2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

# MOSAIC: Multi-objective optimization from zero-shot language reasoning in preference-based RL

Daniel Marta*,1, Simon Holk*,1, and Iolanda Leite[1]

*Abstract*— **Preference-based Reinforcement Learning (RL) enables humans to shape complex goals via preference comparisons between sequences of state-action pairs. Most of the existing approaches focus on a singular objective, overlooking the complex causal reasoning that underpins preferences. However, many real-world challenges are multi-dimensional, and individuals can have different reasons behind their preferences. In this work, we rethink preference-based RL from a multi-objective perspective by distilling human preferences into multiple components. We leverage the zero-shot capabilities of large language models (LLMs) to infer preferences and better align various objectives from text prompts. This allows us to train an ensemble of reward functions, each optimizing for a specific objective. We demonstrate that our approach can address a variety of multi-objective control tasks, improving on approaches that consider a single preference per objective. We show the effectiveness of our approach in better shaping reward functions by utilizing real human preferences and prompts. Our code for the benchmarks, along with additional supplementary details, is available at https://sites.google.com/view/multi-pref/.**

## I. Introduction

Preference-based RL has been recognized as a compelling and intuitive method for training RL agents through pairwise comparisons of sequences of state-action pairs [1], [2], [3]. Looking forward, there are several fundamental challenges to address, stemming from the distinct nature of human preferences and individuality [4]. The diversity in human preferences, expertise, and abilities [5] makes it difficult for a single reward function to represent varied human perspectives [6], [7], [4]. This underpins the inherent multi-objective character of human-aligned robots [8]. In this work, to learn multiple objectives from humans, we adopt a preference-based RL approach [1], [2], [9], [10] as it imposes minimal modality constraints for the reward function, derived solely from human preferences–pairs of state-action sequences and a preference signal for one of the pairs–as it incorporates the essential element of structural alignment, vital for designing intricate objectives [11], [12], [13].

We contend that human preferences are analogous to the varied pieces of a MOSAIC, potentially comprised of different objectives. When we discard the reasoning behind preferences, we may foster, among other outcomes, *causal confusion* [14], [15]—a misalignment in understanding the causal relationships among states, actions, and rewards. This is akin to appreciating a MOSAIC for its overall aesthetic without considering the distinctiveness of each piece. Similarly, *objective collapse* can manifest, comparable to masonry work that repetitively employs one type of stone due to predominant preferences, thereby excluding the unique beauty of other materials. To address these challenges, we explore a more natural way for humans to interact with robots, such as language [16], [17], [18], in an effort to uncover causality in the lens of a multi-objective approach, i.e. optimizing for a compromise of diverse objectives, alongside the easiness of providing preferences [1], [10] serves as an adequate platform to achieve human-aligned robots. To process human prompts and uncover diverse objectives, we leverage recent breakthroughs in large pretrained foundational models, such as BERT [19], CLIP combined with GPT-2 [20], and GPT-3 [21]. These models demonstrate proficiency across a spectrum of tasks–ranging from text completion and sentiment analysis to task descriptions and robot planning–and have established their efficacy in diverse applications [22]. We leverage the zero-shot capabilities of large language models to extract a set of different–and often contradictory–objectives, building policies on a *MOSAIC* through ranking and scalarization. This approach is referred to as **MOSAIC**: **M**ulti-**O**bjective optimization from zero-**S**hot l**A**nguage reason**I**ng in preference-based reinfor**C**ement learning. Figure 1 provides an illustration of the key pillars of MOSAIC in relation to some of the aforementioned analogies. We highlight the main contributions of this paper:

- We present MOSAIC, a stepping stone in a relatively unexplored frontier [23] of acquiring complex policies from non-linear reward functions through language.
- A query sampling strategy for multi-objective preference-based RL, to select highly informative queries through weighted ensemble variance of the different objectives.
- A regularization technique to maximize query information from language in the multi-objective RL.

## II. Related Work

**Learning from human preferences.** Leveraging human preferences for learning has received substantial focus in recent literature [24], [25], [2], showcasing potential as an effective RL method applicable even in high-dimensional robotic settings [10], however, the necessity of extensive human feedback limits its applicability in real-world robotics and other complex scenarios [26], [27], [28] since most of the complexity is derived from the simple nature of preferences by *brute-force*. The reliance on human preferences carries several challenges, as underscored by [15], who, through an empirical study, demonstrates how the infusion of spurious features and an increase in model capacity can inadvertently breed causal confusion pertaining to the true reward function,

*Shared first-authorship.

[1]KTH Royal Institute of Technology, Sweden, {dlmarta, sholk, iolanda}@kth.se

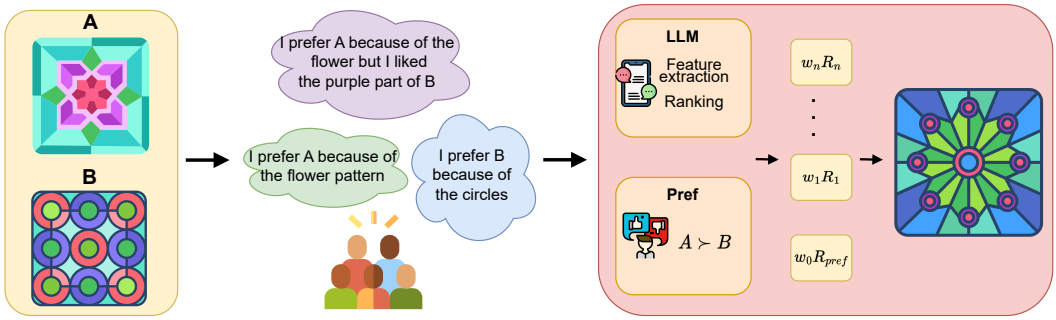


Fig. 1: Macroscopic view of MOSAIC: Humans evaluate two mosaics, figuratively representing state-action sequences A and B, with preferences influenced by various feature-dependent reasons. Analyzing each human prompt for sentiment and distinct objective features, the collective prompts help formulate and rank objectives, which are used to align a policy.

even when utilizing thousands of pairwise preferences. This oversight may pave the way for the emergence of spurious correlations, potentially culminating in reward exploitation [29], [30] or instigating a distributional shift [14].

**Natural Language.** Evidence suggests that suitably expansive language models are capable of executing complex reasoning [31], [22] such as emotional response generation [32]. As a proof of concept, [33] demonstrated that the generation of a thought chain–multiple reasoning steps in a sequential manner–enables the emergence of advanced reasoning abilities in sufficiently expansive language models. Akin to Socratic models [22], it is a common practice to retain specific subcomponents of models—particularly those related to one modality but not others—in a frozen state for downstream tasks [34], [35], [36], [37], [38]. In MOSAIC we aim to integrate an LLM to elaborate on the prompts provided by humans in a variant of zero-shot transfer learning [39], [40], [41], [42].

**Multi-Objective RL (MORL).** Combining multiple objectives in RL enjoys an ever-growing body of literature [43], [44], [45], [23]. In recent years, several methods have been designed to improve efficiency [46] and account for the broad nature of multiple objectives [47], continuous in continuous robot control tasks [48]. Stemming from the interleaved nature of reward and policy training in preference-based RL [1], a multi-objective approach will naturally result in a *dynamic weights setting* [49], where the weights of the different objectives change over time. Several dynamic weighting approaches propose to train a single network [50], [51], [52] to cover the entire preference space as opposed to fix a preference vector during train, but often rely on linearity constraints to cover the entire Pareto front and may not outperform static alternatives [53]. Objective weights may also be implicitly learned through through policy comparisons [54]. In MOSAIC we explore language-based reasoning to uncover both the objectives and their relative importance.

## III. Background

In this section, the fundamentals of MOSAIC are introduced. We explore a scenario wherein a robot, at state $s_t$, initiates an action $a_t$ based on a policy $\pi_\omega(a_t, s_t)$, with parameters $\omega$. After performing the action, the robot earns a vectorized scalar reward $\boldsymbol{r} : s \times \mathcal{A} \rightarrow [r_1, \ldots, r_k], k \in \mathbb{N}_+$ , where $k$ represents the number of different objectives. After the action is taken, we move to the next state $s_{t+1}$ within a multi-objective Markov Decision Process (MOMDP) framework. A policy $\pi : s \rightarrow \mathcal{A}$ within the MOMDP is associated with a specific vector of expected returns $\boldsymbol{\mathcal{J}^\pi} = [\mathcal{J}^\pi_{r_1}, \ldots, \mathcal{J}^\pi_{r_k}]$, linked to the vectorized reward function $\boldsymbol{r}$. It follows that $\mathcal{J}^\pi_{r_k} = \mathbb{E}\left[\sum_{t=0}^{T} \gamma_k^t r_k(s_t, a_t) \middle| s_0 \sim \mathcal{S}_0, a_t \sim \pi(s_t)\right]$ where $\mathcal{J}^\pi_{r_k}$ signifies the return of $r_k$, $T$ denotes the trajectory's horizon, and $\gamma_k$ a discount factor.

**Preference Learning.** We adapt the task of inferring a reward function, $\hat{r}_\psi$, parameterized by $\psi$, from preferences as a supervised learning challenge, following the methodology by [1]. The central goal of preference-based RL, explored thoroughly in [24], is to infer rewards from state-action pair sequences. By defining trajectory segments as sequences of state-action pairs [55], represented as $\sigma^j = ((s_t^j, a_t^j), \ldots, (s_{t+m-1}^j, a_{t+m-1}^j))$, where $j$ denotes the segment index, covering state-action pairs from $t$ to $t+m$ and $m$ indicates the segment length, humans are provided pairs of these segments, $(\sigma^0, \sigma^1)$, and are asked to allocate a preference $\mu \in \{0, 0.5, 1\}$. Here, $\mu = 0$ implies $(\sigma^0 \succ \sigma^1)$, $\mu = 1$ implies $(\sigma^1 \succ \sigma^0)$, and $\mu = 0.5$ indicates equal preference for both segments. Adhering to the Bradley-Terry model [56], the probability of a human preferring $\sigma^0 \succ \sigma^1$, given it is exponentially dependent on the reward sum over the segments' length, is expressed as $P_\psi[\sigma^0 \succ \sigma^1] = \frac{\exp\left(\sum_t \hat{r}_\psi(s_t^0, a_t^0)\right)}{\exp\left(\sum_t \hat{r}_\psi(s_t^0, a_t^0)\right) + \exp\left(\sum_t \hat{r}_\psi(s_t^1, a_t^1)\right)}$. In this context, $\hat{r}_\psi$ is trained as a binary classifier to predict human preferences on new segments, acting as a proxy for the reward function. The preferences, are stored with the corresponding segments in a labeled dataset $\mathcal{D}_\mu$, consisting of triplets $q = (\sigma^0, \sigma^1, \mu)$. While optimizing $\hat{r}_\psi$, samples are drawn from $\mathcal{D}_\mu$, to minimize the binary cross-entropy loss:

$$\mathcal{L}_{CE}(\hat{r}_\psi, \mathcal{D}_\mu) = -\mathbb{E}_{q \sim \mathcal{D}_\mu}[(1-\mu) \log P_\psi(\sigma^0 \succ \sigma^1) + \mu \log P_\psi(\sigma^1 \succ \sigma^0)]$$

## IV. MOSAIC

In MOSAIC we consider different objectives to be optimized as in a MORL approach considering an MOMDP. These objectives are represented by several reward functions which share the initial preference $\mu$, unless otherwise parsed through text prompts (see Sec. IV-B). We provide humans with the option of providing prompts, each prompt denoted

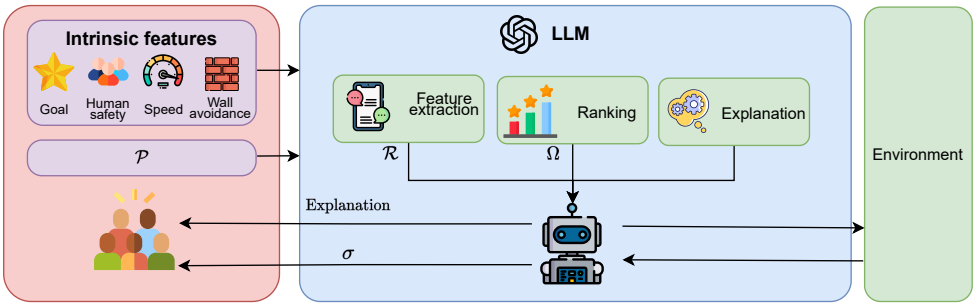


Fig. 2: Overall integration of the LLM with the collective human prompts $\mathcal{P}$ alongside the set of intrinsic objectives $\mathbf{\Omega}$ for a social navigation scenario. Utilizing feature extraction and sentiment analysis, diverse reward functions are trained, subsequently guiding the training of a robot policy.

as $\mathrm{p}_i \in \mathcal{P}$, to augment their preference $\mu$. Here, $i$ indicates the index of the $i$-th prompt, while $\mathcal{P}$ represents the set of all prompts provided by individuals. Consider a set of intrinsic objectives approximated by several reward functions $\mathbf{\Omega} = \{\hat{r}_{\psi_1}, \hat{r}_{\psi_2}, \ldots, \hat{r}_{\psi_n}\}$, where $n \in \mathbb{N}$ represents the different objectives which are environment dependant. In order to obtain the objective weights $\mathbf{\Omega}_w = \{\Omega_{w_1}, \Omega_{w_2}, \ldots, \Omega_{w_n}\}$ we leverage the sentiment analysis capability of the LLM to deduct the relative importance of the objectives according to the full set of prompts $\mathcal{P}$ provided by the users with $\mathrm{LLM}_{\text{weights}} : (\mathcal{P}, \mathbf{\Omega}) \rightarrow \mathbf{\Omega}_w$. To this end, we are interested in policies on the Pareto frontier PF for a given $\mathbf{\Omega}_w$. A Pareto frontier, constitutes all policies $\pi'$ which dominate all other policies $\pi' \succ \pi$ such that $\mathrm{PF} = \{\pi' \mid \forall \pi : \boldsymbol{\mathcal{J}}^{\pi'}_{\hat{\boldsymbol{r}}} \geq \boldsymbol{\mathcal{J}}^{\pi}_{\hat{\boldsymbol{r}}}\}$. We define $\mathrm{PF}^* \subset \mathrm{PF}$ as a subset of policies which also satisfy the human preferred objectives $\mathbf{\Omega}_w$, thus $\mathrm{PF}^* = \{\pi' \in \mathrm{PF} | \exists \mathbf{\Omega}_w \forall \pi (\pi \neq \pi') : \sum_{i=1}^{n} \Omega_{w_i} \mathcal{J}^{\pi'}_{\hat{\boldsymbol{r}}} \geq \sum_{i=1}^{n} \Omega_{w_i} \mathcal{J}^{\pi}_{\hat{\boldsymbol{r}}}\}$. We consider a proxy for a Pareto optimal node to be a parameterized policy $\pi^*_\omega(a_t, s_t)$ that adheres to $\hat{\boldsymbol{r}}_{\mathbf{\Omega}}$, while concurrently considering all objectives, such that $\hat{\boldsymbol{r}}_{\mathbf{\Omega}} = \sum_{i=1}^{n} \Omega_{w_i} \cdot \hat{r}_{\psi_i}(s, a)$.

### A. Query Sampling with MOSAIC

State-of-the-art approaches [1], [57], [10] in preference-based RL, often employ the variance ensemble disagreement as a ranking metric to pick informative queries. To account for multi-objective settings, and given the low sample sizes often available, we propose a modification to ensure higher accuracy in the most relevant objectives. In MOSAIC, we propose weighting the variance of the disagreement by the objective weights. Let $X$ be a random variable representing the sum of predicted rewards for a given segment $\sigma$, as produced by the ensemble. Then, the weighted variance of the ensemble is given by $\mathrm{Var}(X) = \sum_{i=1}^{n} \Omega_{w_i} (\sum_t \hat{r}_{\psi_i}(s_t, a_t) - \bar{\mu})^2$, where $\hat{r}_{\psi_i}(s_t, a_t)$ is the reward prediction produced by the $i$-th estimator for the pair $(s_t, a_t)$, and $\bar{\mu}$ is the mean reward prediction across all estimators in the ensemble, i.e. $\bar{\mu} = \frac{1}{n} \sum_{i=1}^{n} \sum_t \hat{r}_{\psi_i}(s_t, a_t)$.

### B. Prompt-Response Preferences Analysis

Each prompt $\mathrm{p}_i$ is parsed to an LLM alongside a set of objective features $\mathcal{Z} = \{f_1, f_2, \ldots, f_n\}$ that maps to each intrinsic objective $|\mathcal{Z}| = |\mathbf{\Omega}|$, to enhance the initial information for the provided preference $\mu$. We employ an LLM not only to identify intrinsic objective features from human prompts but also to perform sentiment analysis related to those objectives. Thus, we input both the prompt provided by the human and the feature set into the LLM, as expressed by $\mathrm{LLM}_{\text{pref}} : (\mathrm{p}_i \in \mathcal{P}, \mathcal{F}) \rightarrow \mathrm{r}_i \in \mathcal{R}$, where $\mathcal{R}$ corresponds to the set of all possible responses. Each $\mathrm{r}_i$ consists of a set of triplets of the size $n = |\mathcal{F}| = |\mathcal{R}|$, where $\mathrm{r}_i = \{(f_i, \mu_i, \mathrm{y}_i, \mathrm{v}_i), \ldots, (f_n, \mu_n, \mathrm{y}_n, \mathrm{v}_n)\}$, where $f_i \in \mathcal{F}$ is a feature, $\mu_i$ an objective preference, $\mathrm{y}_i \in \{\text{positive}, \text{negative}\}$ is the sentiment associated with feature $f_i$, and $\mathrm{v}_i \in \{\text{low}, \text{high}\}$ is the magnitude associated with the sentiment $\mathrm{y}_1$. In case a feature is not present, it simply inherits the overall preference given by the human. Extending our initial definition of queries, consider *intrinsic objective queries* for each objective $\hat{r}_{\psi_i}$ which are stored on datasets of quintuples $D_{f_i}$ of the form $\{(\sigma^0_i, \sigma^1_i, \mu_i, \mathrm{y}_i, \mathrm{v}_i)\}_{i=1}^{D_{f_i}}$. We process each $\mathrm{r}_i$ into separate datasets for each feature $f_i$. Finally, each of the intrinsic objectives $\hat{r}_{\psi_i}$ is aligned according each dataset $D_{f_i}$. To align to each preference $\mu_i$ we compute the cross entropy of the preferences and regularize the reward objective with the overall sentiment in Sec IV-C. To clarify our formulation we provide a concrete example of a prompt that was evaluated for our experiments (see Sec. V) in the social navigation environment tested with actual human data. For visualisation purposes, the feature set $\mathcal{F}$ is blue, the human prompt $\mathrm{p}_i$ is purple and the output format and response $\mathrm{r}_i$ is green.:

```
Input: You are a robot navigating a corridor with humans
walking around... Which feature(s) was most important
of [end goal, distance to human, speed, distance to
wall] given only the following text given by the user:"B
reached the star but A passed the human more safely and
drove slower."
Respond in the following format for each feature that
is relevant to the text: [feature:insert feature,
alternative: insert 0 or 1, sentiment: insert positive or
negative, value: insert high or low] Alternative explains
which of the alternative of 0/1 the feature is related
to. Sentiment explains if the user thought the robot was
behaving well in regards to the feature, if the robot
behaved well it should be positive, else negative. Value
indicates if the value of the feature was high or low.
Only mention the features that are relevant, disregard
the others.
Output: [feature: distance to human, alternative: 0,
sentiment: positive, value: high]
[feature: end goal, alternative: 1, sentiment: positive,
value: high]
[feature: speed, alternative: 0, sentiment: positive,
value: low]
```

### C. Reward Regularization from Sentiment Analysis

Since humans provide narrower reasoning through each prompt $\mathrm{p}_i$, we propose to highlight sub-sequences within the preferred trajectories akin to temporal cropping [27]. This approach can be considered as incorporating auxiliary tasks, which offer valuable learning signals to improve data efficiency [58], [59], [60]. By enforcing regions of high/low reward, according to sentiment $\mathrm{y}_i$ and intensity $\mathrm{v}_i$, we improve reward query information, shifting from uniformly discounted credit assignment across the entire preferred trajectory. Highlights are constructed as subsequences of segments, represented by $h = \sigma_{i,j} = ((s_i, a_i), \ldots, (s_j, a_j)) \in (s, a)^{j-i}, i, j \in \mathbb{N}_+$ with $0 \leq i \leq j \leq m$. The highlight's length is given by $j-i = L$, wherein $L$ signifies the maximum length for a highlight. To search for highlights within trajectory segments $\sigma_i$, we require several search function $\mathcal{G} = \{g_i\}_{i=1}^{n}$, one for each objective which navigate through segments to detect highlights, depending on the sentiment and intensity, such that $g_i : (\sigma_i^0, \sigma_i^1, \mu_i, \mathrm{y}_i, \mathrm{v}_i) \rightarrow (h^+, h^-)$. We extend the intrinsic objective queries to account for both positive and negative highlights, and store them on a dataset of *highlighted* intrinsic objective queries $D_{h_i} = \{(\sigma_i^0, \sigma_i^1, \mu_i, \mathrm{y}_i, \mathrm{v}_i, h^+, h^-)\}_{i=1}^{D_{h_i}}$. Each dataset $D_{h_i}$ is then used to minimize the regularized loss, and the sum of rewards is discounted along the trajectory of length $L$ by $\lambda^l$, where $0 \leq \lambda \leq 1$:

$$\mathcal{L}_{\text{MOSAIC}} = \mathcal{L}_{CE}(\hat{r}_{\psi_i}, \mathcal{D}_{h_i}) + \mathbb{E}_{h^+ \sim \mathcal{D}_{h_i}} \left[ \sum_{l=0}^{L} \lambda^l \hat{r}_{\psi_i}(s_{j-l}, a_{j-l}) \right] - \mathbb{E}_{h^- \sim \mathcal{D}_{h_i}} \left[ \sum_{l=0}^{L} \lambda^l \hat{r}_{\psi_i}(s_{j-l}, a_{j-l}) \right]$$

### D. Preference learning with MOSAIC

Similar to other preference-based RL methods, MOSAIC (see Alg. 1), interleaves policy with reward learning. We outline in Figure. 2, a high-level representation of the overall integration of learning a reward function from preferences with an LLM. In step A, we train and sample a $\pi_\omega$ as according to PEBBLE [57] to acquire a dataset of trajectory segments $\mathcal{D}_\sigma$. During step B, a feedback session is initiated wherein trajectory segments $\sigma$, are selected according to the weighted variance of the ensemble as in Sec. IV-C. We leverage the LLM to process the human prompts as according to Sec. IV-B to obtain $\mathcal{D}_f$, and further highlights dataset $\mathcal{D}_h$ as introduced in Sec. IV-C. To obtain $\mathbf{\Omega}_w$ we use both global prompts $\mathcal{P}$ and intrinsic features $\mathcal{F}$ with $\text{LLM}_{\text{weights}}$. Notably, while the weights of objectives within $\mathbf{\Omega}_w$ undergo minor adjustments–leading to a dynamic weights setting–in each feedback session, the entire set of prompts is consistently considered. In Step C, gradient descent is performed on parameters $\psi_i$ using Equation III to fine-tune each $r_{\psi_i}$. Upon updating each reward function, they are weighted according to $\mathbf{\Omega}_w$ to derive $\hat{\boldsymbol{r}}_{\mathbf{\Omega}}$. In Step D an explanation is provided by the LLM, taking into account all prompts and the final estimation of the relative importance of objectives within $\mathbf{\Omega}_w$. For prompt examples representing $\text{LLM}_{\text{weights}}$ and $\text{LLM}_{\text{explanation}}$, refer to supplemental materials of the paper at https://sites.google.com/view/multi-pref/.

**Algorithm 1:** MOSAIC

$\mathcal{D}_\sigma, \mathcal{D}_{f_i}, \mathcal{D}_{h_i}, \mathcal{P} \leftarrow \emptyset$;
$\mathcal{Z} \leftarrow \text{getFeatureSet}()$;
**for** *epoch* $= 1, 2, \ldots$ **do**
  /* Train policy $\pi_\omega$ */ ▷ // Step A
  $\pi_\omega \leftarrow \text{train}(\pi_\omega, \hat{\boldsymbol{r}}_{\mathbf{\Omega}})$
  Sample $\pi_\omega$ within an environment to obtain $\mathcal{D}_\sigma^{\text{new}} = \{\sigma^i\}_{i=1}^{D_\sigma}$
  Store new trajectories $\mathcal{D}_\sigma \leftarrow \mathcal{D}_\sigma \cup \mathcal{D}_\sigma^{\text{new}}$
  $\mathcal{D}_\sigma \leftarrow \text{sampleSegments}(\pi_\omega)$
  /* Obtain Human-feedback */ ▷ // Step B
  $\mathcal{D}_{(\sigma_1, \sigma_2)} \leftarrow \text{samplePairs}(\mathcal{D}_\sigma)$
  $\mathcal{P} \leftarrow \mathcal{P} \cup \text{collectPrompts}(\mathcal{D}_{(\sigma_0, \sigma_1)})$
  $\mathcal{R}, \mathbf{\Omega}_w \leftarrow \text{LLM}_{\text{pref}}(\mathcal{P}, \mathcal{F}), \text{LLM}_{\text{weights}}(\mathcal{P}, \mathcal{F})$
  **for** *each* $\hat{r}_{\psi_i}$ **do**
    $D_{f_i} \leftarrow \text{processDataset}(\mathcal{R}) \cup D_{f_i}$
    $D_{h_i} \leftarrow \text{processHighlights}(D_{f_i}, g_i) \cup D_{h_i}$
    /* Train $\hat{r}_{\psi_i}(s_t, a_t)$ */ ▷ // Step C
    $\hat{r}_{\psi_i}(s_t, a_t) \leftarrow \text{train}(\mathcal{D}_{h_i})$ w.r.t $\psi_i$ in Eq. IV-C
explanation $\leftarrow \text{LLM}_{\text{explanation}}(\mathcal{P}, \mathcal{Z})$ ▷ // Step D
**return** $\pi_\omega^*, \hat{\boldsymbol{r}}_{\mathbf{\Omega}}$

## V. Experiments

In this section, we investigate the effectiveness of MOSAIC in adhering to the different objectives. Our hypothesis is that MOSAIC can address challenges such as objective collapse and causal confusion more efficiently. To validate our hypothesis, we ablate MOSAIC on its several components, followed by an empirical study involving real human feedback. Our investigation is primarily driven by the following research questions: **(Q1)**: How impactful are the different components of MOSAIC–query sampling strategy and reward regularization–in better defining the different objectives in MOSAIC?; **(Q2)**: Can MOSAIC mitigate objective collapse and solve multi-objective control tasks through preferences and highlights?; **(Q3)**: Is MOSAIC capable of aligning policies with real human input preferences and prompts?

**Simulated Experiment Setup.** In order to validate the proficiency of MOSAIC in learning multiple objectives, experiments were conducted within two distinct multi-objective MO-Gymnasium [61] environments: `Cheetah` and `Hopper`; and adapted four high-dimensional robotic tasks from Meta-world [62] to account for multiple objectives: `Button Press`, `Button Press Wall`, `Drawer Close`, and `Drawer Close`. To formulate the preferences for different objectives, an oracle was developed, which, upon receiving a pair of segments, dispenses several preferences based on the total reward accumulated for each objective. To simulate the incorporation of highlights, we extended the oracle to provide nuanced feedback over partial trajectories for $1/4$ of the objectives when labelling queries concerning the supplementary objective. We provide concrete hyper-parameters for the models used in the supplemental materials

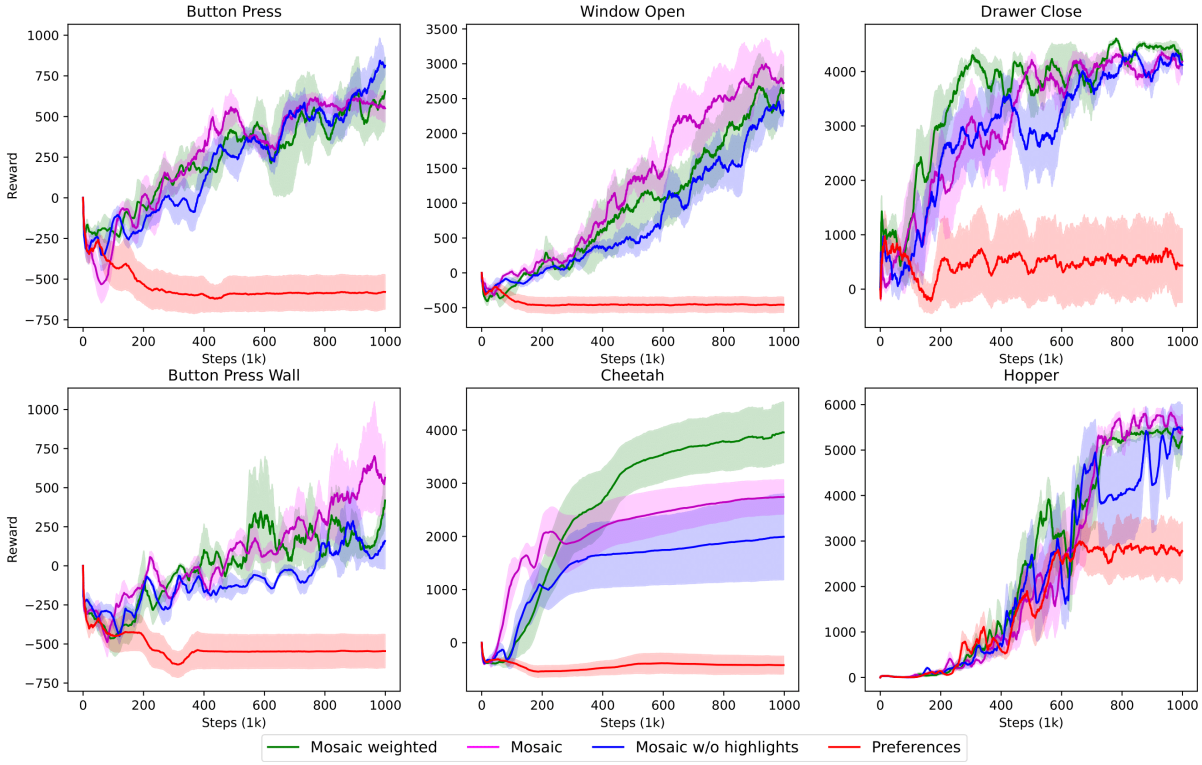


Fig. 3: Learning curves pertaining to the sum of all objectives for all four conditions. The solid lines and shaded areas represent the mean and standard error respectively. MOSAIC consistently surpasses a single preference approach by a large margin.

TABLE I: Averaged scores for each objective and condition, for the environments considered. MOSAIC consistently outperforms learning from a single preference convoluted on different objectives.

| Environment | Objectives | Weighted Mosaic | Mosaic | Mosaic w/o highlight | Preference |
|---|---|---|---|---|---|
| `Button Press` | Total | 660 ± 216 | 563 ± 101 | **715** ± 52 | -572 ± 104 |
| | Env Reward | 922 ± 208 | 769 ± 95 | **946** ± 54 | 50 ± 4 |
| | Closeness | -55 ± 8 | -46 ± 6 | **-42** ± 2 | -247 ± 11 |
| | Avoid left | -207 ± 53 | **-159** ± 59 | -188 ± 35 | -376 ± 107 |
| `Window Open` | Total | 2481 ± 448 | **2654** ± 451 | 2288 ± 120 | -457 ± 105 |
| | Env Reward | 2957 ± 441 | **3137** ± 437 | 2745 ± 117 | 114 ± 8 |
| | Closeness | **-38** ± 4 | -47 ± 9 | -42 ± 4 | -270 ± 11 |
| | Avoid left | -437 ± 13 | -436 ± 22 | -414 ± 15 | **-301** ± 108 |
| `Drawer Close` | Total | 4144 ± 189 | **4443** ± 75 | 3949 ± 239 | 427 ± 748 |
| | Env Reward | 4382 ± 189 | **4646** ± 74 | 4195 ± 222 | 978 ± 769 |
| | Closeness | -81 ± 9 | **-54** ± 4 | -73 ± 7 | -249 ± 11 |
| | Avoid left | -155 ± 15 | **-148** ± 20 | -172 ± 35 | -301 ± 108 |
| `Button Press Wall` | Total | **470** ± 244 | 456 ± 212 | 201 ± 220 | -545 ± 105 |
| | Env Reward | **625** ± 272 | 562 ± 212 | 378 ± 175 | 14 ± 6 |
| | Closeness | -72 ± 7 | **-45** ± 2 | -76 ± 12 | -257 ± 8 |
| | Avoid left | -83 ± 37 | **-59** ± 21 | -100 ± 80 | -301 ± 108 |
| `Cheetah` | Total | **3778** ± 407 | 2360 ± 435 | 1694 ± 629 | -426 ± 168 |
| | Speed | **4123** ± 410 | 2692 ± 441 | 2017 ± 629 | -62 ± 153 |
| | Control | -345 ± 6 | -331 ± 10 | **-322** ± 5 | -364 ± 16 |
| `Hopper` | Total | **5363** ± 148 | 4795 ± 381 | 4141 ± 1109 | 2927 ± 644 |
| | Speed | **2914** ± 206 | 2707 ± 228 | 2023 ± 613 | 1895 ± 437 |
| | Height | **2402** ± 19 | 1993 ± 239 | 2136 ± 504 | 1342 ± 255 |
| | Control | 46 ± 64 | 94 ± 68 | **116** ± 69 | -322 ± 66 |

at https://sites.google.com/view/multi-pref/. To answer Q1 we ablate the different components of MOSAIC in four conditions: **Preference-based [57]**: using a single preference which is given by the oracle when weighting all the objectives, following PEBBLE; **MOSAIC w/o highlights**: employs concrete objective information without regularization; **MOSAIC**: with reward regularization; **Weighted MOSAIC**: with weighted variance as a query sampling mechanism.

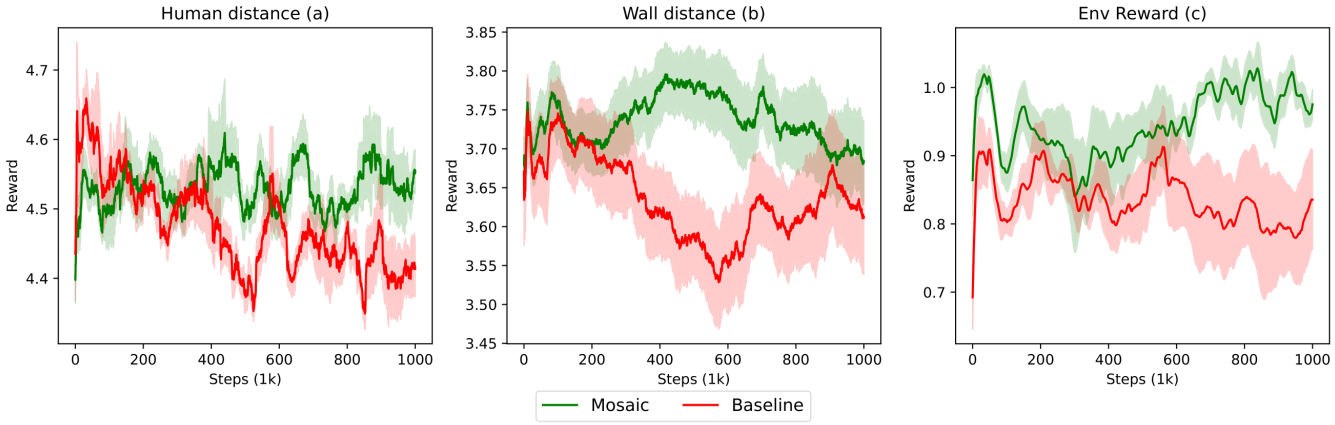


Fig. 4: Experimental of MOSAIC with respect to the social navigation scenario, upon acquiring real human feedback. The baseline uses preferences without textual information, equivalent to the previous preference-based method. Figures (a) depict a reward solely considering the distance to humans, (b) a reward considering wall distance, and (c) the environmental reward.

### A. Benchmark Results

Table I shows the performance of MOSAIC and its various components against a single preference baseline. Across all environments, MOSAIC consistently outperforms the single preference approach on all objectives. In all conditions, MOSAIC, with highlights and/or weighted variance, yields the highest scores. In most environments, the weighted variance sampling strategy, devised in Sec.IV-A, enables higher scores for the most important objective. This experiment demonstrates that providing a single preference, when there are multiple objectives, can lead to causal confusion, thereby supporting Q2. As different objectives may take precedence over the preference, it is akin to introducing noise into the main objective. This is further corroborated in Fig.3, where the preference-based baseline struggles to learn in most environments and converges to a local minimum.

| **Objective weights $\mathbf{\Omega}_w$** | | | |
|---|---|---|---|
| Goal Dist | Human Dist | Speed | Wall Dist |
| 0.3 | 0.4 | 0.2 | 0.1 |

TABLE II: The weights provided by the LLM for each objective based on the human feedback.

### B. Real human experiment results

In order to evaluate how MOSAIC performs when utilizing real human feedback, we make use of a social navigation scenario, SocialNav similar to the one in [63], where the robot learns to navigate in a corridor with humans. The intrinsic objective features, $\mathcal{Z}$, considered for this task include distance to the goal, proximity to humans, distance to walls, and the robot's speed. We monitor three distinct reward functions alongside, each weighted by the significance ascribed by the LLM based on human descriptions, denoted as $\mathbf{\Omega}_w$. The LLM is prompted to assign these weights subsequent to the collection of all feedback. Upon compiling all human preferences, we utilized the LLM, employing all textual feedback to determine the objective weights, as outlined in previous sections. The LLM was observed to prioritize human safety, notably emphasizing the maintenance of a substantial distance. Following human safety, the distance to the goal was identified as the second most crucial objective.

In contrast, the robot's speed and its distance from walls were deemed least vital. Given that the LLM assigned appropriate importance weights (refer to Tab. II), this adequately addresses Q3 and supports our hypothesis that an LLM can assist in resolving causal confusion by contributing causal reasoning regarding the importance of various features. In Fig 4, the benefits of adhering to additional human prompts are apparent, impacting both the tuning and derivation of a more substantively meaningful final policy for the simulated social navigation environment. This highlights the flexibility and robustness of the multi-objective paradigm adopted by MOSAIC. By extracting varied features from human prompts, we ensure the distinctiveness of the features is retained in the final trained policy. Further details regarding the user study, including participant demographics and preference collection, are available in the appendix on the project website: `https://sites.google.com/view/multi-pref/`.

## VI. Discussion and Limitations

In this work, we introduce MOSAIC, a novel multi-objective preference-based RL paradigm that is capable of addressing causal confusion by leveraging the zero-shot capabilities of an LLM to extract and rank objectives from human prompts. With real users, MOSAIC successfully processed human subjective objective preferences, initially prioritising safety and emphasizing goal attainment. Moreover, we have elucidated the capability of further interrogating the LLM concerning the resulting objective weights by employing additional prompts, thereby facilitating valuable explainability to humans.

## ACKNOWLEDGMENT

This research has been carried out as part of the Vinnova Competence Center for Trustworthy Edge Computing Systems and Applications at KTH, and partially supported by the

Swedish Foundation for Strategic Research (SSF FFL18-0199) and the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

## REFERENCES


[1] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," *Advances in neural information processing systems*, vol. 30, 2017.

[2] B. Ibarz, J. Leike, T. Pohlen, G. Irving, S. Legg, and D. Amodei, "Reward learning from human preferences and demonstrations in atari," *Advances in neural information processing systems*, vol. 31, 2018.

[3] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving, "Fine-tuning language models from human preferences," *arXiv preprint arXiv:1909.08593*, 2019.

[4] S. Casper, X. Davies, C. Shi, T. K. Gilbert, J. Scheurer, J. Rando, R. Freedman, T. Korbak, D. Lindner, P. Freire, *et al.*, "Open problems and fundamental limitations of reinforcement learning from human feedback," *arXiv preprint arXiv:2307.15217*, 2023.

[5] A. Peng, A. Netanyahu, M. K. Ho, T. Shu, A. Bobu, J. Shah, and P. Agrawal, "Diagnosis, feedback, adaptation: A human-in-the-loop framework for test-time policy adaptation," in *International Conference on Machine Learning*. PMLR, 2023, pp. 27 630–27 641.

[6] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano, "Learning to summarize with human feedback," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3008–3021, 2020.

[7] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 730–27 744, 2022.

[8] 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, pp. 27–40, 2018.

[9] X. Wang, K. Lee, K. Hakhamaneshi, P. Abbeel, and M. Laskin, "Skill preferences: Learning to extract and execute robotic skills from human feedback," in *Conference on Robot Learning*. PMLR, 2022, pp. 1259–1268.

[10] D. J. Hejna III and D. Sadigh, "Few-shot preference learning for human-in-the-loop rl," in *Conference on Robot Learning*. PMLR, 2023, pp. 2014–2025.

[11] A. Bobu, D. R. Scobee, J. F. Fisac, S. S. Sastry, and A. D. Dragan, "Less is more: Rethinking probabilistic models of human behavior," in *Proceedings of the 2020 acm/ieee international conference on human-robot interaction*, 2020, pp. 429–437.

[12] H. J. Jeon, S. Milli, and A. Dragan, "Reward-rational (implicit) choice: A unifying formalism for reward learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4415–4426, 2020.

[13] S. Booth, S. Sharma, S. Chung, J. Shah, and E. L. Glassman, "Revisiting human-robot teaching and learning through the lens of human concept learning," in *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, 2022, pp. 147–156.

[14] P. De Haan, D. Jayaraman, and S. Levine, "Causal confusion in imitation learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[15] J. Tien, J. Z.-Y. He, Z. Erickson, A. D. Dragan, and D. Brown, "A study of causal confusion in preference-based reward learning," *arXiv preprint arXiv:2204.06601*, 2022.

[16] H. Khayrallah, S. Trott, and J. Feldman, "Natural language for human robot interaction," in *International Conference on Human-Robot Interaction (HRI)*, 2015.

[17] Z. Li, Y. Mu, Z. Sun, S. Song, J. Su, and J. Zhang, "Intention understanding in human–robot interaction based on visual-nlp semantics," *Frontiers in Neurorobotics*, vol. 14, p. 610139, 2021.

[18] S. Pate, W. Xu, Z. Yang, M. Love, S. Ganguri, and L. L. Wong, "Natural language for human-robot collaboration: Problems beyond language grounding," *arXiv preprint arXiv:2110.04441*, 2021.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[20] R. Mokady, A. Hertz, and A. H. Bermano, "Clipcap: Clip prefix for image captioning," *arXiv preprint arXiv:2111.09734*, 2021.

[21] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[22] A. Zeng, M. Attarian, K. M. Choromanski, A. Wong, S. Welker, F. Tombari, A. Purohit, M. S. Ryoo, V. Sindhwani, J. Lee, *et al.*, "Socratic models: Composing zero-shot multimodal reasoning with language," in *The Eleventh International Conference on Learning Representations*, 2022.

[23] C. F. Hayes, R. Rădulescu, E. Bargiacchi, J. Källström, M. Macfarlane, M. Reymond, T. Verstraeten, L. M. Zintgraf, R. Dazeley, F. Heintz, *et al.*, "A practical guide to multi-objective reinforcement learning and planning," *Autonomous Agents and Multi-Agent Systems*, vol. 36, no. 1, p. 26, 2022.

[24] C. Wirth, R. Akrour, G. Neumann, J. Fürnkranz, *et al.*, "A survey of preference-based reinforcement learning methods," *Journal of Machine Learning Research*, vol. 18, no. 136, pp. 1–46, 2017.

[25] K. Amin, N. Jiang, and S. Singh, "Repeated inverse reinforcement learning," *Advances in neural information processing systems*, vol. 30, 2017.

[26] X. Liang, K. Shu, K. Lee, and P. Abbeel, "Reward uncertainty for exploration in preference-based reinforcement learning," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=OWZVD-l-ZrC

[27] J. Park, Y. Seo, J. Shin, H. Lee, P. Abbeel, and K. Lee, "SURF: Semi-supervised reward learning with data augmentation for feedback-efficient preference-based reinforcement learning," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=TfhfZLQ2EJO

[28] R. Hu, S. L. Chau, J. F. Huertas, and D. Sejdinovic, "Explaining preferences with shapley values," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022. [Online]. Available: https://openreview.net/forum?id=-me36V0os8P

[29] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," *arXiv preprint arXiv:1606.06565*, 2016.

[30] D. Hadfield-Menell, S. Milli, P. Abbeel, S. J. Russell, and A. Dragan, "Inverse reward design," *Advances in neural information processing systems*, vol. 30, 2017.

[31] N. Ho, L. Schmid, and S.-Y. Yun, "Large language models are reasoning teachers," *arXiv preprint arXiv:2212.10071*, 2022.

[32] Y. Qian, B. Wang, S. Ma, W. Bin, S. Zhang, D. Zhao, K. Huang, and Y. Hou, "Think twice: A human-like two-stage conversational agent for emotional response generation," in *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, 2023, pp. 727–736.

[33] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[34] T. D. Kulkarni, A. Gupta, C. Ionescu, S. Borgeaud, M. Reynolds, A. Zisserman, and V. Mnih, "Unsupervised learning of object keypoints for perception and control," *Advances in neural information processing systems*, vol. 32, 2019.

[35] P. Florence, L. Manuelli, and R. Tedrake, "Self-supervised correspondence in visuomotor policy learning," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 492–499, 2019.

[36] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer, "Lit: Zero-shot transfer with locked-image text tuning," *arXiv preprint arXiv:2111.07991*, 2021.

[37] M. Tsimpoukelli, J. L. Menick, S. Cabi, S. Eslami, O. Vinyals, and F. Hill, "Multimodal few-shot learning with frozen language models," *Advances in Neural Information Processing Systems*, vol. 34, pp. 200–212, 2021.

[38] K. Zakka, A. Zeng, P. Florence, J. Tompson, J. Bohg, and D. Dwibedi, "Xirl: Cross-embodiment inverse reinforcement learning," in *Conference on Robot Learning*. PMLR, 2022, pp. 537–546.

[39] E. Gavves, T. Mensink, T. Tommasi, C. G. Snoek, and T. Tuytelaars, "Active transfer learning with zero-shot priors: Reusing past datasets for future tasks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2731–2739.

[40] D. Schwab, Y. Zhu, and M. Veloso, "Zero shot transfer learning for robot soccer," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 2018, pp. 2070–2072.

[41] W. Ying, Y. Zhang, J. Huang, and Q. Yang, "Transfer learning via learning to transfer," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5085–5094.

[42] J. W. Soh, S. Cho, and N. I. Cho, "Meta-transfer learning for zero-shot super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3516–3525.

[43] D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley, "A survey of multi-objective sequential decision-making," *Journal of Artificial Intelligence Research*, vol. 48, pp. 67–113, 2013.

[44] P. Vamplew, R. Dazeley, A. Berry, R. Issabekov, and E. Dekker, "Empirical evaluation methods for multiobjective reinforcement learning algorithms," *Machine learning*, vol. 84, pp. 51–80, 2011.

[45] C. F. Hayes, R. Rădulescu, E. Bargiacchi, J. Kallstrom, M. Macfarlane, M. Reymond, T. Verstraeten, L. M. Zintgraf, R. Dazeley, F. Heintz, *et al.*, "A brief guide to multi-objective reinforcement learning and planning," in *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, 2023, pp. 1988–1990.

[46] T. Yu, Y. Tian, J. Zhang, and S. Sra, "Provably efficient algorithms for multi-objective competitive rl," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 167–12 176.

[47] F. Ho and S. Nakadai, "Preference-based multi-objective multi-agent path finding," in *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, 2023, pp. 2000–2002.

[48] J. Xu, Y. Tian, P. Ma, D. Rus, S. Sueda, and W. Matusik, "Prediction-guided multi-objective reinforcement learning for continuous robot control," in *International conference on machine learning*. PMLR, 2020, pp. 10 607–10 616.

[49] S. Natarajan and P. Tadepalli, "Dynamic preferences in multi-criteria reinforcement learning," in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 601–608.

[50] R. Yang, X. Sun, and K. Narasimhan, "A generalized algorithm for multi-objective reinforcement learning and policy adaptation," *Advances in neural information processing systems*, vol. 32, 2019.

[51] A. Abels, D. Roijers, T. Lenaerts, A. Nowé, and D. Steckelmacher, "Dynamic weights in multi-objective deep reinforcement learning," in *International conference on machine learning*. PMLR, 2019, pp. 11–20.

[52] T. Basaklar, S. Gumussoy, and U. Ogras, "Pd-morl: Preference-driven multi-objective reinforcement learning algorithm," in *Deep Reinforcement Learning Workshop NeurIPS 2022*, 2022.

[53] L. Chen, H. Fernando, Y. Ying, and T. Chen, "Three-way trade-off in multi-objective learning: Optimization, generalization and conflict-avoidance," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[54] H. Shao, L. Cohen, A. Blum, Y. Mansour, A. Saha, and M. Walter, "Eliciting user preferences for personalized multi-objective decision making through comparative feedback," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[55] A. Wilson, A. Fern, and P. Tadepalli, "A bayesian approach for policy learning from trajectory preference queries," *Advances in neural information processing systems*, vol. 25, 2012.

[56] 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, pp. 324–345, 1952.

[57] K. Lee, L. Smith, and P. Abbeel, "Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training," *arXiv preprint arXiv:2106.05091*, 2021.

[58] P. Mirowski, R. Pascanu, F. Viola, H. Soyer, A. J. Ballard, A. Banino, M. Denil, R. Goroshin, L. Sifre, K. Kavukcuoglu, *et al.*, "Learning to navigate in complex environments," *arXiv preprint arXiv:1611.03673*, 2016.

[59] E. Shelhamer, P. Mahmoudieh, M. Argus, and T. Darrell, "Loss is its own reward: Self-supervision for reinforcement learning," *arXiv preprint arXiv:1612.07307*, 2016.

[60] L. Pinto and A. Gupta, "Learning to push by grasping: Using multiple tasks for effective learning," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 2161–2168.

[61] F. Felten, L. N. Alegre, A. Nowé, A. L. C. Bazzan, E. G. Talbi, G. Danoy, and B. C. da. Silva, "A toolkit for reliable benchmarking and research in multi-objective reinforcement learning," in *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023.

[62] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine, "Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning," in *Conference on robot learning*. PMLR, 2020, pp. 1094–1100.

[63] D. Marta, C. Pek, G. I. Melsión, J. Tumova, and I. Leite, "Human-feedback shield synthesis for perceived safety in deep reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 7, no. 1, pp. 406–413, 2021.