Zero-shot Personalisation via Dynamic Policy Fusion

Ajsal Shereef
A2I2
Deakin University
Australia
a.palattuparambil@deakin.edu.au

Thommen George Karimpanal
School of IT
Deakin University
Australia
thommen.karimpanalgeorge@deakin.edu.au

Santu Rana
A2I2
Deakin University
Australia
santu.rana@deakin.edu.au

***Abstract*—For many applications, the policy resulting from training a deep reinforcement learning (RL) agent, although optimal for the task, may not cater to the preferences of all users. A naïve solution would be to retrain the agent using a reward function that encodes the user's specific preferences. However, such a reward function is not readily available, and retraining the agent from scratch can be prohibitively expensive, considering the sample-intensive nature of RL. A more practical approach is to adapt the already trained task policy to user-specific needs with the help of human feedback - a process we term *Personalisation*. Specifically, we learn a personalised policy by inferring a user's intended policy (intent-specific policy) through trajectory-level feedback from the user and combining it with the task policy via policy fusion. We also develop a strategy to dynamically re-weight the constituent policies to ensure a fine balance between meeting the user's requirements and completing the task at hand. Further, we establish theoretical bounds on the personalised policy with respect to the task policy. Empirical evaluations in Highway, Pong and 2D Navigation environments show that our proposed dynamic policy fusion approach consistently completes the intended task while adhering to user-specific needs.**

***Index Terms*—Reinforcement Learning, Personalisation, Dynamic Policy Fusion**

I. Introduction

Reinforcement learning (RL) has demonstrated its effectiveness in various real-world scenarios, including computer games [8], inventory management, autonomous driving, robotics [1], healthcare [9], recommendation systems [10] etc. In RL, the learning agent typically learns to maximise its reward by learning an optimal task policy through interactions with the environment. However, in certain scenarios, a user may desire a policy that subtly deviates from this optimal policy to accommodate their personal preference or style.

For example, in a navigation setting, users may want to avoid toll roads, or they may prefer to drive along a coastal route, even if this translates to slightly longer travel times compared to the fastest route, as specified by a task policy trained to minimise the travel time.

A simple solution to accommodate such preferences in a policy would be to retrain an RL agent from scratch, using a reward function that takes users' preferences into account. However, such a reward function could be challenging to design. In addition, the poor sample efficiency of RL could render this approach infeasible. A more practical approach would be to adapt the already trained task policy and adapt (personalise) it to respect a user's specific preferences.

The challenge of respecting additional user-specific objectives could also be formulated as a multi-objective reinforcement learning [11] problem, although such approaches are applicable only when rewards corresponding to each objective are available. However, the present work aims to deal with scenarios where such rewards are unknown or unavailable, with a primary focus on adapting an already trained policy.

We present a policy fusion-based approach to learn such a *Personalised policy* that respects user preferences while also satisfying the task objective. Specifically, by collecting trajectory-level human feedback in the form of trajectory scores, we use an LSTM-based approach to infer an *Intent-specific policy* which encapsulates the user-specific needs and/or preferences. The key idea is then to fuse the task-specific policy with this intent-specific policy, such that the resulting personalised policy completes the task at hand while simultaneously respecting user-specific needs. We note that the same trajectories used for training the task policy can be reused for eliciting human feedback, thereby obviating the need for any additional environment interactions. In addition to this, we posit that such an approach translates to a lower cognitive load for the user, as providing trajectory-level feedback is cognitively less demanding compared to providing feedback at a state-action level [3]. We also theoretically prove that the divergence of the personalised policy with respect to the task policy is bounded.

Despite the mentioned advantages of the described policy fusion-based approach, we show that a naïve policy fusion approach is insufficient, and can lead to undesirable effects. For example, in the navigation scenario, the preference for visiting a particular state may cause the agent to repeatedly visit that state, thereby ignoring the primary navigation objective. In other words, static policy fusion may cause one of the policies to dominate the other, leading to undesired agent behaviours. We address this problem through a novel dynamic policy fusion approach that automatically regulates the dominance of the intent-specific policy by regulating the associated temperature parameter of the Boltzmann distribution used to represent the policy. This helps control the relative contribution of the intent-specific policy in the personalised policy, thereby

allowing the agent to complete the task while respecting user-specific needs/preferences.

Overall, the contributions of this work can be summarised as follows:

- A technique to achieve zero-shot personalisation by adapting an already trained task policy without additional environmental interactions.
- Efficient inference of human intent from trajectory-level feedback, without the need for effort-intensive state-action-level feedback.
- Theoretical analysis establishing the boundedness of the personalised policy with respect to the task policy.
- A modulation mechanism for dynamic policy fusion to balance task performance with user preference.

II. Related Work

Learning from human feedback is of particular interest in the RL community as it leverages human knowledge during the learning process, offering several benefits. Firstly, it improves the efficiency of the system in terms of sample requirements as well as overall performance. In Guan et al. [12], the inclusion of human feedback significantly improved the sample efficiency and performance. Secondly, leveraging human feedback has been shown to enable RL agents to solve complex tasks that are otherwise challenging to manually specify through conventional reward functions. This has been demonstrated in previous works [3], [13], where a reward model is first learned from human preferences on trajectory-level data and subsequently used to train the agent's policy.

Personalisation intersects with many domains, including safety, preferences, customisation, multi-objective reinforcement learning (MORL), transfer learning, and more. For instance, safety can be seen as a form of personalisation, where an agent must obey safety constraints and act in the feasible region to ensure safety. Multi-objective reinforcement learning (MORL) embodies personalisation by determining which objectives take precedence. Addressing conflicting objectives concurrently is MORL's aim, as illustrated in Basaklar et al. [14], which seeks Pareto-optimal solutions across the preference space via a single training process. Related research exists as policy composition, as referenced in Haarnoja et al. [15] and Hunt et al. [16]. Such works explore the concept of learning different policies independently, each with its own reward function. These individual policies are later combined, leading to the emergence of new behaviours.

Unlike these works, our approach does not require any additional interaction or specifically crafting the reward function as in Mo et al. [17] and Bodas et al. [18]. In [17], a personalised dialogue system is developed using transfer learning, adapting common dialogue knowledge from a source domain to a target user, whereas [18] aims to personalise non-player characters (NPCs) to human skill levels to enhance player engagement. This approach designs a composite scalar reward function that implicitly matches the NPC's skill level to that of the human player. However, it necessitates additional environment interaction and NPC training with the engineered reward function.

In contrast to the approaches discussed above, our approach to personalisation looks to adapt an already existing policy towards an inferred intent-specific policy without the need for additional environment interactions. The primary idea is to adapt the trained policy to respect an individual's requirements without compromising the performance of the original task.

III. Preliminaries and Background

We briefly survey some of the related backgrounds that forms the basis for our work. We refer to the policy trained to solve the task as the task-specific policy (π_ϕ) and the policy that captures human intent as the *intent-specific policy*(π_ψ).

A. Reinforcement Learning

We consider tasks to be formulated as MDPs $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$ where $\mathcal{S}$, and $\mathcal{A}$ are the state and action spaces. P is the transition function that captures the transition dynamics of the environment. R is the reward function and γ is the discount factor. At timestep t, the agent in state $s_t \in \mathcal{S}$ takes an action $a_t \in \mathcal{A}$ and obtains a reward $R(s_{t+1}, s_t, a_t)$ and moves to state s_{t+1} according to the transition function $P(s_{t+1}|s_t, a_t)$. A policy $\pi(a|s)$ outputs the probability of taking an action a from a given state s. The episodic discounted return is $G_t = \sum_{t=0}^{T} \gamma^t r_t$, where γ specifies how much the future reward is discounted and T is the total number of timesteps. The agent's objective is to maximise the future expected reward $\mathbb{E}[G_t]$ by learning a Q-function $Q(s_t, a_t)$, which estimates the expected cumulative reward the agent receives from a state s_t by taking action a_t and following policy π.

B. RUDDER

RUDDER [2] addresses the problem of credit assignment and learning from sparse rewards. In RUDDER, an LSTM network analyses the entire trajectory data with a score received at the end to estimate Q-values for individual state-action pairs within the trajectory. These Q-values represent the expected future reward for taking an action in a given state. RUDDER also showed a densified reward can be formulated from the human feedback score as follows,

$$E[r_{t+1}|s_t, a_t, s_{t-1}, a_{t-1}] = Q'(s_t, a_t) - Q'(s_{t-1}, a_{t-1}), \quad (1)$$

where, $Q'(s_t, a_t)$ represents the LSTM-generated Q-values corresponding to the state-action pair at the specific time step t. We use the RUDDER framework to learn the Q-values corresponding to human scores and later convert the Q-values to policies as described in Sections IV-A and IV-B. We use the densified reward as in (1) to modulate the influence of the policy as discussed in Section IV-D.

IV. Methodology

We construct our proposed zero-shot approach for personalisation by first inferring the intent-specific policy via trajectory-level human feedback using RUDDER, followed

by a dynamic policy fusion mechanism that automatically maintains a balance between the inferred policy and a trained task-specific policy.

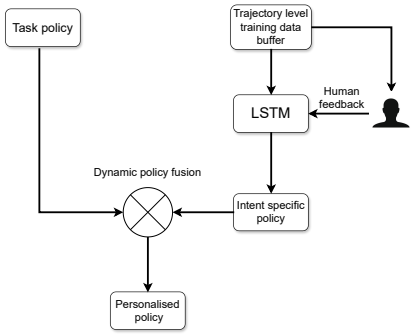

Fig. 1. Summary of our personalisation method. Trajectory-level training data used to train the task policy is labelled by a human user and an LSTM model is employed to identify the intent of the human. The output from the LSTM is converted to an intent-specific policy which is then dynamically fused (Discussed in Section IV-D) with the task-specific policy to obtain the personalised policy.

We assume that the trajectory data used to train task-specific policy is accessible, and reusing this data allows us to perform zero-shot personalisation i.e., without collecting new environment interactions. A subset of this data is sampled with personalised human feedback scores with more desirable trajectories assigned higher scores. We then infer the human intent (intent-specific policy) using the idea discussed in Section IV-A and dynamically fuse it with the task-specific policy. The overview of our method is illustrated in Figure 1.

The subsequent sections provide in-depth details of the components of our personalisation approach.

A. Learning Human Intent using LSTM

To learn the intent-specific policy, as previously described, we leverage the LSTM-based approach of RUDDER. We train this LSTM using human trajectory-level feedback on the same training data used to learn the task-specific policy (Refer Appendix A). Hence, no additional interaction data is required. At each time step, the state-action vector is fed into the LSTM units. In the case of image inputs, we pre-train a Variational Autoencoder (VAE) to reduce the dimensionality of the state. We use FiLM [5] to modulate the state vector with the action vector. In [5], a FiLM network conditions the feature map of the neural network depending on another input signal. Here, we condition the state feature with the action vector and train the LSTM using the human feedback data as training labels. With this training setup, the model approximates the Q-value, which is then converted to the intent-specific policy as described in the next section.

B. Policy construction and fusion

We use a DQN parameterised by ϕ to learn the task and it produces the Q-values (Q) for each state-action pair. Similarly, the LSTM parameterised by ψ also produces the Q-values (Q') corresponding to the human intent. We choose the Boltzmann distribution to convert the Q-values into a policy.

$$\pi_\phi(a|s_t) = \frac{\exp\left(\frac{Q(s_t,a)}{T_\phi}\right)}{\sum_{a\in\mathcal{A}}\exp\left(\frac{Q(s_t,a)}{T_\phi}\right)}, \tag{2}$$

$$\pi_\psi(a|s_t) = \frac{\exp\left(\frac{Q'(s_t,a))}{T_\psi}\right)}{\sum_{a\in\mathcal{A}}\exp\left(\frac{Q'(s_t,a))}{T_\psi}\right)}, \tag{3}$$

where T_ϕ is the temperature corresponding to task-specific policy and T_ψ is the temperature corresponding to intent-specific policy.

Policy fusion is a process of combining two or more policies to produce a new policy as done in Sestini et al. [4]. The policy fusion in the context of personalisation should satisfy two constraints. Firstly, the fused policy should not change if the task objective and the human objective are identical - which we refer to as *invariability property*. Secondly, the fused policy should act on the common support of the policies being fused. Therefore to accommodate the aforementioned constraints, we propose a new fusion method as follows,

$$\pi_f(a|s) = \sqrt{\pi_\phi(a|s_t) \times \pi_\psi(a|s_t)}. \tag{4}$$

Here, π_ϕ and π_ψ represent the probability of choosing an action from a given state. Since these two policies are independent, their joint probability can be expressed as the product of π_ϕ and π_ψ. This ensures that the fused policy acts on the common support. Moreover, when the two policies are the same, the KL-divergence between π_ϕ and π_f is 0, satisfying the invariability property. Please refer Appendix B for more details.

With this policy construction, we can bound the divergence of the π_ϕ and π_f if the divergence of the corresponding Q-values and the temperature is bounded as stated by the following theorem.

Theorem 1. *Let Q and Q' be Q-values corresponding to the task policy and the intent-specific policy respectively. Let $\pi_\psi(a|s)$ and $\pi_\phi(a|s)$ represent the respective policies, with corresponding temperatures T_ψ and T_ϕ. Let $\|Q(s,a) - Q'(s,a)\|_2 < \epsilon \ \forall s \in \mathcal{S}$ and $a \in \mathcal{A}$ and $\|T_\psi - T_\phi\|_2 < \delta$. Then,*

$$KL\left(\pi_\phi(a|s)\|\pi_f(a|s)\right) \leq \log(Z) + \frac{1}{2}\left(\frac{Q^*\delta + \epsilon T_\phi}{T_\phi T_\psi}\right) + \frac{1}{2}\log(\zeta) \ \ \forall a \in \mathcal{A}, s \in \mathcal{S},$$

where Z is the normalising factor of π_f, $Q^* = Q(s, argmax_{a\in\mathcal{A}}Q(s,a))$ and $\zeta = \frac{h(Q',T_\psi)}{h(Q,T_\phi)}$, here $h(Q,T) = \sum_a exp\left(\frac{Q}{T}\right)$.

Despite the invariance properties of such a fused policy, in practice, we found that this approach of statically fusing policies is subject to certain pitfalls, as explained in the next

section. Subsequently, we address the issue through our novel dynamic policy fusion approach in Section IV-D.

C. Pitfalls of Static Fusion

With the naïve static policy fusion techniques described in Section IV-B, which we refer to as static policy fusion methods, a potential challenge arises wherein one of the merged policies over-influences the personalised policy. Consequently, the agent may disproportionately exhibit the corresponding behaviour, resulting in noncompliance with either the task-specific policy or the intent-specific policy.

To illustrate this phenomenon, we consider a 2D Navigation scenario where an agent is tasked with reaching a target location. However, a human may wish some checkpoint state, different from the target state, to be visited before the agent reaches its target. In this case, the task-specific policy corresponds to the actions along the shortest path towards the target, and an intent-specific policy would correspond to one inferred from human feedback, which exclusively favours visiting the checkpoint state. A static policy fusion approach in this case could lead to the agent visiting the checkpoint state indefinitely, thereby failing its original objective of navigating to the target. This motivates the need for fusing the policies in a dynamic fashion, such that the agent respects the human's preferences, while also simultaneously completing the task at hand. We develop such a dynamic fusion technique to control the relative dominance of the individual policies by modulating the temperature parameter T_ψ of the intent-specific policy. We now describe the details of our dynamic fusion strategy.

D. Dynamically Modulating the Policy Fusion

To mitigate unintended policy dominance, we adopt the idea that when the temperature parameter T_ψ in Equation (3) is increased, the probability distribution of actions tends to become uniform, thereby reducing the influence of the intent-specific policy on the personalised policy. We therefore modulate T_ψ depending on whether the personalised policy exhibits over-adherence or under-adherence to the intent-specific policy π_ψ.

However, we note that human intent can be specified in various modes: *preference* (where a state is preferred over others), *avoidance* (where the preference is to avoid a particular state) or *mixed* (where the preference is to avoid certain states and to prefer certain others). In avoidance cases, human feedback assigns lower scores to unfavourable trajectories (those that pass through the state(s) to be avoided) while in the preference case, higher scores are assigned to preferred trajectories (those that pass through the preferred state(s)). Depending on the trajectory score from the human, the LSTM assigns rewards to each state-action pair as in Equation (1) which we refer to as *human-induced rewards*.

To modulate T_ψ, we first compute the shifted rewards as follows:

$$r' = r - mean(r), \tag{5}$$

where r is the vector that contains the rewards for different actions from a given state. The purpose of this shifting operation is to ensure that irrespective of the mode of human intent, the reward vector r' contains elements with both positive as well as negative values. This shifted human-induced reward r' is then used to modulate T_ψ as:

$$T_\psi(t) = \max\left(T_{min}, \frac{T_{max}}{1 + \exp\left(-\left(g(t) - \eta\right)\right)}\right), \tag{6}$$

where T_{min} and T_{max} are the minimum and maximum allowable temperatures and $g(t) = \sum_{t'=0}^{t} r'(s_{t'}, a_{t'})$ is the accumulated shifted human-induced reward (Refer Appendix D).

How does shifting (Equation (5)) enable two-way switching? A high accumulated shifted human-induced reward ($> \eta$) indicates the agent has obeyed the intent-specific policy π_ψ disproportionately more, which flags the need to mitigate the influence of π_ψ. Since the temperature in Equation (6) is monotonic with respect to accumulated reward r', a higher accumulated r' outputs a higher temperature, and this reduces the influence of the π_ψ in the policy fusion step. Conversely, if the accumulated reward $r' < \eta$, the temperature is lowered, which strengthens π_ψ. Therefore, this setup enables a two-way dynamic switching between the policies. Based on the threshold η, the influence of π_ψ is reduced when the accumulated r' is high, and strengthened when it is too low. Our overall algorithm for personalisation is summarised in Algorithm 1.

Algorithm 1 Personalising the action selection in an episode

Input: DQN and LSTM Networks ϕ and ψ, accumulated reward threshold η, DQN temperature T_ϕ, Min and Max temperature T_{min} and T_{max} **Reset** the environment to get the initial state s_0

Initialise: $Q'(s_0, a_0) = 0$, $g(0) = 0$, $T_\psi = max(T_{min}, \frac{T_{max}}{1+exp(\times\eta))})$, $done = False$, $t = 0$

while not $done$ **do**
  $Q_t \leftarrow []$ ▷ DQN Q-values reset the vector to null
  $Q'_t \leftarrow []$ ▷ LSTM Q-values reset the vector to null
  $r_t \leftarrow []$ ▷ Human-induced reward reset the vector to null
  $Q_t \leftarrow \phi(s_t)$ ▷ Invoking DQN
  $Q'_t \leftarrow \phi(s_t, a)$ ▷ Invoking LSTM for each action
  Update r_t as in Equation (1)
  $\pi_\phi \leftarrow BoltzmanDistribution(Q_t, T_\phi)$ ▷ Task-specific policy
  $\pi_\psi \leftarrow BoltzmanDistribution(Q'_t, T_\psi)$ ▷ Intent-specific policy
  $a_t \leftarrow \arg\max_{a \in \mathcal{A}} \sqrt{\pi_\phi(a|s_t) \times \pi_\psi(a|s_t)}$ ▷ Policy fusion
  $r' \leftarrow r_t - mean(r_t)$ ▷ Adjusting as in Eq (5)
  $g(t) \leftarrow \sum_{t'=0}^{t} r'(s_{t'}, a_{t'})$ ▷ Accumulated shifted human-induced reward
  $T \leftarrow max(T_{min}, \frac{T_{max}}{1+exp(-m(g-\eta))})$ ▷ Updating T
  Execute the action a_t from state s_t to transition to s_{t+1}
  $t \leftarrow t + 1$ ▷ Increment the time-step
  if Episode terminate **then**
    $done = True.$
  end if
end while

V. EXPERIMENTS

We demonstrate our approach in our custom 2D Navigation environment, Highway [6], Pong [7]. Specifically, we conduct personalisation experiments in three scenarios: *Avoidance*, *Preference* and *Mixed*, except for Pong, where we only consider the *Preference* scenario. We first describe these environments as follows:

2D Navigation: This is our custom environment where an agent is tasked with navigating to a target state while avoiding undesired states and visiting desired ones. The observation consists of a 40×40 grayscale image with four directional movements allowed. Actions that put the agent out of the frame are invalid. A +1 reward is granted upon reaching the target location, and a 0 reward is given for all other actions. The episode concludes within 20 timesteps or upon reaching the target.

Pong: In Pong, the observations correspond to ball x (horizontal) and y (vertical) position and velocity, player y position and velocity and CPU y position. Actions involve increasing, decreasing, or changing the velocity of the paddle. A +1 reward is obtained upon winning and a 0 reward for losing the game. Pong concludes after 1000 timesteps or upon game outcome (win or loss).

Highway: The goal is to navigate through traffic as fast as possible. The observations consist of a 26-dimensional vector, horizontal (x) and vertical(y) coordinates, corresponding x and y-velocities of five nearby vehicles, and the current lane. The agent can change lanes, stay idle, or adjust speed (move faster or slower). Any action that takes the agent out of frame is void. This environment has a dense reward setting, with positive rewards granted for maximizing speed, and negative rewards incurred upon collision with other vehicles. These rewards are normalized between 0 and 1. Episode termination occurs after 50 timesteps or upon collision with other vehicles (Refer Appendix C).

TABLE I
DIFFERENT PERSONALISATION MODES AND CORRESPONDING HUMAN AND TASK OBJECTIVES.

Env	Mode	Human objective	Task objective
Highway	Preference	Prefer a lane	Move as fast as possible Without colliding other vehicles
	Avoidance	Avoid a lane	
	Mixed	Prefer a lane and avoid another	
2D Navigation	Preference	Prefer a region	Move to the target
	Avoidance	Avoid a region	
	Mixed	Prefer a region and avoid another	
Pong	Preference	Prefer certain paddle positions	Win the game

Table II summarises the different environments with corresponding human and task objectives. To personalise the behaviour, we assume human users have lane preferences or aversions in Highway. Likewise, in 2D Navigation, humans prefer the agent to avoid or visit certain regions. In Pong, personalisation is introduced by favouring specific paddle positions.

Although Section IV-B discussed actions being sampled from policies, in practical applications, it is common to exploit greedy actions once the policy is learned and simplifies the implementation. Hence, in our experiments, we deterministically choose actions. This is akin to using a low-temperature parameter setting in the fused Boltzmann policy.

TABLE II
DIFFERENT PERSONALISATION MODES AND CORRESPONDING HUMAN AND TASK OBJECTIVES.

Env	Mode	Human objective	Task objective
Highway	Preference	Prefer a lane	Move as fast as possible Without colliding other vehicles
	Avoidance	Avoid a lane	
	Mixed	Prefer a lane and avoid another	
2D Navigation	Preference	Prefer a region	Move to the target
	Avoidance	Avoid a region	
	Mixed	Prefer a region and avoid another	
Pong	Preference	Prefer certain paddle positions	Win the game

A. Results

We first demonstrate the pitfalls of static fusion as described in Section IV-C and compare it against dynamic fusion in 2D Navigation. In Table III, we display the results of static fusion using temperatures set to T_{min} and $T_{max}/2$, which were chosen to illustrate the over-influence of intent-specific policy and to match the initial temperature of dynamic fusion as outlined in Equation (6) respectively. In the tables, the up arrow $\uparrow$ indicates higher is better, and the down arrow $\downarrow$ indicates lower is better. The result averaged over 10 seeds with each seed containing 300 episodes. With these temperatures, the static fusion agent performs well in the Avoidance mode, exhibiting behaviours that avoid the undesired region, while achieving high scores. However, as seen in the Preference and Mixed modes, the static fusion agent incurs severe drops in its score and instead exhibits a high number of visits to the desired region. This is indicative of the over-dominance of the intent-specific policy component in the personalised policy. Upon examining the agent's behaviour, we observed that the static fusion agent tends to repeatedly visit the desired region, ignoring the primary navigation task.

TABLE III
PERFORMANCE OF THE STATIC FUSION, DYNAMIC FUSION AND THE DQN IN 2D NAVIGATION.

Mode	Method	Desired region $\uparrow$	Undesired region $\downarrow$	Score $\uparrow$
Preference	Static ($T_\psi = T_{min}$)	$\mathbf{4.987 \pm 0.649}$	-	0.284 ± 0.045
	Static ($T_\psi = \frac{T_{max}}{2}$)	3.788 ± 0.452	-	0.602 ± 0.048
	Dynamic	1.459 ± 0.140	-	$\mathbf{1.000 \pm 1.000}$
	DQN	0.085 ± 0.005	-	$\mathbf{1.000 \pm 0.000}$
Mixed	Static ($T_\psi = T_{min}$)	$\mathbf{2.372 \pm 0.123}$	$\mathbf{0.000 \pm 0.000}$	0.739 ± 0.014
	Static ($T_\psi = \frac{T_{max}}{2}$)	1.116 ± 0.318	$\mathbf{0.000 \pm 0.000}$	0.894 ± 0.035
	Dynamic	0.268 ± 0.058	$\mathbf{0.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	DQN	0.050 ± 0.003	0.183 ± 0.006	$\mathbf{1.000 \pm 0.000}$
Avoidance	Static ($T_\psi = T_{min}$)	-	$\mathbf{0.000 \pm 0.000}$	0.974 ± 0.011
	Static ($T_\psi = \frac{T_{max}}{2}$)	-	0.005 ± 0.005	$\mathbf{1.000 \pm 0.000}$
	Dynamic	-	0.006 ± 0.005	$\mathbf{1.000 \pm 0.000}$
	DQN	-	0.094 ± 0.006	$\mathbf{1.000 \pm 0.000}$

[a]Hyper-parameters $T_\phi = 0.4, T_{min} = 1, T_{max} = 10, \eta = 0$.

On the other hand, the dynamic fusion agent strikes a balance between the primary task and intent-specific policy, assimilating both aspects into the behaviour, irrespective of the mode of personalisation. Although the performance of static fusion could be improved in individual modes of personalisation, it would require extensive tuning of the temperature hyperparameter, which is not practically feasible.

These results establish the fundamental limitations of static fusion and highlight the need for dynamic fusion to satisfy both human and task objectives. We further examine the properties of dynamic fusion in the Highway and Pong environments in Tables IV and V respectively with the results averaged over 10 seeds each with 20 episodes. The rewards

in Table V have been re-scaled to $[0, 1]$. From these tables, we observe that dynamic fusion, while sometimes incurring a small reduction in the optimal score (that achieved by DQN), enables the agent to adhere to the required preference (be it in the Avoiding, Preference or Mixed modes), while completing the primary task at hand.

TABLE IV
RESULTS OF HIGHWAY ENVIRONMENT WITH DIFFERENT VALUES OF η.

Mode	Method	Desired lane ↑	Undesired lane ↓	Hits ↓	Score ↑
Avoidance	DQN	-	10.59 ± 0.96	0.11 ± 0.01	39.59 ± 0.46
	Dynamic ($\eta = 0$)	-	0.28 ± 0.08	$\mathbf{0.06 \pm 0.01}$	$\mathbf{39.83 \pm 0.59}$
	Dynamic ($\eta = 1$)	-	0.35 ± 0.12	0.08 ± 0.02	38.96 ± 0.87
	Dynamic ($\eta = 2$)	-	$\mathbf{0.19 \pm 0.11}$	0.09 ± 0.02	38.80 ± 0.51
Preference	DQN	10.94 ± 0.49	-	0.07 ± 0.02	$\mathbf{40.25 \pm 0.64}$
	Dynamic ($\eta = 0$)	25.14 ± 1.44	-	$\mathbf{0.05 \pm 0.02}$	40.02 ± 0.61
	Dynamic ($\eta = 1$)	24.90 ± 1.52	-	0.10 ± 0.03	38.84 ± 0.84
	Dynamic ($\eta = 2$)	$\mathbf{29.91 \pm 1.38}$	-	0.06 ± 0.01	39.27 ± 0.59
Mixed	DQN	11.83 ± 0.64	9.59 ± 0.82	0.08 ± 0.02	$\mathbf{40.06 \pm 0.79}$
	Dynamic ($\eta = 0$)	21.52 ± 2.00	0.85 ± 0.33	0.12 ± 0.01	38.50 ± 0.43
	Dynamic ($\eta = 1$)	23.96 ± 1.67	0.56 ± 0.13	$\mathbf{0.07 \pm 0.01}$	39.56 ± 0.45
	Dynamic ($\eta = 2$)	$\mathbf{25.21 \pm 2.55}$	$\mathbf{0.27 \pm 0.06}$	0.07 ± 0.04	39.15 ± 0.49

[a]Hyper-parameters $T_\phi = 0.6, T_{max} = 5, T_{min} = 0.3$.

Effect of η: Using Table IV, we examine the effect of η on the performance of an agent under dynamic fusion. In theory, a higher η would make it more likely for T_ψ to remain low, and thus the personalised policy to correspond to the intent-specific policy. Our observations in Table IV are in agreement with this behaviour, as in all preference modes, as η increases, the desired lane visit increases and the undesired lane visit decreases, suggesting that personalised policy is dominated by the intent-specific policy component. In all of our experiments in other environments, we choose $\eta = 0$. However, an ideal choice of η may deviate depending on the environment.

TABLE V
PERFORMANCE OF DYNAMIC FUSION ($\eta = 0$) VS DQN IN THE PONG ENVIRONMENT.

Method	% Desired region ↑	Score ↑
DQN	8.00 ± 0.63	$\mathbf{0.57 \pm 0.01}$
Dynamic	$\mathbf{55.60 \pm 4.48}$	0.47 ± 0.02

Effect of T_{max}: The parameter T_{max} represents the maximum temperature in Equation (6). Theoretically, as T_{max} is increased the intent-specific policy should weaken. The observation in Table VI supports this notion that when T_{max} is increased, the desired lane visit decreases and the undesired lane visit increases suggesting a declining influence of intent-specific policy.

TABLE VI
SENSITIVITY ANALYSIS OF T_{max} IN HIGHWAY ENVIRONMENT MIXED CASE.

T_{max}	Desired lane↑	Undesired lane↓	hits↓	Score↑
10	19.21 ± 1.42	2.02 ± 0.43	0.06 ± 0.02	40.12 ± 0.59
15	17.60 ± 0.90	3.42 ± 0.43	0.09 ± 0.02	39.64 ± 0.45
25	9.53 ± 0.82	4.65 ± 0.52	0.10 ± 0.02	39.51 ± 0.53
35	14.36 ± 0.65	5.32 ± 0.58	0.10 ± 0.02	40.03 ± 0.59

VI. CONCLUSION

We proposed a novel approach to personalise a trained policy to respect human-specified preferences using trajectory-level human feedback, without any further interactions with the environment. Using an LSTM-based approach to infer human intent, we designed a dynamic policy fusion method to ensure the resulting policy completes a given task while also respecting human preferences. We empirically evaluated our approach on the Highway, 2D Navigation and Pong environments and demonstrated that our approach is capable of handling various modes of intent while only minimally compromising the task performance. We believe our approach presents an elegant and scalable solution to the problem of personalising pretrained policies. With a growing focus on personalisation in applications such as chatbots, robotic assistants, self-driving vehicles, etc., we believe our approach has the potential for imminent and widespread impact.

REFERENCES

[1] Han, Dong, Beni Mulyana, Vladimir Stankovic, and Samuel Cheng. "A Survey on Deep Reinforcement Learning Algorithms for Robotic Manipulation." Sensors 23, no. 7 (2023): 3762.

[2] Arjona-Medina, Jose A., Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. "Rudder: Return decomposition for delayed rewards." Advances in Neural Information Processing Systems 32 (2019).

[3] Christiano, Paul F., Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. "Deep reinforcement learning from human preferences." Advances in neural information processing systems 30 (2017).

[4] Sestini, Alessandro, Alexander Kuhnle, and Andrew D. Bagdanov. "Policy fusion for adaptive and customizable reinforcement learning agents." In 2021 IEEE Conference on Games (CoG), pp. 01-08. IEEE, 2021.

[5] Perez, Ethan, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. "Film: Visual reasoning with a general conditioning layer." In Proceedings of the AAAI conference on artificial intelligence, vol. 32, no. 1. 2018.

[6] Leurent, Edouard, "An Environment for Autonomous Driving Decision-Making," GitHub, GitHub repository, 2018.

[7] Tasfi, Norman, "PyGame Learning Environment." GitHub, GitHub repository, 2016.

[8] Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. "Playing atari with deep reinforcement learning." arXiv preprint arXiv:1312.5602 (2013).

[9] Yu, Chao, Jiming Liu, Shamim Nemati, and Guosheng Yin. "Reinforcement learning in healthcare: A survey." ACM Computing Surveys (CSUR) 55, no. 1 (2021): 1-36.

[10] Sun, Yueming, and Yi Zhang. "Conversational recommender system." In The 41st International ACM Sigir conference on research and development in information retrieval, pp. 235-244. 2018.

[11] Roijers, Diederik M., Peter Vamplew, Shimon Whiteson, and Richard Dazeley. "A survey of multi-objective sequential decision-making." Journal of Artificial Intelligence Research 48 (2013): 67-113.

[12] Guan, Lin, Mudit Verma, and Subbarao Kambhampati. "Explanation augmented feedback in human-in-the-loop reinforcement learning." arXiv preprint arXiv:2006.14804 (2020).

[13] Lee, Kimin, Laura Smith, and Pieter Abbeel. "Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training." arXiv preprint arXiv:2106.05091 (2021).

[14] Basaklar, Toygun, Suat Gumussoy, and Umit Y. Ogras. "Pd-morl: Preference-driven multi-objective reinforcement learning algorithm." arXiv preprint arXiv:2208.07914 (2022).

[15] Haarnoja, Tuomas, Vitchyr Pong, Aurick Zhou, Murtaza Dalal, Pieter Abbeel, and Sergey Levine. "Composable deep reinforcement learning for robotic manipulation." In 2018 IEEE international conference on robotics and automation (ICRA), pp. 6244-6251. IEEE, 2018.

[16] Hunt, Jonathan, Andre Barreto, Timothy Lillicrap, and Nicolas Heess. "Composing entropic policies using divergence correction." In International Conference on Machine Learning, pp. 2911-2920. PMLR, 2019.

[17] Mo, Kaixiang, Yu Zhang, Shuangyin Li, Jiajun Li, and Qiang Yang. "Personalizing a dialogue system with transfer reinforcement learning." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1. 2018.

[18] Bodas, Anand, Bhargav Upadhyay, Chetan Nadiger, and Sherine Abdelhak. "Reinforcement learning for game personalization on edge devices." In 2018 International Conference on Information and Computer Technologies (ICICT), pp. 119-122. IEEE, 2018.

APPENDIX A
SIMULATING HUMAN FEEDBACK

Our method involves collecting human feedback for trajectories used for training the task-specific policy. This is done by simulating humans. We conducted experiments in three personalisation modes such as *preference*, *Avoidance* and *Mixed*. In each case, we counted the number of times the agent met the personalisation criteria within a trajectory and this is regarded as the score against the trajectory. For instance, In 2D navigation, if the human wishes the agent to visit a preferred region, we count the number of times the agent visited that state and a positive score is given to that trajectory. Similarly, in the avoidance case, a negative score is given and in the mixed case, we sum both the positives and the negatives. To train LSTM, we randomly sampled a portion of the total trajectory data.

APPENDIX B
OTHER CHOICES OF POLICY FUSION

In this section, we focus on analyzing two common fusion methods: the weighted average and product fusion methods.

The former method is defined as $\alpha\pi_1+(1-\alpha)\pi_2$, where π_1 and π_2 are the policies to be fused, and ζ is a weight parameter. This method satisfies the invariability property, ensuring that the fused policy remains unchanged when both input policies are similar. However, this fusion strategy acts on the union of the support of the input policies, meaning that actions are selected to maximize either one objective and not both simultaneously.

The latter method involves taking the product of the two policies: $\pi_f(a|s) = \pi_1(a|s) \times \pi_2(a|s)$. While this fusion operates on the intersection of the support of the input policies, ensuring that actions are selected to maximize both objectives simultaneously, it fails to satisfy the invariability property as shown below,

Lemma 2. *Let π_ψ and π_ϕ be intent-specific policy and task-specific policies respectively and π_ψ is not a random policy. Then π_ϕ and $\pi_\psi\pi_\phi$ are not invariant policies in any condition.*

Proof.

$$\begin{aligned} KL\left(\pi_\phi|\pi_\phi\pi_\psi\right) &= \sum_a \pi_\phi \log\left(\frac{\pi_\phi Z}{\pi_\phi\pi_\psi}\right) \\ &= \sum_a \pi_\phi \log\left(\frac{Z}{\pi_\psi}\right) \\ &= \log\left(Z\right) - \sum_a \pi_\phi \log \pi_\psi \end{aligned}$$

Where Z is the normalising factor, $Z = \sum_a \pi_\phi\pi_\psi$. Assume the expression above is $0 \Rightarrow \log\left(Z\right) = \sum_a \pi_\phi \log \pi_\psi$

$$\begin{aligned} &\log\left(\textstyle\sum_a \pi_\phi\pi_\psi\right) = \textstyle\sum_a \pi_\phi \log \pi_\psi \\ &\Rightarrow \log\left(E_{\pi_\phi}\left(\pi_\psi\right)\right) = E_{\pi_\phi}\left(\log\left(\pi_\psi\right)\right) \end{aligned}$$

This expression is valid only if π_ϕ is a random policy, which is a contradiction. □

The above Lemma established that even if two policies are the same, the KL-divergence between the task-specific policy and the fused policies produced by the product is not 0 thereby not satisfying the invariability property.

However, the fused policies produced by the product can be bounded as captured by the following theorem,

Theorem 3. *Let Q and Q' be Q-values corresponding to task-objective and human intent objectives. Let $\pi_\psi(a|s)$ and $\pi_\phi(a|s)$ represent the respective policies, with corresponding temperatures T_ψ and T_ϕ. Let $\left\|Q(s,a) - Q'(s,a)\right\|_2 < \epsilon\ \forall s \in \mathcal{S}$ and $a \in \mathcal{A}$ and $\left\|T_\psi - T_\phi\right\|_2 < \delta$. Then,*

$$\begin{aligned} &KL\left(\pi_\phi(a|s)\|\pi_\phi\pi_\psi\right) \leq \log\left(Z\right) + \left(\frac{Q^*\delta + \epsilon T_\phi}{T_\phi T_\psi}\right) \\ &+ \log\left(\zeta\right)\ \ \forall a \in \mathcal{A}, s \in \mathcal{S}, \end{aligned}$$

where Z is the normalising factor of π_f, $Q^* = Q(s, argmax_{a\in\mathcal{A}}Q(s,a))$ and $\zeta = \frac{h(Q',T_\psi)}{h(Q,T_\phi)}$, here $h(Q,T) = \sum_a exp\left(\frac{Q}{T}\right)$.

Proof.

$$\begin{aligned} &KL\left(\pi_\phi\|\pi_\psi\pi_\phi\right) = \sum_a \pi_\phi \log\left(\frac{\pi_\phi Z}{\pi_\psi\pi_\phi}\right) \\ &= \log\left(Z\right) + \sum_a \pi_\phi \left(\log\left(e^{\frac{Q}{T_\phi} - \frac{Q'}{T_\psi}}\right) + \log\left(\zeta\right)\right) \\ &= \log\left(Z\right) + \sum_a \pi_\phi \left(\frac{Q}{T_\phi} - \frac{Q'}{T_\psi} + \log\left(\zeta\right)\right) \\ &= \log\left(Z\right) + \sum_a \pi_\phi \left(\frac{Q}{T_\phi} - \frac{Q'}{T_\psi}\right) + \log\left(\zeta\right) \\ &= \log\left(Z\right) + \sum_a \pi_\phi \left(\frac{QT_\psi - Q'T_\phi}{T_\phi T_\psi}\right) + \log\left(\zeta\right) \\ &= \log\left(Z\right) + \sum_a \pi_\phi \left(\frac{QT_\psi - Q'T_\phi + QT_\phi - QT_\phi}{T_\phi T_\psi}\right) + \log\left(\zeta\right) \\ &\leq \log\left(Z\right) + \sum_a \pi_\phi \left(\frac{Q\delta + \epsilon T_\phi}{T_\phi T_\psi}\right) + \log\left(\zeta\right) \\ &\leq \log\left(Z\right) + \left(\frac{Q^*\delta + \epsilon T_\phi}{T_\phi T_\psi}\right) + \log\left(\zeta\right) \end{aligned}$$

□

The proof for the theorem stated in the main paper can be obtained with minor changes in the above proof.

APPENDIX C
ENVIRONMENTS

We conducted experiments in three environments namely 2D Navigation. Highway [6] and Pong [7] as shown in Figure 2. 2D Navigation is an image-based input and we trained a Variational Autoencoder(VAE) to encode the state-space which

reduces the dimensionality. The other two environments are feature-based.

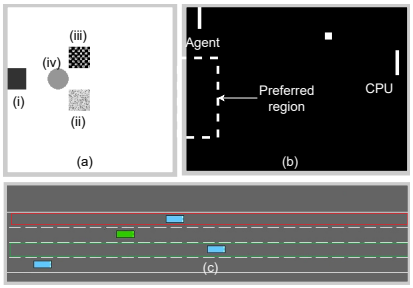

Fig. 2. Snippets from the 2D environment (a), Pong (b) and Highway (c). (i), (ii), (iii), (iv) marked in (a) correspond to the target, desired region, undesired region and agent respectively. In Pong (b), the right paddle is controlled by the CPU and the left is controlled by the agent. The preferred region corresponds to the center, as indicated. In the Highway environment (c), lanes marked with green and red boxes represent desired and undesired lanes respectively.

APPENDIX D
TEMPERATURE MODULATION

We use Equation (6) to modulate the temperature of intent-specific policy. The graph of T_ψ is shown in Figure 3. The accumulated human-induced reward means the agent has not strictly followed the human intent so far which forces the temperature of the intent-specific policy to be less. A low temperature in Boltzmann distribution means a sharp distribution and while fusing this favours more towards intent-specific policy. On the other hand, when the accumulated reward is high, the temperature is also high diminishing the effect of intent-specific policy.

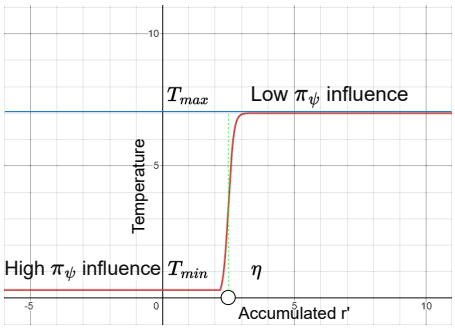

Fig. 3. Plot showing the variation of T_ψ. The temperature rises when the accumulated r' reaches η