P3L: Learning Pluralistically Aligned Policies from Heterogeneous Crowd Feedback

Akansha Kalra[1], **Ellen Novoseller**[2], **Vernon J. Lawhern**[2], **Daniel S. Brown**[1]

[1]**Kahlert School of Computing, University of Utah, Salt Lake City, USA**
[2]**U.S. Army DEVCOM Army Research Laboratory, Aberdeen Proving Ground, MD, USA**

Abstract

Learning from diverse human feedback and personalizing AI models to each user is a hard challenge. Most research on reinforcement learning from human feedback (RLHF) either assumes all preference annotators share the same underlying reward distribution and treat divergent preferences as noise or assume prior knowledge about annotator identity to learn diverse behaviors conditioned on the specific annotator ID. We develop a framework that delivers personalized pluralistic alignment while learning from diverse, potentially conflicting preferences without requiring prior knowledge about annotator identity. Our framework learns a shared basis of preference directions from heterogeneous crowd feedback by combining trajectory representation learning with a PCA decomposition of preference signals. These basis eigenvectors paired with adaptation queries define the personalized reward which is then used to optimize a downstream policy. By leveraging heterogeneous preferences, our approach produces personal alignment that generalizes beyond the offline dataset to out-of-distribution new users and attains competitive RL and test performance with reduced computational costs, and is flexible to varying adaptive query budgets.

1 Introduction

As we witness growing adoption of AI systems and robotic models for downstream applications, aligning these models to an end-user's individual preference is seen as a major challenge (Ji et al., 2023; Casper et al., 2023). One of the most popular approaches for aligning such models is Reinforcement Learning from Human Feedback (RLHF).

RLHF learns a reward function from human preferences and then optimizes an RL policy under the learned reward function (Christiano et al., 2017; Ouyang et al., 2022; Kaufmann et al., 2024). While this approach has been highly successful, it has been shown to be susceptible to over optimization (Anwar et al., 2024) and reward misidentification (Tien et al., 2023), and easily misaligned across diverse users in a crowd (Siththaranjan et al., 2024; Poddar et al., 2024; Chhan et al., 2024). Recent work has tried to solve this challenge by either learning multiple reward functions from divergent crowd-sourced preference (Myers et al., 2022; Shenfeld et al., 2025) or by assuming knowledge about the annotators who provide the preference labels, such as the annotator's id (Poddar et al., 2024) or assuming that preference heterogeneity arises from latent contextual factors that must be modeled explicitly during reward learning, rather than directly enabling policy personalization for new users (Siththaranjan et al., 2024).

This raises a central challenge: *Can we learn from heterogeneous crowd preferences—potentially reflecting conflicting objectives—without annotator identity or prior structural assumptions, while still enabling efficient test-time personalization to new users in embodied control tasks?*

Our key insight is that heterogeneous crowd preferences can be decomposed into a small set of shared preference directions. This reframes pluralistic alignment as estimating user-specific weights

over a shared preference basis representation rather than learning separate reward models for each user. This representation enables personalized reward functions—and corresponding policies—to be constructed for new users at test time without requiring annotator identities, allowing rapid policy personalization from a small number of adaptive queries. Building on the preference vector framework (Luo et al., 2025), which represents human preferences as directions in a latent representation space, we extend this idea to embodied control. We propose a framework that jointly learns trajectory representations and preference directions from heterogeneous crowd feedback. Specifically, we train an end-to-end unsupervised representation encoder that embeds trajectories and use these embeddings to extract a basis of preference eigenvectors capturing variation in user objectives.

Unlike prior work such as Variational Preference Learning (VPL) (Poddar et al., 2024), which assumes access to annotator identities to infer user-specific rewards, our framework does not require annotator information. At the same time, it can incorporate annotator identities when available, enabling it to operate under both privileged and non-privileged settings. Together, these ideas enable a simple and effective framework for pluralistic alignment in embodied control. We summarize our contributions as follows:

- **Learning preference bases from heterogeneous feedback.** We introduce a framework that learns a shared basis of preference directions from heterogeneous, potentially conflicting crowd preferences. Our method leverages an unsupervised trajectory encoder to construct latent trajectory embeddings and extracts preference eigenvectors that capture variation in user objectives without requiring annotator identities.
- **Personalized policy alignment.** We demonstrate that the learned preference basis enables test-time construction of personalized reward functions and policies, allowing alignment to individual users across multiple objectives.
- **Empirical evaluation.** Across multi-objective control environments, our method achieves strong alignment to both in-distribution and out-of-distribution users. Compared to prior work such as VPL (Poddar et al., 2024), it achieves higher test-time preference accuracy and competitive RL performance while significantly reducing training cost and allowing flexible adaptation to varying numbers of test-time queries.

2 Related Work

Reinforcement Learning from Human Feedback (RLHF) addresses the challenge of specifying reward functions by instead learning them from human feedback. Human feedback can take several forms, including pairwise preferences (Christiano et al., 2017; Poddar et al., 2024), rankings (Brown et al., 2019; Myers et al., 2022), and corrective feedback (Bajcsy et al., 2017; Mehta & Losey, 2024). Among these approaches, learning reward functions from pairwise preferences has become one of the most widely adopted paradigms. Most RLHF methods assume that preference data originates from a single underlying reward function, effectively treating disagreement among annotators as noise. Some work attempts to account for imperfect human feedback. For example, Lee et al. (2021) introduces stochastic preference models to capture noisy or inconsistent human judgments. Nevertheless, these approaches still assume that annotators ultimately share a common objective. By contrast, real-world preference data is often collected from diverse users whose objectives may fundamentally differ or even conflict. This motivates studying methods that can learn from heterogeneous crowd feedback while still enabling alignment to individual users.

Personalized RLHF and Pluralistic Alignment Recent work has begun to investigate how to model heterogeneous human preferences and personalize AI behavior to different users. Several approaches have explored preference diversity primarily in the context of large language models (LLMs) (Luo et al., 2025; Shenfeld et al., 2025; Siththaranjan et al., 2024). These methods typically model preference variation through latent variables or preference decompositions, but focus on reward modeling rather than policy personalization in embodied control domains. Distributional Preference Learning (DPL) (Siththaranjan et al., 2024) studies preference learning in the presence

of *hidden context*, where annotator identity or other contextual factors influencing preferences are not observed by the reward model. While this approach explicitly models preference heterogeneity through latent contextual factors during reward learning, its primary goal is to characterize and detect hidden context rather than to construct personalized policies for new users. Only a small number of works study pluralistic alignment in embodied control tasks. Variational Preference Learning (VPL) (Poddar et al., 2024) models user-specific reward functions by introducing a latent variable conditioned on annotator identity, requiring access to annotator IDs during training. Alternatively, Bahlous-Boldi et al. (2025) learns a Pareto set of reward hypotheses using lexicase selection to represent trade-offs among multiple objectives. Despite these advances, existing approaches either require annotator identities, explicitly model latent contexts during reward learning, or produce sets of reward hypotheses rather than directly enabling personalization for new users. In contrast, our framework learns a shared basis of preference directions from heterogeneous crowd feedback without requiring annotator identities. This representation allows personalized reward functions—and corresponding policies—to be constructed for previously unseen users at test time from a small number of preference queries.

3 Preliminaries

Reinforcement Learning from Human Feedback (RLHF). Given an environment, we model it as a finite-horizon Markov Decision Process (MDP) (Puterman, 1994) with state space S, action space A, horizon T, and reward function $r : S \times A \to \mathbb{R}$. The true reward function r is unknown and must be inferred from human preferences over trajectories. A trajectory τ is defined as a sequence of state–action pairs over the horizon T, expressed as $\tau = (s_1, a_1, s_2, a_2, \ldots, s_{T-1}, a_{T-1})$. Given a dataset D of annotated preferences over trajectory pairs of the form $\tau_i \prec \tau_j$, traditional RLHF first learns a reward function using a maximum likelihood objective based on the Bradley–Terry–Luce (BTL) model (Bradley & Terry, 1952; Luce et al., 1959), written as $p(\tau_i \succ \tau_j) = \frac{e^{r_\theta(\tau_i)}}{e^{r_\theta(\tau_i)}+e^{r_\theta(\tau_j)}}$, where θ denotes the parameters of the learned reward model. A policy π is then optimized under the learned reward $\pi^\star_{r_\theta} = \arg\max_\pi \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t r_\theta(s_t, a_t)\right]$. This formulation learns a *single* reward function that aggregates all preferences. Consequently, divergent preference signals are typically treated as noise rather than as evidence of genuinely different user objectives.

Variational Preference Learning (VPL). Prior work by Poddar et al. (2024) addresses learning from crowd preferences by introducing a latent preference variable z associated with each annotator. The reward model is then conditioned on this latent variable via a contextual BTL model $p(\tau_i \succ \tau_j \mid z) = \frac{e^{r_\theta(\tau_i,z)}}{e^{r_\theta(\tau_i,z)}+e^{r_\theta(\tau_j,z)}}$. However, this formulation assumes access to annotator identities during training in order to infer the latent preference variable. Another limitation of VPL is that it requires a fixed number of preference queries for adaptation to a new user, referred to as the *context set* (Poddar et al., 2024). This constraint arises from the VAE-based latent preference inference architecture, whose encoder requires a fixed-size input. In practice, this requirement increases labeling cost and precludes test-time learning from variable numbers of adaptation queries.

Decomposed Reward Models (DRM) Luo et al. (2025) reformulate preference learning by decomposing rewards into multiple orthogonal preference directions in representation space rather than learning a single scalar reward. However, their analysis focuses primarily on reward prediction accuracy and does not consider optimizing policies under personalized reward functions. High accuracy on held-out preference queries does not necessarily translate to behaviorally aligned policies, as widely documented in prior RLHF work (Christiano et al., 2017; Tien et al., 2023). Furthermore, DRM relies on powerful pretrained embedding models trained on internet-scale data. In practice, they employ reward models derived from large language models such as Gemma-2B-RM and Llama3-8B-RM, which have undergone multiple stages of training including pretraining and supervised instruction fine-tuning before preference alignment (Ziegler et al., 2020; Ouyang et al., 2022). These pretrained models provide rich representations capable of capturing diverse human

preference factors. In contrast, many embodied control domains lack such pretrained representation models. A key challenge is therefore learning meaningful trajectory representations directly from limited task-specific data. Our work investigates how to learn such representations and preference decompositions in settings where large pretrained embedding models are unavailable.

3.1 Problem Formulation

Learning from diverse crowdsourced feedback presents a more challenging setting than standard RLHF. Vanilla RLHF aggregates preferences from multiple annotators into a single reward function, implicitly assuming that all preference labels arise from a shared underlying objective. However, in many real-world settings different users may have fundamentally different goals or behavioral preferences. Aligning an AI system to the preferences of a specific user at test time in the presence of heterogeneous crowd feedback is known as *pluralistic alignment* (Sorensen et al., 2024).

Pluralistic Alignment with Heterogeneous Users Consider a population of N annotators, where each user u_i has an unknown ground-truth reward function r_i. For each preference query m, an unobserved annotator index $z^m \in \{1, \ldots, N\}$ generates a label $y^m \in \{0, 1\}$ over a trajectory pair (τ_A^m, τ_B^m) according to that annotator's latent reward function r_{z^m}, where $y^m = 1$ indicates that τ_A^m is preferred over τ_B^m. Because users may have different objectives, conflicting labels for the same pair of trajectories can naturally occur, e.g.,$(\tau_A, \tau_B, y = 0)$ and $(\tau_A, \tau_B, y = 1)$, when the two labels are produced by different annotators. We are given an offline dataset of crowd preferences $D = \{(\tau_A^m, \tau_B^m, y^m)\}_{m=1}^M$ while the annotator identities $\{z^m\}_{m=1}^M$ may or may not be available.

Learning Objective Our goal is to learn from this heterogeneous preference dataset and construct policies that can be personalized to individual users. In the primary setting, we do not assume access to annotator identities during training. However, if annotator information is available, the framework should be able to leverage it to improve representation learning and alignment.

Under this setting, we aim to address the following questions: 1) **Can we infer personalized reward functions $\hat{r}_i$ for individual users from heterogeneous preference data without requiring annotator identities?** 2) **Can we optimize policies $\hat{\pi}_i$ under these inferred rewards to produce behaviors aligned with diverse user preferences?** 3)**Can such personalization generalize to both in-distribution users and previously unseen users relative to the offline preference dataset?**

This formulation defines the task of learning representations of heterogeneous preferences that enable efficient test-time personalization of reward functions and policies for new users.

4 Preference-PCA Pluralistic Learning

We call our proposed approach Preference-PCA Pluralistic Learning (P3L). The overall pipeline of our approach is shown in Fig 1 and is inspired by the Decomposed Reward Model (Luo et al., 2025). We consider the joint problem of pre-training an expressive embedding extractor and leveraging it to learn preference vectors from diverse heterogeneous crowd preferences in control domains to provide personalized alignment to test time users on both test time held out set as well as a personalized policy optimized under the personalized derived reward.

4.1 Pretraining-Embedding Extractor/Representation Encoder

To learn a latent embedding of a trajectory τ, we employ the Masked Decision Prediction (MaskedDP) (Liu et al., 2022) as our representation encoder ϕ. The encoder in MaskedDP first embeds states and actions as separate linear layers, adds positional embeddings followed by random masking and self-attention. MaskedDP encoder, in its original form/architecture, outputs $2T$ tokens for a trajectory τ of length T. We obtain a single embedding for a given trajectory τ using the frozen MaskedDP encoder in followings 2 ways : **Privileged Traj-MaskedDP** and **Non-Privileged CLS-MaskedDP**.

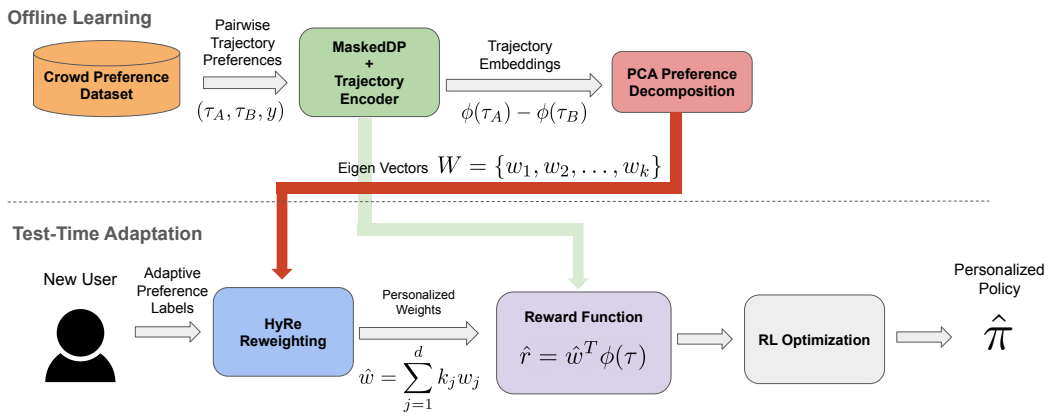


Figure 1: **Preference-PCA Pluralistic Learning (P3L)**

Privileged Traj-MaskedDP In this setting, we assume that annotator id is known. This is the exact same assumption made in VPL by Poddar et al. (2024). We term this approach as **Privileged** because of this assumption about access to privileged information about annotator identity.

We leverage prior knowledge about annotator-identity on preference labels to learn a linear summarization of the trajectory. Formally, given a preference dataset D=$\{(\tau_b, \tau_c), (\tau_d, \tau_e), ...(\tau_o, \tau_m)\}$ of size M with knowledge that annotator A_1 prefers $\tau_b \succ \tau_c$ and $\tau_d \succ \tau_e$, we sample τ_b and τ_d from A_1's set to serve as the anchor and positive trajectories respectively while for another annotator A_2 given that $\tau_o \succ \tau_m$, then we can use τ_o as the negative trajectory.Their latent embeddings generated by frozen MaskedDP encoder, are fed into linear layer trained using contrastive learning via triplet loss. Mathematically, the triplet loss can be formalized as

$$\mathcal{L} = \max\left(0,\ \|\phi(\tau^{A_1}_{anchor}) - \phi(\tau^{A_1}_{pos})\|_2^2 - \|\phi(\tau^{A_1}_{anchor}) - \phi(\tau^{A_2}_{neg})\|_2^2 + \alpha\right) \tag{1}$$

where α is the triplet margin. Note, we refer to this approach later on as only **Traj-MaskedDP**.

Non-Privileged CLSMaskedDP In this setting, we assume no knowledge about annotator or their id- we term this approach as **Non-Privileged** as we have no access to the privileged information about annotator. We learn the singular trajectory embedding in BERT-style, by feeding the outputs of frozen MaskedDP encoder through the self-attention module, thereby learning a CLS token in a completely end-to-end unsupervised manner. We learn the summary token by employing different masking ratios in MaskedDP in the triplet loss. Given two trajectories τ_1, τ_2 from the preference dataset,we use ϕ_{τ_1} with mask ratio m_1 for anchor and with another mask ratio m_2 as positive while ϕ_{τ_2} with mask m_1 is used as negative.

$$\mathcal{L} = \max\left(0,\ \|\phi((\tau_1)_{m1}) - \phi((\tau_1)_{m2})\|_2^2 - \|\phi((\tau_1)_{m1}) - \phi((\tau_2)_{m1})\|_2^2 + \alpha\right) \tag{2}$$

This approach is a completely end-to-end unsupervised embedding learning method and we refer to this approach later on as only **CLS-MaskedDP**.

4.2 Learning from Crowd-Sourced Feedback

Given the crowd sourced preference dataset D with preferences over trajectories of form $\tau_i \prec \tau_j$, we can model heterogeneity in human choices as a preference vector via the 2 step process below :

Learning eigenvectors We first apply PCA to set of latent difference embeddings ϕ_{diff} derived from pairwise preference queries in the offline heterogeneous preference dataset to obtain a set

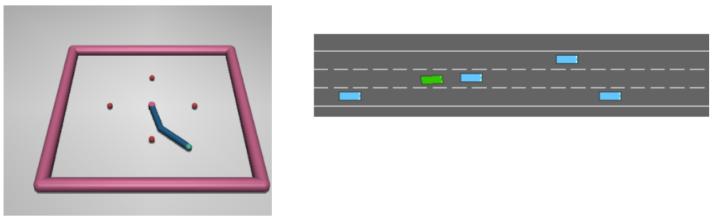


Figure 2: **Environments** MOReacher with 4d reward and MO-Highway with 3d reward.

of principal basis vectors $W = [w_1, w_2, ..., w_d]$, where each orthogonal basis vector captures a principal direction of variation in the observed preference signals.

$$W = \{w_1, w_2, \ldots, w_k\} = \text{PCA}(\phi_{\text{diff}}), \quad \phi_{\text{diff}} = \begin{cases} \phi(\tau_1) - \phi(\tau_2), & \text{if } y = 0 \\ \phi(\tau_2) - \phi(\tau_1), & \text{if } y = 1 \end{cases} \tag{3}$$

Both positive w_i and negative $-w_i$ of the eigenvector are taken to account for fact that PCA eigenvectors are direction agnostic while human preference is inherently not. Thus any human preference can be represented as linear combination of these basis vector $\hat{w} = \sum_{j=1}^{d} k_j w_j$ where $k_1, k_2, ...k_d$ are weights on these individual vectors w_i.

Each w_i can be combined with the embedding $\phi(\tau)$ to give reward of trajectory τ in that dimension and this total reward of trajectory τ can be expressed as $\hat{r} = \hat{w}^T \phi(\tau)$.

Adapting to Test time User At test time, we actively query the test-time user for their preferences to form an adaptive set of queries D_{adapt}. We then leverage these adaptive queries to compute coefficients $k_1, k_2, ...k_d$ on the principal eigenvectors using Hypothesis Reweighing (HyRe) (Lee et al., 2024) to form the complete preference vector of the test-time user. Mathematically, $k_i = \frac{\exp(z_i)}{\sum_d \exp(z_d)}$, where $z_i = \mathbb{E}_{(\tau_A, \tau_B, y) \sim D_{\text{adapt}}} \left[\log \sigma \left(w_i^\top (\phi(\tau_A) - \phi(\tau_B)) \right) \right]$. In this manner, weights k_j shift dynamically according to the adaptive preference dataset to favor the new user by assigning greater weight to directions that capture more of his preference direction i.e. assign higher weight to eigenvectors w_j that minimize the loss while down-weighting those associated with higher loss.

Learning Personalized Policy Once we have the weights k_j on our basis eigenvectors w_j, we can then learn a personalized policy $\hat{\pi}$ for the new user i that maximizes the expected performance under $\hat{r}_i$. Formally $\hat{\pi}_i^\star = \arg\max_{\pi_i} \mathbb{E}_{\tau \sim \pi_i} [\hat{r}_i(\tau)] = \arg\max_{\pi_i} \mathbb{E}_{\tau \sim \pi_i} \left[\hat{w}_i^\top \phi(\tau) \right]$.

5 Experimental Setup

Environments To evaluate pluralistic alignment, we use multi-objective environments as recommended by Sorensen et al. (2024). To learn from crowdsourced feedback with multiple—possibly conflicting—objectives, we consider the following environments Fig 2 from MO-Gymnasium as described by Felten et al. (2023): i) MO-Reacher provides a 4-dimensional reward vector, where each dimension corresponds to the reward for reaching a specific goal. Some goals induce competing preferences; for example, reaching the left goal conflicts directly with reaching the right goal. ii) MO-Highway has a 3-dimensional reward with 1st reward dimension rewarding high-speed driving while 2nd dimension scores on driving in right lane and the 3rd reward dimension represents the collision reward. We use the Fast Highway variant with fixed timestep of 30 timesteps with 3 lanes and 20 vehicles. In each of these environments we ignore the terminal condition and make these tasks fixed time horizon to avoid leakage of information and to prevent termination bias in reward learning from preferences as advised by Freire et al. (2020). Based on the number of underlying objectives in an environment we report the performance across all objectives where each individ-

	Baselines		MaskedDP	
	SingleRM	VPL	Traj-MaskedDP	CLS-MaskedDP
User 1	47.333 ± 2.494	74.000 ± 18.062	96.000 ± 6.464	87.333 ± 3.887
User 2	62.000 ± 4.000	24.000 ± 18.785	93.333 ± 2.981	92.667 ± 2.494
User 3	19.333 ± 2.494	45.333 ± 28.488	87.333 ± 6.464	99.333 ± 1.333
User 4	90.000 ± 5.578	45.333 ± 31.312	95.333 ± 3.399	94.667 ± 2.667
Average User Accuracy	54.667 ± 0.667	47.167 ± 3.274	93.000 ± 1.453	93.500 ± 1.434

Table 1: **Accuracy of Held out Test-Set Preference queries on MO-Reacher** across 5 seeds with 60 adaptive preference queries and 30 held out preference test queries. Both of our approaches,**Traj-MaskedDP** which assumes privilege annotator id is given like VPL, as well as **CLS-MaskedDP** which does not assume any access to annotator identity, outperforms both the baselines Single Reward Model and VPL, for individual users, as well as on average. The blue background indicates the best performance, while the gray background denotes the second-best performance.

ual user prefers one specific objective i.e. User 1 represents preferring objective 1 over any other objective and so on.

Baselines We evaluate both our privileged and non privileged approach against 1) Single Reward Head Model learned through conventional Bradley Terry loss Model, 2) VPL. Originally VPL was trained with input as only state, and for consistency with original work, we keep the input as state only. For Mo-Reacher we used only 60 adaptive queries and 600 pairwise training preference queries. while for MO-Highway we used 300 pairwise training preference queries and 30 adaptive queries. VPL uses adaptive queries to infer the latent preference variable on which it conditions its reward and policy while our approaches,inspired from (Luo et al., 2025) uses HyRE (Lee et al., 2024), to infer the test-time user's preference vector. We optimize RL policies under learned reward function using Augmented Random Search (Mania et al., 2018) and Proximal Policy Optimization(PPO) (Schulman et al., 2017) from Stable Baselines3 (Raffin et al., 2021) with episodic rewards to for 7e6 timesteps for MO-Reacher and with 2e5 timesteps for MO-Highway respectively. We employ episodic rewards i) as preference dataset is is constructed using fixed horizon to mitigate termination bias (Freire et al., 2020) ii) since reward is computed using PCA preference vector and our representation encoder that summarizes the whole trajectory (see Appendix D for full details). Original VPL is an offline framework which uses Implicit Q-Learning(IQL) (Kostrikov et al., 2022) to learn an offline policy optimized for per timestep reward. We use discrete IQL from TorchRL (Bou et al., 2024). For fair comparison to our frameworks, we learn both an offline IQL policy as well as an online episodic policy for VPL.

Metrics We compare the performance of our approach against earlier frameworks on these metrics: (1) accuracy on adapting to a new held out test-time user's preferences; (2) performance of RL policy optimized under new test-time user's preferences; (3) time taken on NVIDIA GeForce RTX-GT3090 to train these models; (4) Ablation with respect to the number of adaptive queries(For ablation with respect to training queries and number of PCA dimensions, please refer to Appendix C and Appendix A respectively); and 5) Adaptation to out of distribution users where we train on 3 users and test the adaptation to new user both on test and RL policy

Note that in the Original VPL paper (Poddar et al., 2024), the adaptation query set size is referred to as context size in VPL and is noted in their appendix as the main hyperparameter to tune. Therefore, metric (4) directly evaluates prior work under its key adaptation hyperparameter, and we include this ablation to ensure a fair comparison against VPL.

Table 2: **Accuracy of Held out Test-Set Preference queries on MO-Highway** across reported as mean and standard deviation across 5 seeds with 30 adaptive preference queries and 15 held out preference test queries.Both of our approaches,**TrajMaskedDP** which assumes privilege annotator id is given like VPL, as well as **CLSMaskedDP** which does not assume any access to annotator identity, outperforms VPL, for individual users, as well as on average. Single Reward Model achieves competitive performance for one user as our approaches but on average fails to align with each user.

	Baselines		MaskedDP	
	SingleRM	VPL	Traj-MaskedDP	CLS-MaskedDP
User 1	68.000 ± 13.597	30.667 ± 9.043	96.000 ± 5.333	85.333 ± 11.470
User 2	77.333 ± 9.978	12.000 ± 12.220	77.333 ± 16.111	77.333 ± 9.043
User 3	65.333 ± 13.597	53.333 ± 9.428	82.667 ± 3.266	73.333 ± 4.216
Average User Accuracy	70.222 ± 3.326	32.000 ± 8.503	85.333 ± 4.787	78.667 ± 4.576

6 Results

6.1 In-Distribution

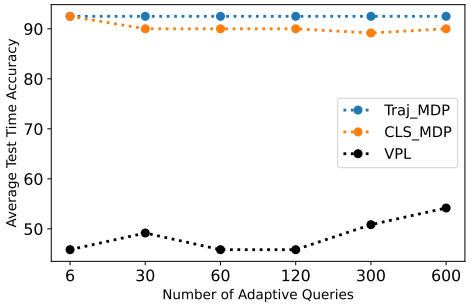
Test-Time Preference Alignment Accuracy In Table 1 and Table 2 for MO-Reacher and MO-Highway, we present the accuracy of adapting to test-time held out preference queries in terms of mean and standard deviation across 5 seeds. We use 2 PCA eigenvectors across both our frameworks.For more results on number of PCA components, please refer to Appendix A. We observe that for both environments, our approaches Traj-MaskedDP and CLS-MaskedDP outperform both Single Reward Model and VPL, where Traj-MaskedDP makes same assumptions about prior knowledge of annotator identity as VPL while CLS-MaskedDP approach has no such prior knowledge. Thus we show better alignment to held out queries when learning from diverse crowd feedback with no information about annotator identity.

RL performance under Test-Time Preference Alignment Table 3 on MO-Reacher and Table 4 on MO-Highway compare performance of RL policies learned using ARS as online RL algorithm across all frameworks except for VPL, where we also optimized an offline IQL policy, for fair comparison to Poddar et al. (2024) who used IQL. We report the RL performance as mean and standard deviation over 100 rollouts across 5 seeds for each individual user as well as average across all users all seeds per model. We observe from Table 3 that for MO-Reacher, both of our frameworks, for most of users on average, deliver a more personalized policy than VPL or Single Reward Model except for User 2, where VPL gets a higher score when learning with per-timestep offline RL IQL policy than any other models which optimize an online RL episodic policy. However, this performance appears to stem from the offline IQL policy optimization rather than the VPL reward model itself since optimizing an online ARS policy under the VPL reward, fails to achieve comparable performance. We can see from Table 4, for 2 users User 1 and User 2, who prefer high speed driving and driving in right most lane respectively, that RL performance under CLS-MaskedDP is highest. This implies that CLS-MaskedDP yields a more personalized policy than any other model, even though it has no prior information about annotator identity unlike VPL and Traj-MaskedDP. User 3 prefers driving most conservatively to avoid collision, and interestingly we notice that Single Reward Model here does better than any other personalized framework. We hypothesize that might be due to the limited diversity of trajectories in the offline preference dataset for this behavior, since this behavior uses sparse rewards: 0 for avoiding collisions and -1 for collisions. Additionally, since the environment is fixed horizon and does not terminate on collision, negative rewards for this user keep accumulating. But on average across all users, RL performance of our privilege approach Traj-MaskedDP, still outperforms all baselines.

Training Time Comparison We report the time taken to train all frameworks in Table 3a on Nvidia GPU 3090. From the table we see that VPL incurs the highest training cost,taking up to

	Baselines	MaskedDP	
	VPL	Traj-MaskedDP	CLS-MaskedDP
Training Time	225.21 ± 9.33	7.83 ± 0.97	4.16 ± 0.06

(a)



(b)

Figure 3: (a) **Training time** in seconds reported as mean and standard deviation across 5 seeds: the lower time it takes, the faster the framework is. (b) **Test-time accuracy vs. the number of queries used for adaptation** demonstrates that both our frameworks Traj-MaskedDP and CLS-MaskedDP achieve high accuracy around 90% with smaller number of preference queries as with bigger adaptation query set size whereas VPL fails to achieve a comparable performance even when fed 100x adaptation queries per user.

Table 3: **RL Performance on MO-Reacher** when evaluated over 100 episodes across 5 seeds reported in terms of mean +/- std. Both our frameworks, Traj-MaskedDP and CLS-MaskedDP achieve higher average RL performance than the baselines, demonstrating improved policy alignment.

	Baselines			MaskedDP	
	SingleRM-ARS	VPL-OfflineRL-IQL	VPL-OnlineRL-ARS	Traj-MaskedDP-ARS	CLS-MaskedDP-ARS
User 1	13.871 ± 7.598	1.491 ± 7.342	14.201 ± 9.170	37.845 ± 6.244	29.024 ± 7.352
User 2	19.369 ± 8.004	35.556 ± 13.217	-1.613 ± 8.899	31.247 ± 4.934	31.216 ± 1.413
User 3	17.436 ± 6.164	12.122 ± 5.609	12.241 ± 9.300	31.031 ± 3.827	26.740 ± 6.324
User 4	16.140 ± 6.729	14.780 ± 12.189	-0.118 ± 6.961	26.708 ± 1.444	27.432 ± 4.787
Average	16.704 ± 2.286	15.987 ± 14.196	6.178 ± 8.557	31.708 ± 4.520	28.603 ± 1.996

Table 4: **RL performance MO-Highway** presented in terms of mean and standard deviation across 5 seeds with 100 rollouts per seed.We achieve competitive performance as compared to baselines on individual users and on average except User 3 which prefers strictly conservative driving.

	Baselines			MaskedDP	
	SingleRM-PPO	VPL-IQL	VPL-OnlineRL-PPO	Traj-MaskedDP-PPO	CLS-MaskedDP-PPO
User 1	3.335 ± 4.459	3.117 ± 4.725	5.805 ± 3.840	5.120 ± 3.651	6.105 ± 3.984
User 2	15.850 ± 12.777	24.545 ± 9.388	0.717 ± 2.170	29.024 ± 3.604	29.736 ± 1.043
User 3	-6.276 ± 9.282	-6.520 ± 9.863	-20.280 ± 7.502	-10.980 ± 10.782	-20.462 ± 7.263
Average RL performance	4.303 ± 11.110	7.047 ± 16.028	-4.586 ± 13.744	7.721 ± 19.972	5.126 ± 24.823

50× longer to train than CLS-MaskedDP and up to 30× longer than Traj-MaskedDP, even though Traj-MaskedDP makes similar assumptions. CLS-MaskedDP is consistently the fastest, requiring the least training time as reported across 5 seeds. This has significant practical implications with respect to computational cost, energy consumption, and compute budget.

Test-Time Preference Alignment vs. Adaptation Query Budget In Fig 3b, we compare and contrast impact of number of adaptive queries on test accuracy across our approaches against VPL. Both of our frameworks, Traj-MaskedDP and CLS-MaskedDP, adapt to a new user using any number of preference queries, allowing flexible adaptation at test time. By contrast, VPL requires a fixed

Table 5: **Out-of-distribution adaptation to new users**. Results report mean ± std over 2 seeds for test preference accuracy (15 queries) and RL performance (100 evaluation rollouts). Our methods generalize better to unseen users, achieving higher preference accuracy and stronger downstream RL performance than VPL.

Setup			Baselines	MaskedDP	
OOD New User	Trained on Users	Eval Metric	VPL	Traj-MaskedDP	CLS-MaskedDP
1	2-3-4	Test	66.667±0.000	100 ± 0.00	100 ± 0.00
		RL	8.299 ± 3.809	32.369 ± 3.093	27.203 ± 2.146
2	1-3-4	Test	66.667 ±0.00	100 ±0.00	100 ±0.00
		RL	8.746 ± 3.492	31.195 ± 1.304	26.647 ± 0.343
3	1-2-4	Test	100.00±0.00	100.00±0.00	76.667±3.333
		RL	17.515 ± 11.603	34.047 ± 4.927	30.622 ± 11.509
4	1-2-3	Test	0.00±0.00	100.00±0.00	100.00±0.00
		RL	1.484 ± 10.833	35.999 ± 2.872	14.032 ± 4.555
Average		Test	58.333 ± 36.324	100.00±0.00	94.167 ± 10.104
		RL	9.011 ± 5.691	33.403 ± 1.809	24.626 ± 6.303

number of queries to infer the latent preference for a new user. This stems from VPL architecture which uses a VAE with a linear layer as first layer, rendering it unable to handle variable sized adaptation query sets. Consequently, we train a new VPL model for each adaptation query size. Our results demonstrate that both our frameworks, can not only work with any variable size of adaptation queries but can also achieve competitive performance of 90% accuracy with smaller number of adaptation queries as they obtain with bigger query sets; whereas VPL remains highly sensitive to the adaptation size and performs substantially worse even with significantly larger query sets of 100x per user.For per-user adaptation ablation , please ref to Appendix B

6.2 Personalization to Out-of-Distribution New Users

We explore personalization to Out-Of-Distribution(OOD) new user at test time by training on any possible set of 3 users in MO-Reacher and testing adaptation to new user with an unseen behavior, both with respect to test set accuracy as well as RL performance. For adaptation to new user to get k_j for our frameworks and to infer latent preference variable for VPL, we use 30 adaptation queries. And since this is an online adaptation to new user out of training distribution, we only optimize an RL ARS policy under VPL unlike in-distribution VPL-IQL. We report mean and standard deviation across 2 seeds for 15 test preference queries and 100 rollouts for RL for the two metrics respectively in Table 5. Our approach, Traj-MaskedDP outperforms VPL on both test set accuracy as well as on downstream RL performance for 4 different OOD users as well as on average, while making same assumptions about access to annotator identity in offline diverse dataset. It also outperforms CLS-MaskedDP, which makes no such assumption so is trained under weaker settings. Notably, despite being trained without annotator identity information, CLS-MaskedDP still achieves higher RL performance than VPL when adapting to new OOD user. It also outperforms VPL on average in test set accuracy.

7 Discussion and Conclusion

We present a framework for pluralistic alignment for personalized policy learning from heterogeneous crowd feedback both with and without privileged information about annotator identities. Experiments in multi-objective control environments show improved preference alignment, competitive RL performance,significantly reduced training cost compared to prior approaches, as well as generalization to both in-distribution and out-of-distribution users. Our method also supports flexible adaptation at test time thus lowering the human interaction burden at test time.

References

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, et al. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*, 2024.

Ryan Bahlous-Boldi, Li Ding, Lee Spector, and Scott Niekum. Pareto optimal learning from preferences with hidden context. In *Reinforcement Learning Conference*, 2025. URL https://openreview.net/forum?id=7R0azwu9Fx.

Andrea Bajcsy, Dylan P. Losey, Marcia K. O'Malley, and Anca D. Dragan. Learning robot objectives from physical human interaction. In Sergey Levine, Vincent Vanhoucke, and Ken Goldberg (eds.), *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pp. 217–226. PMLR, 13–15 Nov 2017.

Albert Bou, Matteo Bettini, Sebastian Dittert, Vikash Kumar, Shagun Sodhani, Xiaomeng Yang, Gianni De Fabritiis, and Vincent Moens. TorchRL: A data-driven decision-making library for pytorch. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QxItoEAVMb.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Daniel S. Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International conference on machine learning*, pp. 783–792. PMLR, 2019.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

David Chhan, Ellen Novoseller, and Vernon J Lawhern. Crowd-prefrl: Preference-based reward learning from crowds. *arXiv preprint arXiv:2401.10941*, 2024.

Paul F Christiano, Jan Leike, Tom B Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *NIPS*, 2017.

Florian Felten, Lucas N. Alegre, Ann Nowé, Ana L. C. Bazzan, El Ghazali Talbi, Grégoire Danoy, and Bruno 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.

Pedro Freire, Adam Gleave, Sam Toyer, and Stuart Russell. Derail: Diagnostic environments for reward and imitation learning. *arXiv preprint arXiv:2012.01365*, 2020.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2024.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=68n2s9ZJWF8.

Kimin Lee, Laura Smith, Anca Dragan, and Pieter Abbeel. B-pref: Benchmarking preference-based reinforcement learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL https://openreview.net/forum?id=ps95-mkHF_.

Yoonho Lee, Jonathan Williams, Henrik Marklund, Archit Sharma, Eric Mitchell, Anikait Singh, and Chelsea Finn. Test-time alignment via hypothesis reweighting. *arXiv preprint arXiv:2412.08812*, 2024.

Fangchen Liu, Hao Liu, Aditya Grover, and Pieter Abbeel. Masked autoencoding for scalable and generalizable decision making. *Advances in Neural Information Processing Systems*, 35:12608–12618, 2022.

R Duncan Luce et al. *Individual choice behavior*, volume 4. Wiley New York, 1959.

Feng Luo, Rui Yang, Hao Sun, Chunyuan Deng, Jiarui Yao, Jingyan Shen, Huan Zhang, and Hanjie Chen. Rethinking diverse human preference learning through principal component analysis. *arXiv preprint arXiv:2502.13131*, 2025.

Horia Mania, Aurelia Guy, and Benjamin Recht. Simple random search of static linear policies is competitive for reinforcement learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/7634ea65a4e6d9041cfd3f7de18e334a-Paper.pdf.

Shaunak A Mehta and Dylan P Losey. Unified learning from demonstrations, corrections, and preferences during physical human–robot interaction. *ACM Transactions on Human-Robot Interaction*, 13(3):1–25, 2024.

Vivek Myers, Erdem Biyik, Nima Anari, and Dorsa Sadigh. Learning multimodal rewards from rankings. In Aleksandra Faust, David Hsu, and Gerhard Neumann (eds.), *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 342–352. PMLR, 08–11 Nov 2022. URL https://proceedings.mlr.press/v164/myers22a.html.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=gRG6SzbW9p.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition, 1994. ISBN 0471619779.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL http://jmlr.org/papers/v22/20-1364.html.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Idan Shenfeld, Felix Faltings, Pulkit Agrawal, and Aldo Pacchiano. Language model personalization via reward factorization. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=E7Tu5yjqXw.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in RLHF. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=0tWTxYYPnW.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. Position: A roadmap to pluralistic alignment. In *International Conference on Machine Learning*, pp. 46280–46302. PMLR, 2024.

Jeremy Tien, Jerry Zhi-Yang He, Zackory Erickson, Anca Dragan, and Daniel S. Brown. Causal confusion and reward misidentification in preference-based reward learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=R0Xxvr_X3ZA.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. URL https://arxiv.org/abs/1909.08593.

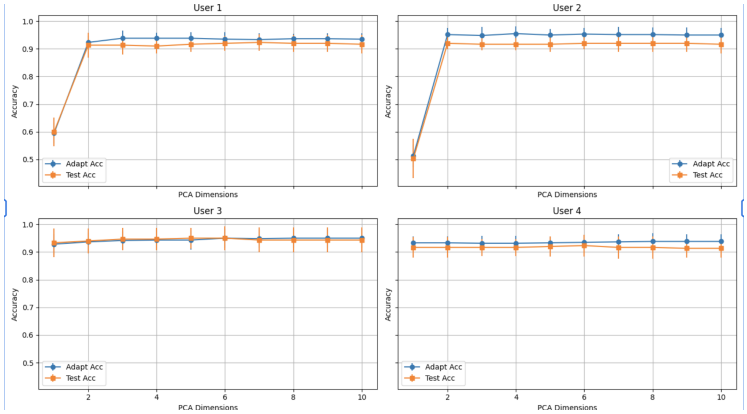


Figure 4: **Per User Test Accuracy vs Number of PCA basis eigenvectors** calculated per user across 10 seeds for each PCA size with standard deviation plotted as error bars.

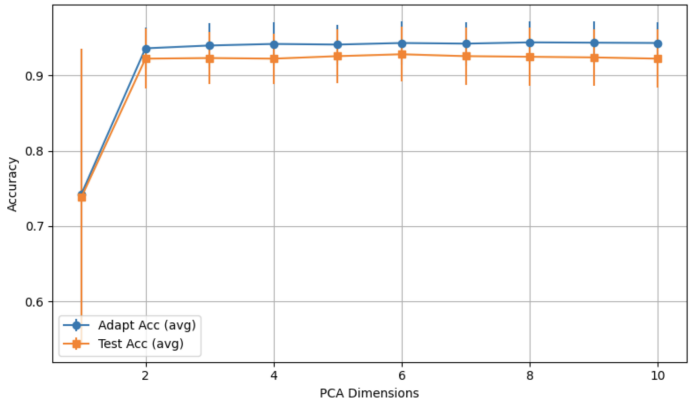


Figure 5: **Average Test Accuracy vs Number of PCA basis eigenvectors** calculated across all users over 10 seeds for each PCA size with standard deviation plotted as error bars.

A Ablation on PCA dimensions

We show test time accuracy as no of PCA dimensions are varied for each individual user in Fig 4 and averaged across all users in Fig 5 respectively. For all PCA dimensions, Traj-MaskedDP PCA is trained with 600 queries , adapted with 60 and tested with fixed set of 30 queries per user using Traj-MaskedDP .

B Ablation Per-User Adaptive queries

All models are trained with 600 preference queries and tested on fixed set of 30 queries per user. We observe similar trend as we do for average test time accuracy vs adaptation query size, that our frameworks can do comparably high even with smaller adaptation number of queries as they can with $100x$ queries for adaptation whereas VPL us very sensitive to no of queries used for inferring its latent preference .

C Ablation on Training queries

All models are adapted with 60 preference queries and tested on fixed set of 30 queries per user. Our framework in both settings outperforms VPL as shown in Fig 7

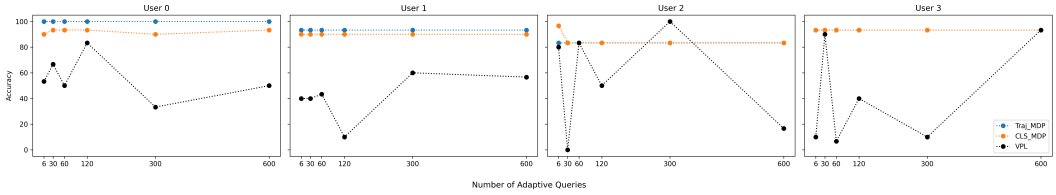


Figure 6: **Per-User Test Time Accuracy vs No of adaptation queries** where each user favors a distinct goal in the environment.

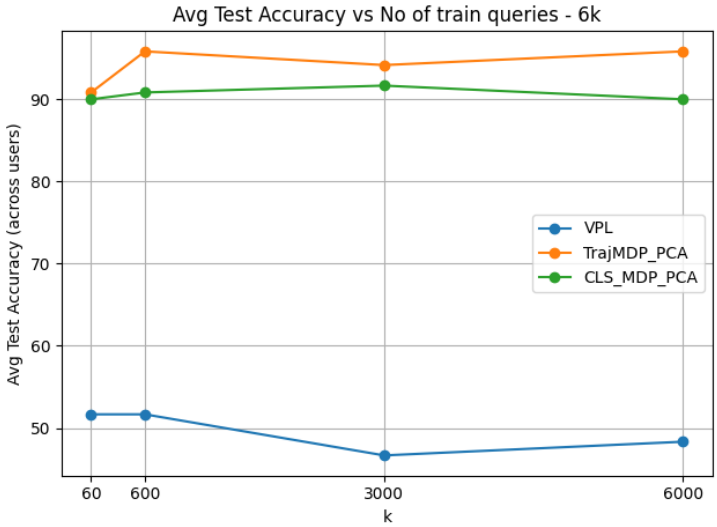


Figure 7: **Training Query size ablation**

D RL experiment details

MO-Reacher is set to have fixed horizon of 50 and does not any terminal condition by default. We use ARS with default parameters to optimize an RL policy using episodic rewards. MO-FastHighway on other hand by default ends if there's a collision or if lapses 30 timesteps. So to avoid early termination bias, we ignore the terminal condition and set the environment to have fixed horizon of 30. We use Vectorized Norm with $norm_obs = True$ and $norm_reward = True$ from SB3 to train PPO policies with normalized obs and normalized rewards in Fast Highway.