2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

SEQUEL: Semi-Supervised Preference-based RL with Query Synthesis via Latent Interpolation

Daniel Marta*[1], Simon Holk*[1], Christian Pek[2], and Iolanda Leite[1]

***Abstract*— Preference-based reinforcement learning (RL) poses as a recent research direction in robot learning, by allowing humans to teach robots through preferences on pairs of desired behaviours. Nonetheless, to obtain realistic robot policies, an arbitrarily large number of queries is required to be answered by humans. In this work, we approach the sample-efficiency challenge by presenting a technique which synthesizes queries, in a semi-supervised learning perspective. To achieve this, we leverage latent variational autoencoder (VAE) representations of trajectory segments (sequences of state-action pairs). Our approach manages to produce queries which are closely aligned with those labeled by humans, while avoiding excessive uncertainty according to the human preference predictions as determined by reward estimations. Additionally, by introducing variation without deviating from the original human's intents, more robust reward function representations are achieved. We compare our approach to recent state-of-the-art preference-based RL semi-supervised learning techniques. Our experimental findings reveal that we can enhance the generalization of the estimated reward function without requiring additional human intervention. Lastly, to confirm the practical applicability of our approach, we conduct experiments involving actual human users in a simulated social navigation setting. Videos of the experiments can be found at https://sites.google.com/view/rl-sequel**

I. Introduction

Recent robot learning advances in RL lean towards leveraging human knowledge and guidance as an interactive and efficient medium [1], [2]. By inferring reward functions from humans, robot policies can be made user-specific [3], adapted efficiently [4], [5] and even aligned with natural language [6], [7]. Many recent works propose to leverage demonstrations, preferences, or combinations of both [8]–[14]. Preference learning [15]–[17] poses as a data-efficient learning approach which is able to convey subtle and multi-modal nuances [18]. Indeed, preference-driven teaching introduces the critical component of structural alignment [19]–[21], while also fostering a significant diversity in trajectory paths (state-action sequences), both essential for robot learning.

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.

* These authors contributed equally to this work.

1 Authors are with the Division of Robotics, Perception and Learning, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, 114 28 Stockholm, Sweden. The authors are also affiliated with Digital Futures. Email: {dlmarta, sholk, iolanda}@kth.se

2 Authors are with the Dept. of Cognitive Robotics, TU Delft, 2628 CD Delft, The Netherlands. Email: c.pek@tudelft.nl

While preference learning allows human users to label large amounts of state-action sequences, current preference-based approaches [16], [22] require large amounts of actual human feedback to be applicable to realistic robot tasks [23]. While there are recent works which focus on pre-training [23], [24] to tackle this issue, many feedback efficiency challenges remain. For example, small sample sizes of labeled preferences can lead to reward exploitation — a failure in reward inference which can lead to sub-par behaviours [25], [26]. Moreover, preferences exhibit a strong correlation with causality [27], [28]. That is to say, there is often an underlying reason if one behaviour is favoured over another. Ignoring this aspect can result in a distributional shift, which leads to a phenomenon known as *causal confusion* [29], [30], where additional interactions with the environment can degrade performance. Moreover, Tien et al. [31] showed in their study on preference-based RL that reward functions often succumb to false correlations and reward hacking, achieving minimal test errors but failing to extend to out-of-distribution states. This leads to additional human burden and ultimately necessitates the extraction of internal and implicit representations via brute force by over-querying humans.

We introduce SEQUEL: SEmi-Supervised Preference-based RL with QUEry Synthesis via Latent Interpolation. In this work, we view the feedback inefficiency and generalization problem of preference-based RL through the lens of representation learning and semi-supervised learning [32]. Akin to the idea of using generative models to learn environment dynamics for policy learning [33]–[35], we address the problem of limited human-labeled query data. We propose to leverage a readily available policy space and expand the available queries in a comprehensive way by interpolating between query elements (trajectories) in the latent space. There are numerous works which have endorsed the effectiveness of latent space interpolation [36]–[39], particularly in intricate input spaces like images, thereby facilitating a seamless transition between data points. Through this process, we can derive new queries from the interpolated latent space vectors, which maintain an intrinsic connection to the original queries, allowing us to delve into the decision boundaries of the reward model. The objective is to introduce minor yet correlated variations to the labeled queries, significantly boosting the quality and diversity of the training data, enhancing the reward function out-of-distribution robustness and expediting learning. To demonstrate the effectiveness of our method, we provide empirical evidence, including comparisons with state-of-the-art approaches in preference-based RL. We also evaluate SEQUEL against actual human

data, further reinforcing its merits and performance.

II. Related Work

Learning from preferences in RL. Preference-based RL has an active research front and enjoys from an ever growing body of literature [15]. Prior work in preference learning [40]–[42] presents utility functions as linear functions and offers closed-form solutions for the expected utility of selection of pairs of demonstrations to better understand the expert's preferences and align a policy. There is also previous work which improves policies through preferences either by pre-defining features [43] or through Bayesian approaches [44], [45]. However, the constraints and assumptions imposed on the reward function space can be hard to adapt to the intricate objectives of modern robotic tasks [23]. Contemporary research approaches impose few restrictions on reward function modality [13], [16], [46] and show promising results on robotic benchmarks. Nonetheless, they require extensive usage of human feedback which still limits their applicability to real robotic tasks [23], [47]–[49]. While there is recent research which approaches the feedback inefficiency challenge through pre-training [23], [24], or bi-level optimization [50], we explore a representation learning approach to exploit labeled human queries. While we do not approach directly the preference explanation [49] problem, our work can comprehensively utilize the latent space of labeled queries to improve the robustness of preference-based RL reward functions and can be used in other frameworks.

Data augmentation and semi-supervised learning. Semi-supervised learning is a known and striving field in machine-learning [32], [51]. Prior work considered entropy minimization [52], consistency regularization [53], or pseudo-labelling [53], [54] to improve generalization on sets of unlabeled data. Kingma et al. [55] popularized generative models to improve semi-supervised learning performance in computer vision tasks. Unsupervised data augmentation has been considered [56], where a clear link emerges between better data augmentation leading to significantly better semi-supervised learning. In RL, even simple data augmentation techniques such as in the form of input perturbations [9], [57] can effectively regularize and improve the robustness of policies [58]. Unsupervised representations have also been used to improve data efficiency and generalization of policies [59], [60]. Closer to our work in preference-based RL, Park et al. [48] propose to label unlabeled queries in a semi-supervised learning approach by utilizing pseudo-labelling and temporal cropping. We build on his work by proposing to leverage generative models to synthesize newer queries to improve out-of-distribution performance. Our work can also be seen as a form of distillation [61]–[63] where the generative (teacher) model can be used as an effective regularizer to train a reward (student) model.

III. Preliminaries

From a state s_t, a robot provides an action a_t following policy $\pi_\omega(a_t, s_t)$ parameterized by ω. This action prompts a reward $r(s_t, a_t)$ and a new state s_{t+1} from an environment which is modelled as a Markov decision process (MDP). The objective of the robot is to obtain an optimal policy $\pi^*_\omega(a_t, s_t)$ which maximizes the expected discounted sum of rewards. In this work, we estimate a model for the reward function from humans in a feedback-efficient manner.

Preference-based RL. As in [16] we formulate the problem of estimating a reward function $\hat{r}_\psi$ parameterized by ψ from preferences as a supervised learning problem. The goal of preference-based RL [15], [16] is to infer state-action reward information from pairs of trajectory segments. Trajectory segments [64] consist of sequences of state-action pairs, denoted as $\sigma^j = ((s^j_t, a^j_t), \ldots, (s^j_{t+l-1}, a^j_{t+l-1}))$, where j denotes the index of the segment, which contains state-action pairs ranging from t to $t+l$, where l is the length of the segment. Pairs of trajectory segments, denoted as (σ^0, σ^1), are given to humans who then assign a preference $y \in \{0, 0.5, 1\}$. If the human prefers σ^0 over σ^1 they provide $y = 0$ which is noted as $\sigma^0 \succ \sigma^1$, conversely $y = 1$ reads as $\sigma^1 \succ \sigma^0$ and $y = 0.5$ denotes equal preference of both segments. Following the Bradley-Terry model [65], the probability of a human preferring $\sigma^0 \succ \sigma^1$ assuming it depends exponentially on the sum of rewards over the length of the segments is given by:

$$P_\psi[\sigma^0 \succ \sigma^1] = \frac{\exp(\sum_t \hat{r}_\psi(s^0_t, a^0_t))}{\exp(\sum_t \hat{r}_\psi(s^0_t, a^0_t)) + \exp(\sum_t \hat{r}_\psi(s^1_t, a^1_t))} \quad (1)$$

In this formulation, the reward model $\hat{r}_\psi$ can be trained as a binary classifier to predict human preferences on unseen segments, and used as a proxy for the reward function. The preferences provided by humans are stored alongside the corresponding segments and stored on a labeled dataset $\mathcal{D}_l$ consisting of triplets (σ^0, σ^1, y). When optimizing $\hat{r}_\psi$ we sample from $\mathcal{D}_l$ and minimize the binary cross-entropy loss:

$$\begin{aligned}\mathcal{L}_{CE}(\hat{r}_\psi, \mathcal{D}_l) = - \mathop{\mathbb{E}}_{(\sigma^1, \sigma^2, y) \sim \mathcal{D}_l} [(1\text{–}y) \log P_\psi(\sigma^1 \succ \sigma^2) \\ + y \log P_\psi(\sigma^2 \succ \sigma^1)]\end{aligned} \quad (2)$$

A. Semi-supervised learning techniques in preference-based RL

Semi-supervised learning aims at leveraging unlabeled samples to improve a model's robustness and generalization. Consider $\mathcal{L}_{su}$ as a supervised loss and $\mathcal{L}_u$ as an unsupervised loss, the semi-supervised loss [32] can be typically written as $\mathcal{L}_{SSL} = \mathcal{L}_{su} + \lambda \mathcal{L}_u$, where $\lambda \in [0, 1]$ is a balancing parameter between both losses. Next, we discuss two techniques presented in SURF [48].

Pseudo-labeling. We revisit pseudo-labeling [53], [54] in the context of preference-based RL [48]. While we directly obtain preferences y from humans, we may take an estimation $\hat{y}$ for unlabeled queries by picking the segment which is more likely to be chosen. Consider unlabeled segments σ^0_u and σ^1_u, then we define $\hat{y}$ as:

$$\hat{y} = \begin{cases} 0, & \text{if } P_\psi\ [\sigma^0_u \succ \sigma^1_u] > 0.5 \\ 1, & \text{if otherwise.} \end{cases} \quad (3)$$

Moreover, only queries which have a high confidence level $P_\psi[\sigma^0_u \succ \sigma^1_u] > \tau, \tau \in [0, 1]$ are stored in triples alongside

unlabeled segments to form D_u. The semi-supervised loss is as follows:

$$\mathcal{L}_{SSL} = \mathcal{L}_{CE}(\hat{r}_\psi, \mathcal{D}_l) + \lambda \mathcal{L}_{CE}(\hat{r}_\psi, \mathcal{D}_u) \tag{4}$$

Temporal cropping. Temporal cropping is widely used as a semi-supervised technique [53], [66]. In the context of preference-based RL [48], considering our initial definition of a segment, *temporal crop* of σ^j is another trajectory segment $\sigma^{j'} = ((s^j_{t'}, a^j_{t'}), \ldots, (s^j_{t'+l'-1}, a^j_{t'+l'-1}))$, where $t \leq t' < t' + l' \leq t + l$ and $l' \leq l$. Thus, a temporal crop of $\sigma^{j'}$ is any sub-sequence of σ^j.

IV. SEQUEL

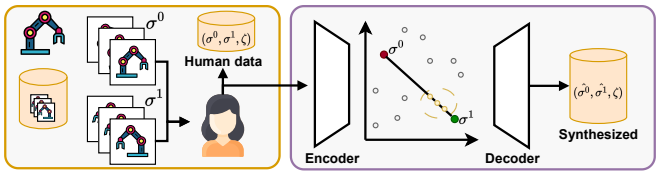

Fig. 1: An overview of SEQUEL: We start by collecting segments and presenting them to humans, which in turn provide preferences. Then we train a VAE to interpolate and synthesise new queries to improve the robustness of $\hat{r}_\psi$.

In this section, we formalize **SEQUEL**: **SE**mi-Supervised Preference-based RL with **QUE**ry Synthesis via **L**atent Interpolation. Our proposition is to leverage the structural composition of query elements to synthesize novel queries by interpolating the latent space of labeled queries. Figure 1 provides a visual representation of our framework, while Algorithm 1 offers an overview of the technique combined with preference-based RL.

A. SEQUEL's algorithm

First we start by acquiring trajectory segments σ obtained from consecutively sampling a policy π_ω and store them in a dataset $\mathcal{D}_\sigma = \{\sigma^i\}_{i=1}^{D_\sigma}$. Then, we sample segments from $\mathcal{D}_\sigma$ to query humans resulting in triplets (σ^0, σ^1, y) stored in $\mathcal{D}_l$. As in standard preference-based RL, the policy training is interleaved with reward training. We sample $\mathcal{D}_l$ to obtain a new estimate of $\hat{r}_\psi$ through Eq. 2, and use this estimation to resume train policy of π_ω following PPO [67].

To build an informative latent space, we explore a compact representation for segments. More concretely, we propose to leverage a Variational Autoencoder (VAE) [68] on segments collected from $\mathcal{D}_\sigma$. We start by defining an encoder $q_\phi(z|\sigma^j)$ which maps a trajectory segment σ^j to a distribution over latent representations $\mathcal{Z}$, where ϕ are the parameters of the encoder network. Then, we define a decoder $p_\theta(\sigma^j|z)$, parameterized by θ which aims at reconstructing the original segment σ^j. The objective is to precisely reconstruct the original segments while encouraging the encoding distributions $\mathcal{Z}$ to be close to a standard normal distribution. Following Kingma et al. [68], the point-wise loss of σ^j can be written as:

$$\begin{aligned}\mathcal{L}_{VAE}(\theta, \phi; \sigma^j) = &\mathop{\mathbb{E}}_{z \sim q_\phi(z|\sigma^j)}[-\log p_\theta(\sigma^j|z)] \\ &+ \beta \cdot D_{KL}(q_\phi(z|\sigma^j)||\mathcal{N}(0, I))\end{aligned} \tag{5}$$

where D_{KL} is the Kullback-Leibler divergence and β is a hyperparameter which balances both objectives. Then we sample mini-batches of size N to compute the total loss where $\mathcal{L}_{VAE}(\theta, \phi) = \sum_{i=1}^{N} \mathcal{L}_{VAE}(\theta, \phi; \sigma^i)$. The fine-tunning process of the VAE is interleaved with training both $\hat{r}_\psi$ and π_ω, i.e., we obtain new segments by sampling π_ω, then we train the VAE on the full dataset $\mathcal{D}_\sigma$ and elicit feedback from humans to train a new estimation of $\hat{r}_\psi$.

Algorithm 1 SEQUEL

Require: Queries per session M, Interpolation increment δ, classification boundary τ, unlabeled batch ratio μ, loss weight λ

for epoch $= 1, 2, \ldots$ **do**
  /* Train policy π_ω */
  Collect trajectories $\mathcal{B}^{\text{temp}}$ by sampling π_ω
  $\mathcal{B}^{\text{temp}} \leftarrow \{(s_t, a_t, s_{t+1}, r_\psi(s_t, a_t))\}_{j=1}^{B_{\text{temp}}}$
  Optimize π_ω via PPO with $\mathcal{B}^{\text{temp}}$
  Sample π_ω to obtain $\mathcal{D}_\sigma^{\text{new}} = \{\sigma^i\}_{i=1}^{D_\sigma}$
  Store new trajectories $\mathcal{D}_\sigma \leftarrow \mathcal{D}_\sigma \cup \mathcal{D}_\sigma^{\text{new}}$
  /* Train enc. $q_\phi(z|\sigma)$ and dec. $p_\theta(\sigma|z)$ */
  Train $q_\phi(z|\sigma)$ and $p_\theta(\sigma|z)$ via Eq.5 with $\mathcal{D}_\sigma$
  /* Obtain Human-feedback */
  Form pairs of segments $\{(\sigma_0, \sigma_1)\}_{i=1}^{M}$ by sampling $(\sigma_0, \sigma_1) \sim \mathcal{D}_\sigma^{\text{new}}$
  Inquire humans for preferences and store them on $\mathcal{D}_l \leftarrow \mathcal{D}_l \cup \{(\sigma_0, \sigma_1, y)\}_{i=1}^{M}$
  /* Train reward function $\hat{r}_\psi$ */
  for each $(\sigma_0, \sigma_1, y) \in \mathcal{D}_l^{\text{new}}$ **do**
    $\zeta \leftarrow 0$, $z_0 \sim q_\phi(z|\sigma_0)$, $z_1 \sim q_\phi(z|\sigma_1)$
    repeat
      $\zeta \leftarrow \zeta + \delta$
      $z^0_{\text{interp}} = (1 - \zeta) \cdot z^0 + \zeta \cdot z^1$
      $z^1_{\text{interp}} = (1 - \zeta) \cdot z^1 + \zeta \cdot z^0$
      $\hat{\sigma}^0_{\text{interp}} \sim p_\theta(\sigma|z^0_{\text{interp}})$
      $\hat{\sigma}^1_{\text{interp}} \sim p_\theta(\sigma|z^1_{\text{interp}})$
      Store interpolated queries $\mathcal{D}_u \leftarrow \mathcal{D}_u \cup (\hat{\sigma}^0_{\text{interp}}, \hat{\sigma}^1_{\text{interp}}, y)$
    until $P_\psi[\hat{\sigma}^0_{\text{interp}} \succ \hat{\sigma}^1_{\text{interp}}] < \tau$
  end for
  Train $\hat{r}_\psi$ via Eq.4 with $\mathcal{D}_l$ and $\mathcal{D}_u$ and hyperparameters λ, μ
end for

B. Query-based latent space interpolation for semi-supervised preference-based RL

A significant limitation of relying solely on pseudo-labeling (see Sec. III) in preference-based RL is that the reduced query sizes from human labelers hinder the ability to develop an effective segment representation. This challenge makes it difficult to distinguish similar segments, and it enforces an overly strict classifier decision boundary when pseudo-labeling. Consequently, this can lead to over-fitting and noisy predictions, issues that SURF partially addresses

through temporal cropping. In SEQUEL, we chose not to focus on temporal cropping, since it only yielded improvements of $\sim$ 4% on average on Cheetah and Walker2d, when combined with our interpolation approach. In order to interpolate and synthesize newer queries from labeled data, we use both the encoder $q_\phi(z|\sigma)$ for interpolation and the decoder $p_\theta(\sigma|z)$ for query augmentation to obtain a more robust reward function $\hat{r}_\psi$. Consider a query (σ^0, σ^1, y), from the posterior latent distribution $q_\phi(z|\sigma)$ we sample both z^0 and z^1, such as $z^0 \sim q_\phi(z|\sigma^0)$ and $z^1 \sim q_\phi(z|\sigma^1)$ respectively. Following the reparameterization trick [68], we take the output of the encoder as a mean μ and perform element-wise multiplication with a standard normal such as $z^0 = \mu^0 + \epsilon^0 \cdot \sqrt{\Sigma^0}$ where Σ^0 is sampled from $\mathcal{N}(0,1)$ and has the same size as μ^0.

We synthesize new queries by interpolating near the decision boundary of both σ^0 and σ^1, more concretely, we complement pseudo-labeling by allowing a much larger τ by inferring on segments which are equally spaced in the latent space. Given z_0 and z_1, we can interpolate such that:

$$z^0_{\text{interp}} = (1-\zeta)\cdot z^0 + \zeta \cdot z^1, \tag{6}$$

$$z^1_{\text{interp}} = (1-\zeta)\cdot z^1 + \zeta \cdot z^0 \tag{7}$$

where ζ is an interpolation parameter. Then, for each interpolation, we decode it back to the original space such that $\hat{\sigma}^0_{\text{interp}} \sim p_\theta(\sigma|z^0_{\text{interp}})$ and $\hat{\sigma}^1_{\text{interp}} \sim p_\theta(\sigma|z^1_{\text{interp}})$. The VAE is capable of learning continuous and smooth latent space representations, where similar data points are positioned closely together. However, the resulting structure may not exhibit linear properties. To safeguard against noisy latent interpolations, we vary ζ in small increments δ until the decoded representations form a query which is under the decision boundary τ, i.e $P_\psi[\hat{\sigma}^0_{\text{interp}} \succ \hat{\sigma}^1_{\text{interp}}] < \tau, \tau \in [0,1]$. This approach allows for a nuanced handling of the environment-dependent variations in the latent space structure (refer to Sec.V-C). Finally, we pseudo-label by maintaining the initial label of the original segments, and store the generated queries $(\hat{\sigma}^0_{\text{interp}}, \hat{\sigma}^1_{\text{interp}}, y)$ into the unlabeled dataset $\mathcal{D}_u$. Following [48], [53] we sample a larger labeled minibatch by a factor of μ and optimize $\hat{r}_\psi$ following Eq. 4.

V. Experimental Evaluation

We proceed to investigate the effectiveness of SEQUEL: (1) We assess the performance of SEQUEL against a baseline and semi-supervised techniques introduced by SURF [48]; then (2) we elicit actual human feedback to access the real world applicability of our approach; and finally (3) we explore the latent space of the queries produced by SEQUEL during the human feedback collection.

A. Synthetic Benchmark Performance

Environments. We benchmark SEQUEL on a variety of tasks: four environments, Hopper, Walker, Cheetah and Swimmer [69] which were used for testing in the original preference-based RL baseline [16]; Reacher from the DeepMind Control Suite [70]; and four complex robotic manipulation task from Meta-world [71]: DoorClose, DrawerClose, WindowClose, 3DReacher.

Implementation Details. We compare our approach with two algorithms: (1) a baseline without data augmentation following Christiano et al. [16] and refer to it as **PPO**; (2) we implement pseudo-labeling and temporal cropping introduced in SURF but use PPO [67] instead of SAC [72], which we refer to as **SURF-PPO**. We keep the same hyperparameters and implementation design across the experiments for the policy π_ω trained following PPO, the reward function $\hat{r}_\psi$, the query selection strategy, and the amount of feedback collected. We optimize all networks using ADAM [73]. To select queries, we use a strategy based on the variance across ensemble members, i.e. ensemble disagreement, of the estimated reward model $\hat{r}_\psi$. For the semi-supervised learning parameters, we use the same loss weight $\zeta = 1$ and unlabeled batch ratio $\mu = 4$ for both SURF-PPO and SEQUEL, and use their parameters for both the and threshold parameter $\tau = 0.99$. Where in SURF their ablation study finds lower values of τ yields lower learning performance, our approach is able to set lower values of τ. We set $\tau = 0.9$ and make interpolation increments of $\delta = 0.025$.

Results Discussion. To test the query efficiency of SEQUEL, we perform extensive testing on a variety of environments with different dynamics and settings to test if our interpolation hypothesis of considering latent representations of segments holds. The full results of these experiments can be seen in Figure 2. At first glance, we see both semi-supervised approaches to improve on the baseline, indicating these methods help reward function training and consequently policy learning. For all environments, we observe the asymptotic performance of SEQUEL to be either on-par or above SURF-PPO and baseline. The largest differences in performance stem from Walker and DoorClose. Across environments, we also observe a common trend of SEQUEL accelerating learning earlier than other methods. While when considering pseudo-labeling, noisy reward estimations at the beginning of training will not yield high levels of confidence τ of choosing a segment over another $P_\psi\ [\sigma^0 \succ \sigma^1]$, our method is able to immediately create synthetic queries near the original boundary improving reward function robustness. To probe the effects of SEQUEL on achieving a more regularized reward function, and to distinguish the role of the latent queries created by SEQUEL in enhancing segment representations, we examine the accuracy of the reward function. In Figure 3, we depict the reward model accuracy throughout all phases of training for the three most complex tasks. During each feedback session, which occurs every 20K timesteps, we sample a thousand queries from the policy π_ω and plot the reward accuracies for each condition. Our analysis confirms that SEQUEL consistently performs at a higher level when compared to the other conditions.

B. Eliciting Human-feedback with SEQUEL

To access the effectiveness of SEQUEL with actual human feedback, we use SocialNav [74], a social navigation task environment developed in Unity [75] implemented by the

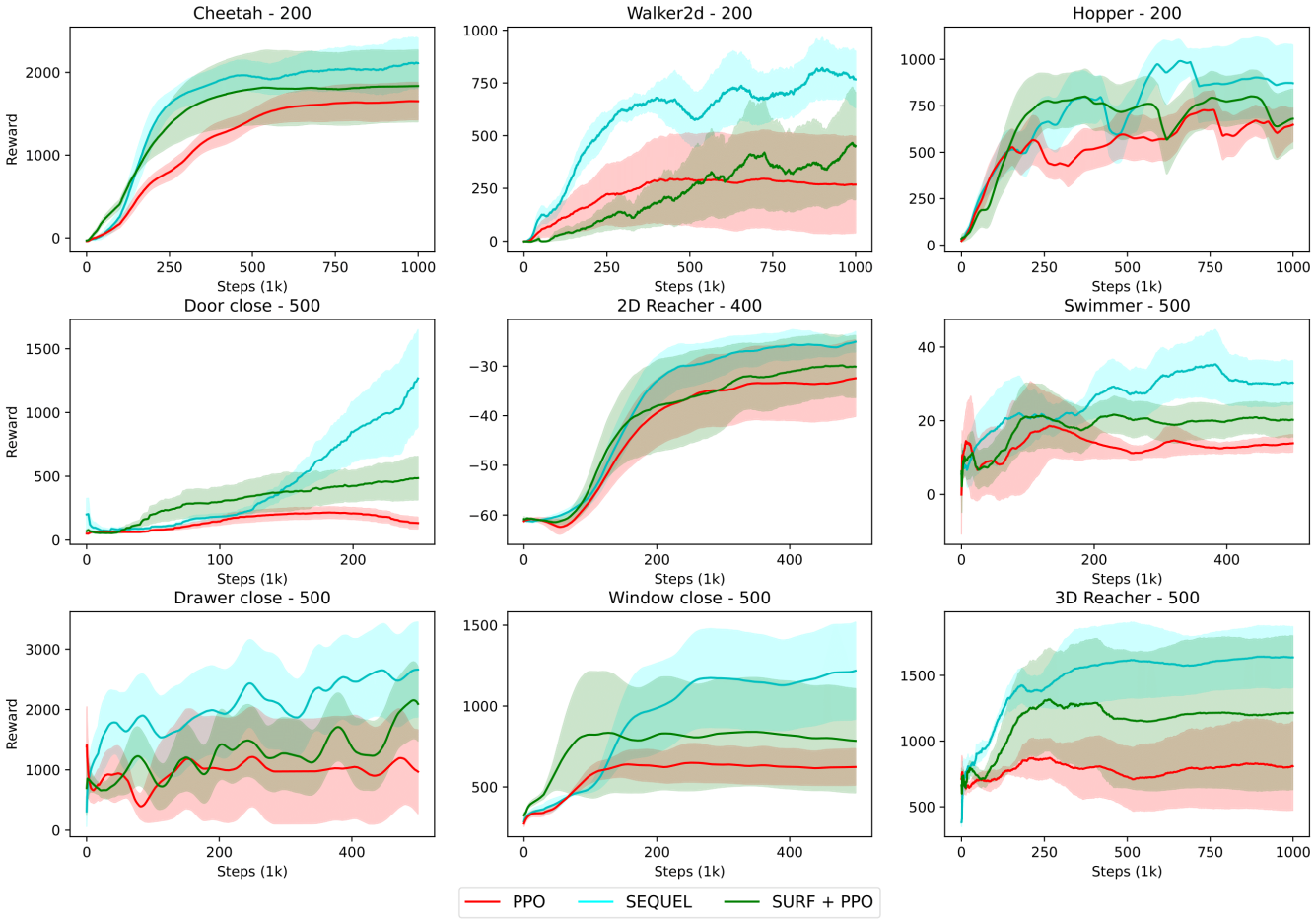

Fig. 2: Learning curves for the different environments. The titles indicate the environment used followed by the total queries collected from a synthetic oracle. The solid lines represent the mean and shaded areas of the standard error.

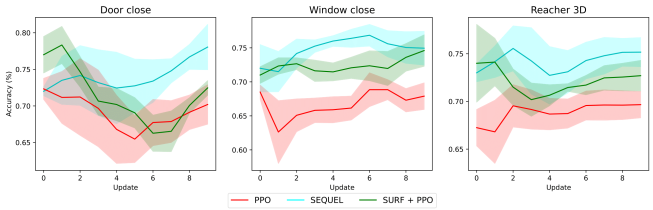

Fig. 3: Curves showing the accuracy on queries sampled after each training loop of the reward function, showing the generalization to unseen queries.

	SocialNav		
	SEQUEL	SURF+PPO	PPO
Rwd.	0.813	0.719	0.717
Acc.	67.52%	64.94%	62.11%

TABLE I: Reward and accuracy observed in SocialNav using real human feedback.

authors to better visualize the synthesized queries produced by SEQUEL in Sec. V-C. As noted by the findings of [23] when performing preference-based RL with real users, maximizing information when picking queries by ensemble disagreement, can produce queries difficult to answer for being hard to distinguish. We followed their suggestion by introducing a "skip" option and increased the number of randomly sampled uniform queries to 30%. We start from a pre-trained policy π_ω trained on the default reward of the environment for 5×10^5 timesteps to acquire a pool of segments which are more suitable for preference elicitation, saving human queries in the process. We conduct a comparative analysis between baseline, SURF-PPO, and SEQUEL. We gathered feedback by requesting humans to favor segments in which the robot collects stars safely as it traverses through the corridor. To evaluate if a reward function generated by SEQUEL is superior in forecasting the selection of segments, thus offering better generalization, we reserve 20% of the collected queries for accuracy testing. We proceed with training under all conditions and the reward results are presented in Table I. We notice that SEQUEL demonstrates the highest final reward, closely trailed by SURF-PPO and the baseline, substantiating the findings made in Section V-A. Furthermore, we also confirm a marginally increased accuracy, backing our hypothesis of a more robust reward function.

C. *Latent space of queries produced by SEQUEL*

We delve deeper into the unlabelled queries synthesized by SEQUEL. A downside to relying exclusively on reward estimations for pseudo-labeling of unlabelled queries is that in the early stages of training, these estimations can be fairly

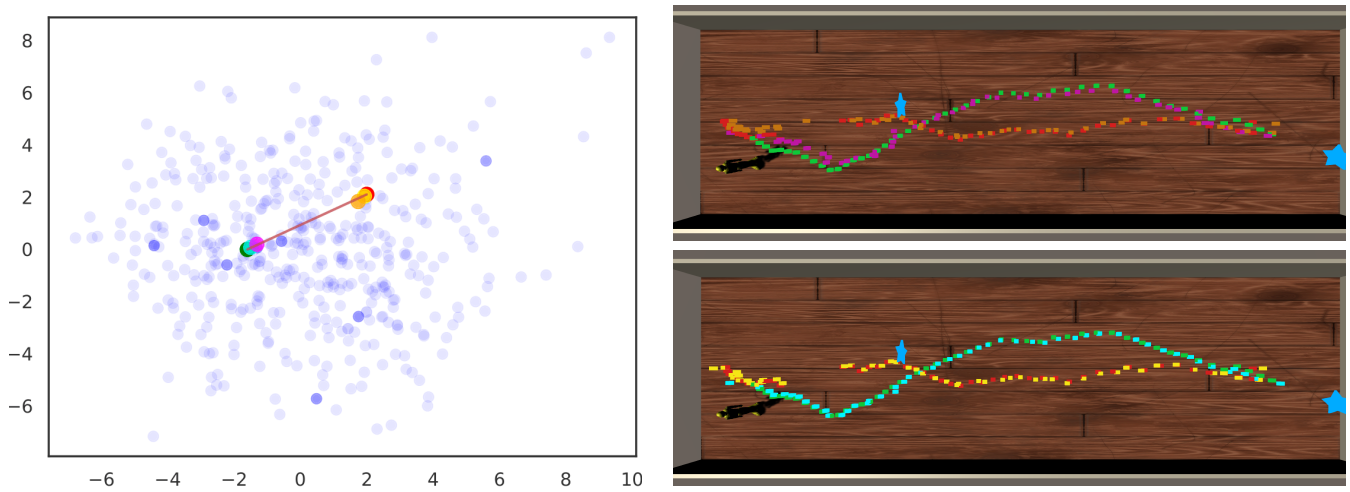

Fig. 4: On the left side, the TSNE projection of query pairs. On the right, the reconstructed query segments. The original preferred segment is highlighted in green, and the non-preferred one in red. The segments colored in cyan and magenta represent interpolated and decoded segments that are $\zeta = 0.025$ (2.5%) and $\zeta = 0.05$ (5%) respectively, away from the preferred segment. Similarly, yellow and orange depict segments that are 2.5% and 5% distant from the non-preferred segment.

uncertain due to data scarcity. Consequently, the network's intermediate layers may not have a precise representation of the input, which can hinder the differentiation between similar segments. To mitigate this issue, SURF employs a high threshold parameter τ of 0.99. SEQUEL offers a solution to this challenge by utilizing the latent space of segments, which can be acquired from a dataset of segments $\mathcal{D}_\sigma$. This dataset is gathered through a sequential sampling of a policy, eliminating the necessity for further human intervention. We can capitalize on our prior knowledge of the distribution of segments to interpolate and improve the generalizability of our reward function. This is realized by incorporating correlated variations, enhancing the speed of developing an internal representation of the reward function. We use our SocialNav environment to intuitively illustrate SEQUEL's strength. We start by pre-training π_ω with the default reward function of the environment. We then follow the SEQUEL Algorithm 1 and proceed to acquire $\mathcal{D}_\sigma$, and both the VAE encoder $q_\phi(z|\sigma)$ and decoder $p_\theta(\sigma|z)$. In Figure 4, the latent space of labeled and synthesized queries is explored by illustrating trajectories, using only the position variables (x, y) extracted from the complete state. Consider an example of a query presented to humans, characterized by encoded representations z_0 and z_1, and graphically represent it in green and red respectively, in Figure 4. We interpolate on both query elements with interpolation parameter $\delta = 0.025$. Observe that the interpolated points do not align perfectly, as this is a TSNE projection originating from a space with more than two dimensions. Finally, we decode the interpolated points back to the original input space following $p_\theta(\sigma|z)$. By inspection, we observe all the reconstructed interpolated segments to be highly correlated to the original ones, supporting our hypothesis of being able to extend the preference of the labelled query provided by the human to the interpolated ones. We observe a gradient to be expected from $\zeta = 0.025$ to $\zeta = 0.05$ on the synthesized segments. However, at $\zeta = 0.05$ we get a correspondent confidence level of $P_\psi[\hat{\sigma}^0_{\text{interp}} \succ \hat{\sigma}^1_{\text{interp}}] \approx 0.96$ which would be discarded if we considered a cut-off of $\tau = 0.99$ by sampling using pseudo-labeling.

VI. Discussion

Future Research. SEQUEL seamlessly integrates with any preference-based RL framework, offering improved reward function generalization without the need for additional human-labeled samples. While SEQUEL poses as a stepping stone in improving preference-based RL, it requires some assumptions and leaves future research open.

Quality of the latent representation. If feature vectors are poorly represented in the latent space, it can result in incorrect interpolation of latent segments, causing our approach to revert to pseudo-labeling. This highlights an intriguing area of research: the potential to train generative models in a manner that enables the latent space to reflect not only the policy space but also to align with the perspectives and annotations of human evaluators.

Active Learning. While we chose ensemble disagreement as a query selection strategy, we could potentially leverage the latent representation of segments to create queries that maximize reward information while not being too similar for humans to be able to distinguish.

Conclusion. We presented SEQUEL, a novel semi-supervised learning approach for preference-based RL which advances the field. As shown by our experiments, our approach is on-par or above performance when compared to the state-of-the-art, on a variety of complex scenarios relevant to robotics. We also presented visual evidence to better understand the effectiveness of our method, and hope to motivate other works in exploring the query space produced by labeled queries to better reason through human intents. Finally, we tested SEQUEL with human feedback and corroborated our previous findings on the synthetic benchmarks.

REFERENCES

[1] J. MacGlashan, M. K. Ho, R. Loftin, B. Peng, G. Wang, D. L. Roberts, M. E. Taylor, and M. L. Littman, "Interactive learning from policy-dependent human feedback," in *Int. Conf. on Machine Learning*. PMLR, 2017, pp. 2285–2294.

[2] R. Zhang, D. Bansal, Y. Hao, A. Hiranaka, J. Gao, C. Wang, R. Martín-Martín, L. Fei-Fei, and J. Wu, "A dual representation framework for robot learning with human guidance," in *6th Annual Conference on Robot Learning*, 2022. [Online]. Available: https://openreview.net/forum?id=H6rr_CGzV9y

[3] D. Shah, A. Bhorkar, H. Leen, I. Kostrikov, N. Rhinehart, and S. Levine, "Offline reinforcement learning for visual navigation," in *6th Annual Conference on Robot Learning*, 2022. [Online]. Available: https://openreview.net/forum?id=uhIfIEIiWm_

[4] A. Xie, A. Singh, S. Levine, and C. Finn, "Few-shot goal inference for visuomotor learning and planning," in *Conference on Robot Learning*. PMLR, 2018, pp. 40–52.

[5] D. Marta, S. Holk, C. Pek, J. Tumova, and I. Leite, "Variquery: Vae segment-based active learning for query selection in preference-based reinforcement learning," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 7878–7885.

[6] T. Sumers, R. D. Hawkins, M. K. Ho, T. L. Griffiths, and D. Hadfield-Menell, "How to talk so AI will learn: Instructions, descriptions, and autonomy," 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=ZLsZmNe1RDb

[7] S. Holk, D. Marta, and I. Leite, "Predilect: Preferences delineated with zero-shot language-based reasoning in reinforcement learning," *arXiv preprint arXiv:2402.15420*, 2024.

[8] D. Brown, W. Goo, P. Nagarajan, and S. Niekum, "Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations," in *International conference on machine learning*. PMLR, 2019, pp. 783–792.

[9] D. S. Brown, W. Goo, and S. Niekum, "Better-than-demonstrator imitation learning via automatically-ranked demonstrations," in *Conference on robot learning*. PMLR, 2020, pp. 330–359.

[10] T. Fitzgerald, P. Koppol, P. Callaghan, R. Q. J. H. Wong, R. Simmons, O. Kroemer, and H. Admoni, "INQUIRE: INteractive querying for user-aware informative REasoning," in *6th Annual Conference on Robot Learning*, 2022. [Online]. Available: https://openreview.net/forum?id=3CQ3Vt0v99

[11] A. Taranovic, A. G. Kupcsik, N. Freymuth, and G. Neumann, "Adversarial imitation learning with preferences," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=bhfp5GlDtGe

[12] E. Bıyık, D. P. Losey, M. Palan, N. C. Landolfi, G. Shevchuk, and D. Sadigh, "Learning reward functions from diverse sources of human feedback: Optimally integrating demonstrations and preferences," *The International Journal of Robotics Research*, vol. 41, no. 1, pp. 45–67, 2022.

[13] 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.

[14] D. Marta, S. Holk, C. Pek, J. Tumova, and I. Leite, "Aligning human preferences with baseline objectives in reinforcement learning," in *International Conference on Robotics and Automation*, 2023.

[15] 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.

[16] 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.

[17] D. S. Brown and S. Niekum, "Deep bayesian reward learning from preferences," *arXiv preprint arXiv:1912.04472*, 2019.

[18] V. Myers, E. Biyik, N. Anari, and D. Sadigh, "Learning multimodal rewards from rankings," in *Conference on Robot Learning*. PMLR, 2022, pp. 342–352.

[19] 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 *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2022, pp. 147–156.

[20] 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.

[21] 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.

[22] K. Lee, L. Smith, A. Dragan, and P. Abbeel, "B-pref: Benchmarking preference-based reinforcement learning," *arXiv preprint arXiv:2111.03026*, 2021.

[23] 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.

[24] 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.

[25] 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.

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

[27] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman, *Causation, prediction, and search*. MIT press, 2000.

[28] F. Eberhardt, "Introduction to the foundations of causal discovery," *International Journal of Data Science and Analytics*, vol. 3, no. 2, pp. 81–91, 2017.

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

[30] 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.

[31] J. Tien, J. Z.-Y. He, Z. Erickson, A. Dragan, and D. S. Brown, "Causal confusion and reward misidentification in preference-based reward learning," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=R0Xxvr_X3ZA

[32] Y. Chen, M. Mancini, X. Zhu, and Z. Akata, "Semi-supervised and unsupervised deep visual learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[33] D. Ha and J. Schmidhuber, "Recurrent world models facilitate policy evolution," *Advances in neural information processing systems*, vol. 31, 2018.

[34] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=S1lOTC4tDS

[35] R. Sekar, O. Rybkin, K. Daniilidis, P. Abbeel, D. Hafner, and D. Pathak, "Planning to explore via self-supervised world models," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8583–8592.

[36] D. Berthelot, C. Raffel, A. Roy, and I. Goodfellow, "Understanding and improving interpolation in autoencoders via an adversarial regularizer," *arXiv preprint arXiv:1807.07543*, 2018.

[37] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," *arXiv preprint arXiv:1511.06349*, 2015.

[38] D. Ha and D. Eck, "A neural representation of sketch drawings," *arXiv preprint arXiv:1704.03477*, 2017.

[39] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville, "Adversarially learned inference," *arXiv preprint arXiv:1606.00704*, 2016.

[40] R. Akrour, M. Schoenauer, and M. Sebag, "Preference-based policy learning," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011. Proceedings, Part I 11*. Springer, 2011, pp. 12–27.

[41] ——, "April: Active preference learning-based reinforcement learning," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*. Springer, 2012, pp. 116–131.

[42] R. Akrour, M. Schoenauer, M. Sebag, and J.-C. Souplet, "Programming by feedback," in *International Conference on Machine Learning*. JMLR. org, 2014, pp. 1503–1511.

[43] M. Cakmak, S. S. Srinivasa, M. K. Lee, J. Forlizzi, and S. Kiesler, "Human preferences for robot-human hand-over configurations," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2011, pp. 1986–1993.

[44] D. Sadigh, A. D. Dragan, S. Sastry, and S. A. Seshia, "Active preference-based learning of reward functions." in *Robotics: Science and Systems*, 2017.

[45] E. Bıyık, N. Huynh, M. J. Kochenderfer, and D. Sadigh, "Active preference-based gaussian process regression for reward learning," in *Robotics: Science and Systems (RSS)*, 2020.

[46] 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.

[47] 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

[48] 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

[49] 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

[50] R. Liu, F. Bai, Y. Du, and Y. Yang, "Meta-reward-net: Implicitly differentiable reward learning for preference-based reinforcement learning," 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=OZKBReUF-wX

[51] D. Shin, A. Dragan, and D. S. Brown, "Benchmarks and algorithms for offline preference-based reward learning," *Transactions on Machine Learning Research*, 2022.

[52] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," *Advances in neural information processing systems*, vol. 17, 2004.

[53] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020.

[54] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, 2013, p. 896.

[55] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, "Semi-supervised learning with deep generative models," *Advances in neural information processing systems*, vol. 27, 2014.

[56] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," *Advances in neural information processing systems*, vol. 33, pp. 6256–6268, 2020.

[57] D. Yarats, I. Kostrikov, and R. Fergus, "Image augmentation is all you need: Regularizing deep reinforcement learning from pixels," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=GY6-6sTvGaf

[58] M. Laskin, K. Lee, A. Stooke, L. Pinto, P. Abbeel, and A. Srinivas, "Reinforcement learning with augmented data," *Advances in neural information processing systems*, vol. 33, pp. 19 884–19 895, 2020.

[59] M. Laskin, A. Srinivas, and P. Abbeel, "Curl: Contrastive unsupervised representations for reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5639–5650.

[60] Y. Wu, M. Mozifian, and F. Shkurti, "Shaping rewards for reinforcement learning with imperfect demonstrations using generative models," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 6628–6634.

[61] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He, "Data distillation: Towards omni-supervised learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4119–4128.

[62] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves imagenet classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 687–10 698.

[63] Y. Chen, X. Zhu, W. Li, and S. Gong, "Semi-supervised learning under class distribution mismatch," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 3569–3576.

[64] 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.

[65] 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.

[66] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, "Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=HklkeR4KPB

[67] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[68] D. P. Kingma, M. Welling *et al.*, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[69] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.

[70] S. Tunyasuvunakool, A. Muldal, Y. Doron, S. Liu, S. Bohez, J. Merel, T. Erez, T. Lillicrap, N. Heess, and Y. Tassa, "dm_control: Software and tasks for continuous control," *Software Impacts*, vol. 6, p. 100022, 2020.

[71] 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.

[72] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proceedings of the 35th International Conference on machine learning (ICML-18)*, 2018, pp. 1861–1870.

[73] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization 3rd int," in *Conf. for Learning Representations, San*, 2014.

[74] 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.

[75] A. Juliani, V.-P. Berges, E. Teng, A. Cohen, J. Harper, C. Elion, C. Goy, Y. Gao, H. Henry, M. Mattar, and D. Lange, "Unity: A general platform for intelligent agents," 2020.