $\boldsymbol{\lambda}$ into a normalized solution space, thereby aligning preferences with multi-objective value solutions $\boldsymbol{Q}$. First, solutions for the maximum expectable values of $\boldsymbol{Q}(s, a, \boldsymbol{\lambda})$ are obtained for selected key preferences, initializing a radial basis function interpolation with a linear kernel. The interpolator is subject to continuous updates throughout the training process.

Another feature is the angle term $g(\boldsymbol{\lambda}_p, \boldsymbol{Q})$, a component of the loss function designed to minimize the directional angle between the interpolated preference vectors $\boldsymbol{\lambda}_p$ and the multi-objective vector $\boldsymbol{Q}$:

$$g(\boldsymbol{\lambda}_p, \boldsymbol{Q}(s,a,\boldsymbol{\lambda};\theta)) = \cos^{-1}\left(\frac{\boldsymbol{\lambda}_p^T \boldsymbol{Q}(s,a,\boldsymbol{\lambda};\theta)}{\|\boldsymbol{\lambda}_p\|\|Q(s,a,\boldsymbol{\lambda};\theta)\|}\right) \quad (1)$$

This mechanism ensures that the resulting policies align with the specified preferences, regardless of the varying scales of individual objectives' rewards. The actor network is updated by maximizing the term $\boldsymbol{\lambda}\boldsymbol{Q}$, where $\boldsymbol{\lambda}$ is the original convex preference vector and $\boldsymbol{Q}$ is the critic network's output, while simultaneously minimizing the directional angle term.

The loss functions for the critic is defined as:

$$\begin{aligned} L_{\text{critic}}(\theta_i) = \; & \mathbb{E}_{(s,a,r,s',\boldsymbol{\lambda})\sim D}\left[y - Q(s,a,\boldsymbol{\lambda};\theta_i)\right] \\ & + \mathbb{E}_{(s,a,\boldsymbol{\lambda})\sim D}\left[g(\boldsymbol{\lambda}_p, Q(s,a,\boldsymbol{\lambda};\theta_i))\right] \end{aligned}$$

where $\boldsymbol{y} = \boldsymbol{r} + \gamma \arg_{\boldsymbol{Q}} \min_{i=1,2} \boldsymbol{\lambda}\boldsymbol{Q}(s', \tilde{a}, \boldsymbol{\lambda}; \theta_i')$ is the target value obtained from the target networks. The actor loss is

$$\begin{aligned} \nabla_\phi L_{\text{actor}}(\phi) = \; & \mathbb{E}_{(s,a,r,s',\boldsymbol{\lambda})\sim D}\left[\nabla_a \boldsymbol{\lambda}^T Q(\cdot)|_{a=\pi(\cdot)} \nabla_\phi \pi(\cdot)\right] \\ & + \alpha \mathbb{E}_{(s,a,\boldsymbol{\lambda})\sim D}\left[\nabla_a g(\boldsymbol{\lambda}_p, Q(\cdot))|_{a=\pi(\cdot)} \nabla_\phi \pi(\cdot)\right] \end{aligned}$$

with $\boldsymbol{Q}(\cdot) = \boldsymbol{Q}(s,a,\boldsymbol{\lambda};\theta_1)$ and $\pi(s,\boldsymbol{\lambda};\phi)$ and $\alpha = 10$.

To efficiently learn across the entire preference space in MORL-TD3, a hindsight experience replay mechanism retrospectively generates alternative preference vectors for past experiences, effectively augmenting the training data diversity in the preference domain and enabling the policy to adapt to various objectives more effectively.

Moreover, the training process involves running a number of C_p environments in parallel for N time steps, resulting in a collection of $N \times C_p$ transitions. Each environment is tailored to explore a distinct segment of the preference vector space. This parallelization strategy increases data collection efficiency, which is crucial for training robust and effective multi-objective reinforcement learning models.

1) State and Action Space: The state space contains information about the local goal location, the human, and surrounding obstacles observed through the lidar sensor. More specifically, we provide the robot-centric relative 2D goal location $\boldsymbol{p}_g$ and human position $\boldsymbol{p}_h$ in polar coordinates. The 360° lidar scan is limited to a scanning range of $4\ \mathrm{m}$ and is min-pooled from its original resolution of 720 rays down to $N_{\text{lidar}} = 30$ rays. We merge the poses of goal and human with the lidar distance values $\mathcal{L}_t = \{d_i^t | 0 \le i < N_{\text{lidar}}\}$ in the state space vector as $s_t = \left(\boldsymbol{p}_g, \boldsymbol{p}_h, \mathcal{L}_t\right)$.

The robot is controlled with action commands of linear and angular velocity as $a_t = (v, \omega)$ that lie within a range of $v \in [0, 0.5]\ \mathrm{m\,s^{-1}}$ and $\omega \in [-\pi, +\pi]\ \mathrm{rad\,s^{-1}}$. The inference loop runs at $5\ \mathrm{Hz}$.

2) Networks: The networks of actor, critic, behavior cloning policy, and reward model (see below) are fully-connected multi-layer perceptron (MLP) networks that share the same architecture of 4 layers with 256 neurons each.

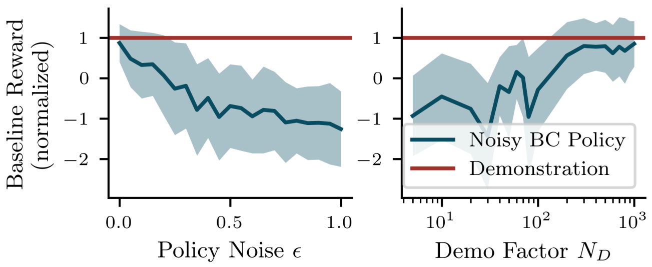

Fig. 2: Exploration of D-REX-related demonstration parameters averages over 20 trajectory rollouts. **a)** The execution of the ϵ-greedy noise-injected behavior cloning (BC) policy trained with a demonstration augmentation factor of $N_D = 1{,}000$ reveals a degradation of goal-approaching performance measured by the normalized baseline reward with growing strength of the injected noise. **b)** The demonstration augmentation factor N_D indicates how many times the human-centric oracle demonstration trajectory (see Sec.III-D.3) was rolled out with randomized obstacle placement. Better generalization comes with higher randomization efforts. Both plots compare against the optimal demonstration behavior's reward.

C. Incorporating Demonstrations

We incorporate demonstration data by distilling demonstration trajectories τ synthetically into a reward model that natively integrates into MORL as one of the objectives and guides the learning agent to demonstration-like behavior. In a typical RL from human feedback setting, such a reward model is derived from pairwise A≻B preference queries in a human feedback process via a ranking loss. Here however, we employ the disturbance-based reward extrapolation (D-REX) approach by Brown *et al.* [16] to generate a reward model from demonstrations. D-REX solves the problem of non-existent ranking for demonstration data, by artificially ranking over noise-injected demonstration trajectories: First, a behavior cloning (BC) policy π_{BC} is trained from N_D demonstration trajectories. Subsequently, the BC policy $\pi_{BC}(\cdot|\epsilon)$ is executed in the environment with increasing level of ϵ-greedy policy noise $\epsilon \in \mathcal{E} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_d)$ with $\epsilon_1 > \epsilon_2 > \ldots > \epsilon_d$. Trajectory rollouts with lower noise are automatically ranked superior compared to their higher-noise counterparts. Finally, a rich preference-ranking dataset

$$D_{\text{rank}} = \{\tau_i \prec \tau_j | \tau_i \sim \pi_{BC}(\cdot|\epsilon_i), \tau_j \sim \pi_{BC}(\cdot|\epsilon_j), \epsilon_i > \epsilon_j\}$$

is obtained. From D_{rank}, we train a reward model $\hat{R}(s,a) \in [0,1]$ using the Bradley-Terry model [17] implemented as a binary cross entropy loss such that $\sum_{s\in\tau_i} \hat{R}_\theta(s,a) < \sum_{s\in\tau_j} \hat{R}_\theta(s,a)$ when $\tau_i \prec \tau_j$.

Based on initial testing, we choose a noise range $\mathcal{E} = (0, \ldots, 0.2)$ for the ranking dataset D_{rank} and obtain $N_D = 1000$ demonstration augmentations from a single human-centric demonstration trajectory pattern, compare Fig. 2.

D. Reward

The scope reward functions covers navigational baseline objectives, and three distinct style objectives based on quantifiable metrics and demonstrations. In our MORL setup, the baseline objectives are summed and occupy the first entry in the reward vector $\boldsymbol{r}_t$ with a static preference weight of one. Note that this is neglected in further notations of $\boldsymbol{\lambda}$ to focus on the tuneable objectives. For the other objectives occupying entries in the reward vector, the preference weights are dynamic.

The reward vector for our MORL framework consists of

$$\boldsymbol{r}_t = (\underbrace{r^t_{\text{baseline}}}_{\text{static}}, \underbrace{r^t_{\text{demo}}, r^t_{\text{distance}}, r^t_{\text{efficiency}}}_{\text{dynamic objectives}}), \tag{2}$$

as explained below.

1) Baseline: The basal navigation behavior required independent of preferences is goal pursuance and collision avoidance. Goal-oriented navigation can be achieved with a continuous potential reward $r^t_{\text{goal}} = c_g \cdot (d^t_g - d^{t-1}_g)$, using the difference in distance of robot and goal $d_g = |\boldsymbol{p}_g|$ between two subsequent time steps and $c_g = 125$. Note that the total non-discounted cumulative goal reward $R = \sum_{t=0}^{T} r^t_{\text{goal}}$ is independent of the number of steps taken to reach the goal. This way, the baseline reward is not biased towards shortest path driving behavior. For collision avoidance, a sparse collision penalty of $r^t_{\text{collision}} = -1000$ upon contact between robot and any obstacle is typically used. Our baseline reward function becomes $r^t_{\text{baseline}} = r^t_{\text{goal}} + r^t_{\text{collision}}$.

2) Tuneable Style Objectives: Our three style objectives cover proxemics, efficiency, and demonstration-reflection: An essential comfort factor in a human-robot navigation scenario are proxemics. We therefore include human distance keeping as one of our tuneable objectives. To do so, we define a quadratic penalty for positional closeness $d_h = |\boldsymbol{p}_h|$ to the human within a range $d_{\text{thresh}} = 2$ m as

$$r_{\text{distance}} = \frac{p_{\max}}{(d_{\text{thresh}} - d_{\min})^2}(d_h - d_{\text{thresh}})^2 \text{ if } d_h \leq d_{\text{thresh}}, \tag{3}$$

else zero, with $p_{\max} = -10$ and $d_{\min} = 0.3$ m. A quadratic term was chosen over a linear one, as the steps required to circumnavigate the human are directly proportional to the radius. Where circumnavigating the human at a larger distance at the cost of more steps should give a smaller cumulative penalty, only a quadratically decaying term works. This was confirmed in preliminary experiments.

The second style objective is navigational efficiency, or shortest path navigation, implemented with a constant time penalty $r^t_{\text{efficiency}} = -10$. This can be in conflict with distance-keeping to path-obstructing obstacles, and usually means passing by closely for minimal detours.

The third and last objective is demonstration-like behavior r^t_{demo}, as elaborated below.

3) Demonstration Acquisition and Reward: By integrating the demonstration-based reward model $r^t_{\text{demo}} = c_d \cdot (\hat{R}_\theta(s_t, a_t) - b_d)$, the amount of demonstration reflection becomes tuneable.

The demonstrations may capture nuanced navigation styles which are difficult to express using an analytical reward function, such as a characteristic trajectories when approaching the user. In this work, we rely on a demonstration oracle that generates an optimal demonstration pattern, see Fig. 3. Specifically, the robot circumnavigates the human in a distinct circular manner. Initially, the robot advances directly towards the human. Upon reaching a proximity of $d_h = 1$m, it executes a 90° left turn and proceeds to orbit the human clockwise at a radius d_h. Once the goal is abeam to the robot, it rotates 90° away from the human and proceeds directly towards the target. While these are not user-demonstrations, the distinct oracle pattern enables a clearer performance analysis, as its behavior is by design contradictory to the other two objectives. It produces longer trajectories in contrast to the time-efficiency penalty $r^t_{\text{efficiency}}$. Also the distance to the human is quite close at $d_h = 1$ m, contradicting the distance penalty r^t_{distance} with an impact radius of 2 m. As the demonstration is anchored solely around the human and goal position, we can easily augment the single demonstration trajectory by rolling it out N_D times in randomized obstacle configurations, while recording only collision-free rollouts. The resulting dataset is handed to the D-REX pipeline, as elaborated in Sec. III-C The final reward term becomes $r^t_{\text{demo}} = c_d \cdot (\hat{R}_\theta(s_t, a_t) - b_d)$, with experimentally found scaling factor $c_d = 10$ and offset $b_d = -1$.

IV. Experimental Evaluation

With our experimental evaluation, we aim to answer the following research questions:

- Q1: Does our framework produce an adaptable yet reliable navigation policy?
- Q2: What is the influence of different navigation preference weights on the navigation behavior of the robot?
- Q3: How well can the demonstrations be reflected by the D-REX-based reward model?

We will do so, by first performing a quantitative analysis that discusses the behavior diversity on selected navigation scenarios. Secondly, a quantitative analysis averages navigation task-relevant metrics over a larger number of trajectories.

A. Training and Environment

To generate our human-centered training environments, we randomly sample the robot start and goal position to have a distance in the range of 6 and 12 m. Subsequently, the human is placed at the center between start and goal. We sample a total of three static obstacles in form of rectangular-shaped boxes around the scene, while avoiding the start, goal, and human positions. On its way to the goal, the robot has to avoid the human, while keeping clear of the obstacles may conflict with the human distance-keeping objective.

In total, we train our agent with $\gamma = 1.0$ on $N \times C_p = 600$k environment steps jointly collected by $C_p = 3$ environments and use the final model for evaluation.

B. Qualitative Navigation Analysis

In the depicted navigation scenarios of Fig. 3, the navigation strategies of our MORL agent are illustrated under varying preference weights and obstacle configurations. Three distinct subplot rows linearly interpolate convex preferences between pairwise combination for two of the three navigation objectives, while keeping the third at zero, respectively. For row 1) interpolating between distance and efficiency, the preference vector parameterized by $\mu \in [0, 1]$ would read $\boldsymbol{\lambda}_1(\mu) = (0, \mu, 1 - \mu)$. The other rows follow analogously by their respective pairwise vector index combinations. The resulting set of $\boldsymbol{\lambda}_i(\mu)$ is $\Lambda_i = \left\{ \left(\frac{i}{N}, 1 - \frac{i}{N}, 0\right) \mid \frac{i}{N} = \mu, i = 0, 1, \ldots, N \right\}$ with $N = 10$, jointly forming the test set $\Lambda = \Lambda_1 \cup \Lambda_3 \cup \Lambda_3$, see Sec. IV-C.

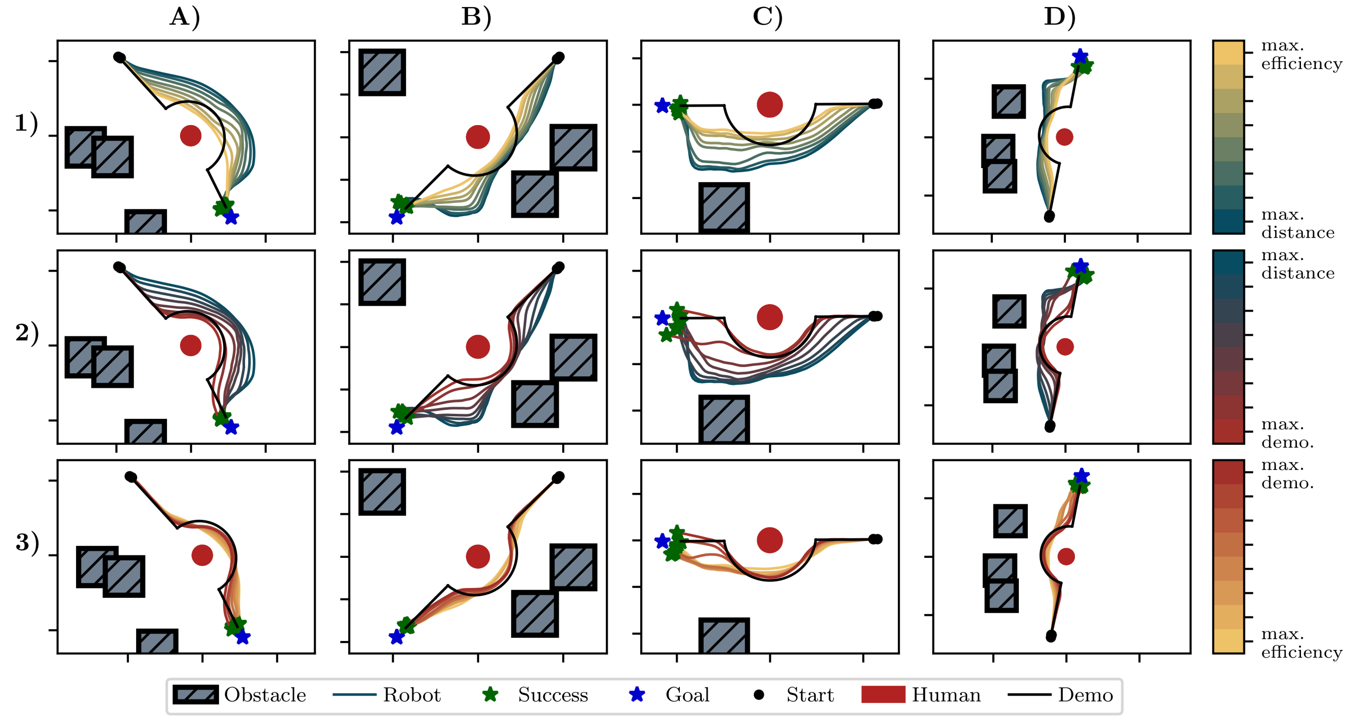

Fig. 3: Trajectory rollouts for different preference vectors (**columns**) for different scene setups (**rows**). As can be seen, the navigation policy shifts its behavior according to the set preference $\boldsymbol{\lambda}$. The colorbars on the right indicate the interpolated preference space Λ_i for each plot row.

The plots show the resulting set of trajectories representing the robot's path from an initial location (black dot) to a designated target (blue star), in the context of static obstacles and the presence of a human (red circle). Also, the original oracle demonstration trajectory is included (black line).

For the shift from efficiency to distance keeping (Fig. 3.1), the trajectory set indicates an increasing distance to the human during traversal. Vice versa, the robot traverses with increasing closeness to the human, until finally it barely passes without collision. A resulting reduction in path length can be observed, as expected from the efficiency penalty $r^t_{\text{efficiency}}$. Note that the agent learned to occasionally stay close to nearby obstacles under the maximum human distance objective, followed by distinct goal-ward turn when the obstacle was passed.

For the shift from distance keeping to demonstration-like behavior (Fig. 3.2), a decrease of minimum distance to the human d_h can be observed. In contrast to the shift into the efficiency objective above, the trajectories now shape into the characteristic demonstration pattern of straight approaching and circular circumnavigation of the human, followed by a bending in in the line between human an goal. While the sharp corners of the in front and behind the human are not as pronounced as in the demonstration trajectory, the navigation behavior clearly approaches demonstration-like patterns, answering Q3.

The overall picture completes, as we shift preferences from demonstration back into efficiency, see Fig. 3.3. While in column A) the demonstration driven trajectories do a nuanced bend around the human, the efficiency driven trajectories pilot right towards the goal after barely passing the human.

Whenever obstacles are close to the human, the agent avoids collisions at the cost of reduced distance to the human. Under maximum distance preference, before and after obstacles, the distance to the human is maintained. Notably, all trajectories traverse with the human on the right hand site, as defined by the demo.

To conclude with respect to Q1 and Q2, the results demonstrate the robot's capability to modulate its path planning strategy from human-distant navigation over demonstration-obedient to efficiency-focused direct paths.

C. *Quantitative Analysis*

We performed a quantitative evaluation of the preference-reflecting agent with respect to multiple performance- and navigation-based metrics, see Fig. 4. The agent was evaluated for 100 episodes in random environment setups for different preference weights $\boldsymbol{\lambda} \in \Lambda$. The again interpolated convex preference vectors as described in Sec. IV-B are indicated as colored fractions in Fig. 4e. Statistical t-test significance for the difference in means between the maximum preferences is indicated. Regarding the navigational performance with respect to Q1, we observe success rate of $100\ \%$, without timeouts or collisions, compare Table I, first entry. As the preference for distance-keeping to the human increases, both the measured minimum human distance and navigation time increase, see Fig. 4a+d. In other words, the robot successfully trades both longer demonstration-like and efficiency-driven trajectories for keeping a higher distance as desired. To measure how well the original demonstration trajectory is reflected with respect to Q3, we compute the Fréchet distance [20] to the executed trajectory, see Fig. 4b. The Fréchet distance finds its minimum at $0.41\ \mathrm{m}$ as the demonstration preference is maximized. Compared to the maximum distance keeping objective, also the efficiency objective reduces the demonstration Fréchet distance, which is expected with

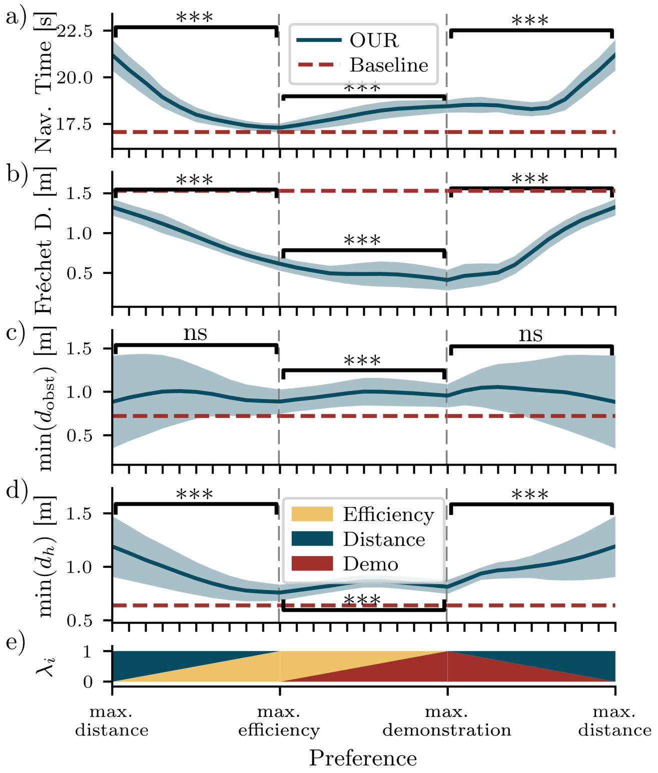

Fig. 4: Quantitative metrics of OUR agent for different preference configurations **(e)**, tested for statistical significance for dissimilar mean between the maximum preferences, with *** for $p \leq 0.001$, and ns for no significance. **a)** The navigation time is smallest for maximized efficiency preference, as expected. **b)** The Fréchet distance to the demonstration trajectory decreases as the demonstration preference increases. **c)** The minimum distance to any obstacle is taken directly from the lidar sensor. **d)** The minimum distance to the human grows with its preference weight. A non-MORL baseline agent (red dotted line) that only obeys the baseline reward term is included in each plot.

the demonstration trajectory closely passing by the human. Comparing the diverging trends of the minimum lidar-recorded distance to any obstacle $\min(d_{\text{obst}})$ in Fig. 4c against the minimum human distance $\min(d_h)$ Fig. 4d, the policy has clearly learned to distinguish between the human and other static obstacles for distance-keeping. While the minimum human distance increases together with its preference weight, the minimum distance to surrounding obstacles does not. The quantitative analysis complements the picture obtained from the qualitative, finding measurable evidence for Q1 to Q3.

1) Baseline: A non-multi-objective policy that only obeys the navigational baseline rewards r_{baseline} of goal pursuance and collision avoidance was trained. The quantitative results of the baseline agent are included in Fig. 4 as a red dotted line. Here, two metrics stand out: While the MORL agent trades human for obstacle distance, the baseline agent lacking the human-distance reward treats both the human and static obstacles similarly. For this reason, a similar minimum value for $d_h = 0.64$ m and $d_{\text{obst}} = 0.72$ m is measured for the baseline agent, in clear contrast to the MORL agent. Also, a higher demonstration Fréchet distance confirms the lack of demonstration knowledge. Is is also notable that the baseline agent has a lower success rate of 97 %.

	$\boldsymbol{\lambda}$	OUR	-NH	-RM	-RM-NH	-γ
SR↑ [%]	$\boldsymbol{\lambda}_i \in \Lambda$	**100**	96.8	100	79.6	96.0
CR↓ [%]	$\boldsymbol{\lambda}_i \in \Lambda$	**0**	2.7	0	11.4	4.0
TR↓ [%]	$\boldsymbol{\lambda}_i \in \Lambda$	**0**	0.5	0	9.0	0.0
$\min(d_h)$↑ [m]	$\boldsymbol{\lambda}_{\text{dist}}$	1.18	0.52	1.16	0.48	**1.22**
Fréchet↓ [m]	$\boldsymbol{\lambda}_{\text{demo}}$	**0.41**	0.57	0.46	0.49	0.50
Nav. time↓ [s]	$\boldsymbol{\lambda}_{\text{eff}}$	17.3	**16.9**	17.4	19.2	18.3

TABLE I: Ablation study with respect to the state space and reward model. For the ablation identifiers and preference vectors $\{\boldsymbol{\lambda}_{\text{dist}}, \boldsymbol{\lambda}_{\text{demo}}, \boldsymbol{\lambda}_{\text{eff}}\}$, please refer to Sec. IV-C.2. For brevity, the identifiers are shortent after OUR, so that, e.g., -NH corresponds to OUR-NH. The results were averaged over 100 trajectories for single $\boldsymbol{\lambda}$, and for the success rate (SR), collision rate (CR), and timeout rate (TR) additionally over all $\boldsymbol{\lambda}_i \in \Lambda$.

2) Ablation Study: We ablated the architecture with respect to the state space and demonstration reward model, compare Table I. The state space changes apply to all involved models: D-REX BC policy, D-REX reward model, actor and critic. Firstly, we exclude the human position from the state space in OUR-NH. Secondly, the reward model reasons only about the state without the corresponding action as $r^t_{\text{demo}} = \hat{R}_\theta(s_t)$ in OUR-RM. Thirdly, the above two ablations are combined in OUR-RM-NH. Fourthly, we lower the discount value to $\gamma = 0.99$ in OUR-γ. Note that the the maximum preference vectors in Table I are $\boldsymbol{\lambda}_{\text{demo}} = (1, 0, 0)$, $\boldsymbol{\lambda}_{\text{dist}} = (0, 1, 0)$, $\boldsymbol{\lambda}_{\text{eff}} = (0, 0, 1)$, respectively.

Compared to OUR, we observe a decrease in the distance-reflection capabilities upon removal of the human position from the state space in OUR-NH and OUR-RM-NH. Given the intrinsic correlation between demonstration and distance preferences relative to the human position, this result appears to be a natural consequence. It furthermore underlines that the MLP networks cannot differentiate the human from obstacles in the low-resolution lidar. With similar collision-free performance, the state-only reward model in OUR-RM is a contender to our flagship approach, but at slightly weaker preference-reflection. A slightly better human distance-keeping ability is revealed by OUR-γ, but at the cost of more collisions and weaker demonstration and efficiency adaption. Note that we have also experimented with lower values of discount value γ, e.g., at $\gamma = 0.95$ without convergence of training. We attribute this result with the agent's willingness to defer goal rewards into the future under higher γ for better compliance with preference objectives.

V. Conclusion

In conclusion, our paper introduces an innovative framework combining multi-objective reinforcement learning (MORL) with demonstration-based learning for adaptive, personalized robot navigation in human environments. Our approach successfully modulates the conflicting objectives of distance keeping, navigational efficiency and demonstration reflection without retraining, demonstrating excellent navigational performance.

While the amount of demonstration-reflection can be modulated, a limitation of our approach is the inability to alter the demonstration data itself without retraining. We believe this poses an interesting avenue for future research.

We believe that designing robot policies in a multi-objective, therefore tuneable manner is a step forward in developing robots capable of seamlessly integrating into human-centric spaces.

References

[1] P. Vamplew, R. Dazeley, C. Foale, S. Firmin, and J. Mummery, "Human-aligned artificial intelligence is a multiobjective problem," *Ethics and Information Technology*, vol. 20, no. 1, Mar. 2018.

[2] J. de Heuvel, N. Corral, L. Bruckschen, and M. Bennewitz, "Learning Personalized Human-Aware Robot Navigation Using Virtual Reality Demonstrations from a User Study," in *2022 31th IEEE International Conference on Robot Human Interactive Communication (RO-MAN)*, 2022.

[3] J. de Heuvel, N. Corral, B. Kreis, J. Conradi, A. Driemel, and M. Bennewitz, "Learning Depth Vision-Based Personalized Robot Navigation From Dynamic Demonstrations in Virtual Reality," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2023.

[4] C. F. Hayes, R. R adulescu, E. Bargiacchi, J. Källström, M. Macfarlane, M. Reymond, T. Verstraeten, L. M. Zintgraf, R. Dazeley, F. Heintz, E. Howley, A. A. Irissappane, P. Mannion, A. Nowé, G. Ramos, M. Restelli, P. Vamplew, and D. M. Roijers, "A practical guide to multi-objective reinforcement learning and planning," *Autonomous Agents and Multi-Agent Systems*, vol. 36, no. 1, Apr. 2022.

[5] D. Marta, S. Holk, C. Pek, J. Tumova, and I. Leite, "Aligning Human Preferences with Baseline Objectives in Reinforcement Learning," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, May 2023.

[6] R. Yang, X. Sun, and K. Narasimhan, "A Generalized Algorithm for Multi-Objective Reinforcement Learning and Policy Adaptation," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[7] T. Basaklar, S. Gumussoy, and U. Y. Ogras, "PD-MORL: Preference-Driven Multi-Objective Reinforcement Learning Algorithm," 2023.

[8] J. Xu, Y. Tian, P. Ma, D. Rus, S. Sueda, and W. Matusik, "Prediction-Guided Multi-Objective Reinforcement Learning for Continuous Robot Control," in *Proceedings of the 37th International Conference on Machine Learning*. PMLR, Nov. 2020.

[9] F. Felten, L. N. Alegre, A. Nowe, A. Bazzan, E. G. Talbi, G. Danoy, and B. C. da Silva, "A Toolkit for Reliable Benchmarking and Research in Multi-Objective Reinforcement Learning," *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[10] X. He and C. Lv, "Toward personalized decision making for autonomous vehicles: A constrained multi-objective reinforcement learning technique," *Transportation Research Part C: Emerging Technologies*, vol. 156, Nov. 2023.

[11] S. Huang, A. Abdolmaleki, G. Vezzani, P. Brakel, D. J. Mankowitz, M. Neunert, S. Bohez, Y. Tassa, N. Heess, M. Riedmiller, and R. Hadsell, "A Constrained Multi-Objective Reinforcement Learning Framework," in *Proceedings of the 5th Conference on Robot Learning*. PMLR, Jan. 2022.

[12] A. Ballou, X. Alameda-Pineda, and C. Reinke, "Variational meta reinforcement learning for social robotics," *Applied Intelligence*, vol. 53, no. 22, Nov. 2023.

[13] J. Choi, C. Dance, J.-e. Kim, K.-s. Park, J. Han, J. Seo, and M. Kim, "Fast Adaptation of Deep Reinforcement Learning-Based Navigation Skills to Human Preference," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2020.

[14] K. Lee, S. Kim, and J. Choi, "Adaptive and Explainable Deployment of Navigation Skills via Hierarchical Deep Reinforcement Learning," May 2023.

[15] L. Keselman, K. Shih, M. Hebert, and A. Steinfeld, "Optimizing Algorithms from Pairwise User Preferences," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2023.

[16] D. S. Brown, W. Goo, and S. Niekum, "Better-than-Demonstrator Imitation Learning via Automatically-Ranked Demonstrations," in *Proceedings of the Conference on Robot Learning*. PMLR, May 2020.

[17] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. The method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, 1952.

[18] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," Feb. 2018.

[19] C. Li, F. Xia, R. Martín-Martín, M. Lingelbach, S. Srivastava, B. Shen, K. Vainio, C. Gokmen, G. Dharan, T. Jain, A. Kurenkov, C. K. Liu, H. Gweon, J. Wu, L. Fei-Fei, and S. Savarese, "iGibson 2.0: Object-Centric Simulation for Robot Learning of Everyday Household Tasks," *arXiv:2108.03272 [cs]*, Nov. 2021.

[20] H. Alt and M. Godau, "Computing the fréchet distance between two polygonal curves," *International Journal of Computational Geometry & Applications*, vol. 05, no. 01n02, Mar. 1995.

[21] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "ROS: An open-source Robot Operating System," in *ICRA Workshop on Open Source Software*, vol. 3. Kobe, Japan, 2009.

[22] D. Fox, W. Burgard, F. Dellaert, and S. Thrun, "Monte carlo localization: Efficient position estimation for mobile robots," *Aaai/iaai*, vol. 1999, no. 343-349, 1999.

[23] G. Grisetti, C. Stachniss, and W. Burgard, "Improving Grid-based SLAM with Rao-Blackwellized Particle Filters by Adaptive Proposals and Selective Resampling," in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*. IEEE, 2005.