# Identifying Latent State-Transition Processes for Individualized Reinforcement Learning

Yuewen Sun[1,2], Biwei Huang[3], Yu Yao[4], Donghuo Zeng[5], Xinshuai Dong[2], Songyao Jin[3], Boyang Sun[1], Roberto Legaspi[5], Kazushi Ikeda[5], Peter Spirtes[2], Kun Zhang[1,2]

[1]Mohamed bin Zayed University of Artificial Intelligence, [2]Carnegie Mellon University, [3]University of California San Diego, [4]The University of Sydney, [5]KDDI Research

yuewen.sun@mbzuai.ac.ae

# Background

- **What is Reinforcement Learning (RL)?**
  - RL is a method where agents learn to make decisions by interacting with an environment
  - The agent observes a current state, takes an action, and transitions to a new state, receiving a reward

- **Why is individualization crucial?**
  - Individualized RL tailors decisions based on unique characteristics, like preferences or physiological traits, which affect state transitions
  - Healthcare
    - Individual-specific factors like genetic makeup impact responses to treatment
    - Identifying these unique factors can personalize treatment plans, leading to improved health outcomes
  - Education
    - Differences in learning styles (e.g., visual vs. hands-on) affect how students absorb information
    - RL can use these insights to recommend tailored learning activities, enhancing educational effectiveness

# Motivation

- **Individual-specific factors in RL are often latent and unobserved**
  - Patient's genetic traits may impact their response to treatment but remain hidden from observation
  - Challenge to fully understand each individual's unique influence on state transitions
- **Identify these latent factors can better optimize personalized policies**
  - RL can tailor educational content to suit each student, improving learning outcomes
  - Allow RL systems to adapt more effectively to individual needs and improve outcomes
- **Contributions**
  - Introduction of Individualized Markov Decision Processes
  - Theoretical guarantees for identifying individual-specific latent factors
  - Practical generative method to estimate these factors and optimize policies

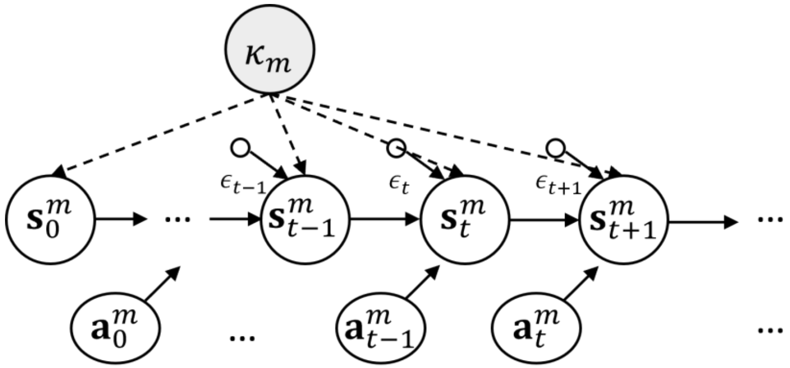
# Problem Formulation

## ■ Individualized Markov Decision Process (iMDP)

- ☐ *State and Action Spaces*: Common across individuals
- ☐ *Individual-specific Factor ($\kappa$)*: A latent variable unique to each group that influences state transitions
- ☐ *Group and Individual Uniqueness*: Individuals are grouped based on shared latent factors, while each individual has unique identifiers

## ■ Objective

- ☐ Identify latent individual-specific factors $\kappa$ from observed trajectories
- ☐ Derive individualized policies for each agent and realize policy adaptation for newcomers




iMDP for individual $m$

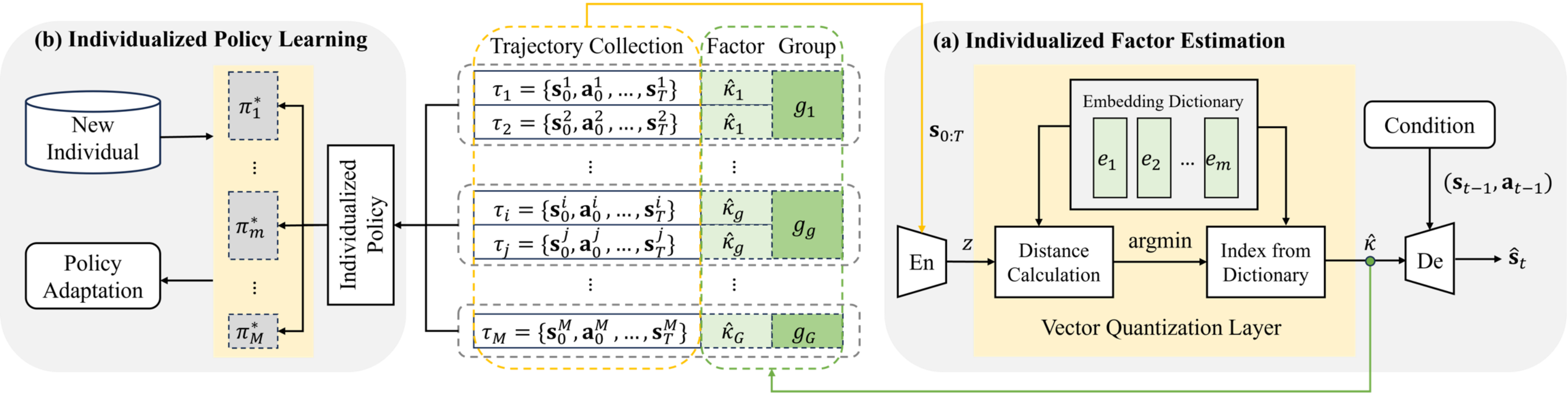
# Identifiability Theorem

- **Purpose**
  - Guarantee that latent individual-specific factors $\kappa$ can be uniquely identified from observed trajectories

- **Key Conditions**
  - *Finite Latent Factors*: Identifiability guaranteed if the latent factor $\kappa$ has a finite set of values and individuals are grouped accordingly
  - *Infinite Latent Factors*: For complex cases with infinite or continuous latent factors, identifiability is achieved under rank deficiency within post-nonlinear temporal model

# Methodology

- **1st Phase: Latent Factor Estimation**
  - *Objective*: Identify latent individual-specific factors that impact state transitions
  - *Process*: Encode trajectories into latent representations and quantize using an embedding dictionary
  - Provides a foundation for personalized policy adaptation by capturing unique, unobserved influences
- **2nd Phase: Individualized Policy Learning**
  - *Objective*: Develop policies tailored to individual characteristics
  - *Process*: Initialize policy using latent factors, then adapt through online interaction new individual
  - Enhances policy effectiveness by aligning decisions with each individual's unique traits



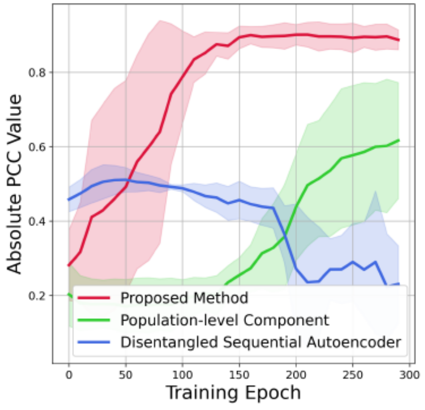
# Experiment Result: Latent Factor Estimation

## ■ Conclusion

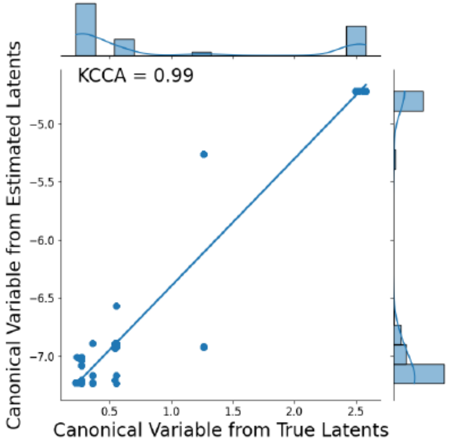
- ☐ Our method effectively estimates latent factors with strong correlation to true values, supporting reliable individualization

## ■ Results

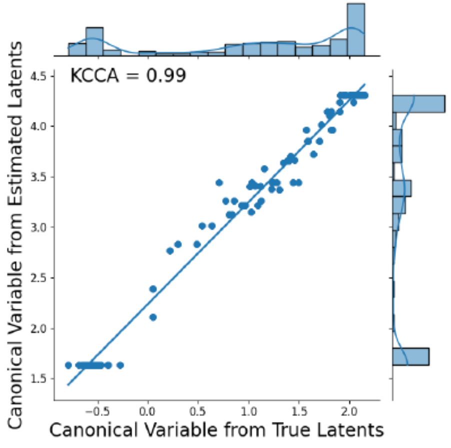
- ☐ *Fig (a)*: Our method achieves higher PCC values over time, outperforming baselines
- ☐ *Fig (b-c):* Kernel Canonical Correlation Analysis (KCCA) scatterplots indicate a near-perfect correlation between estimated and true latent factors
- ☐ *Fig (d):* Larger sample sizes (M) and trajectory lengths (T) improve identifiability



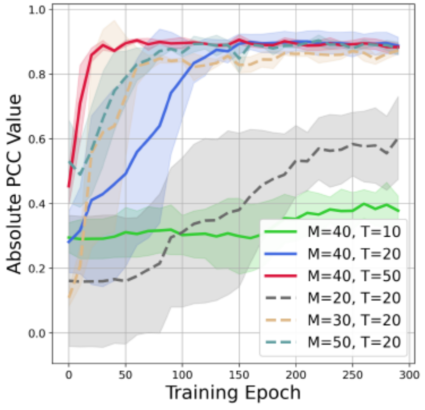

(a)




(b)




(c)




(d)

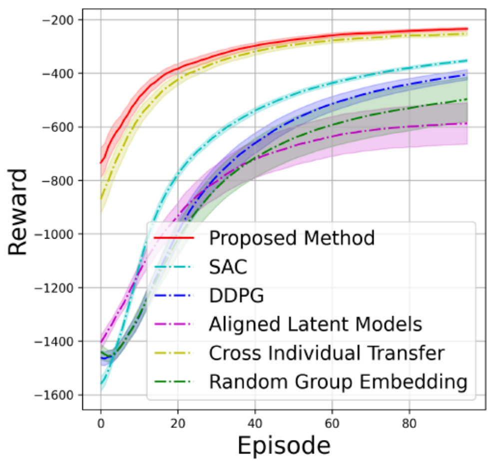
# Experiment Result: Policy Learning Improvement

## ■ Conclusion

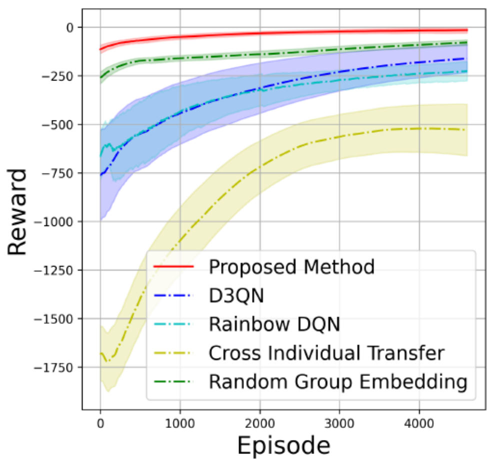
- ☐ The proposed method demonstrates superior policy learning, leading to higher rewards and faster convergence across tasks

## ■ Results

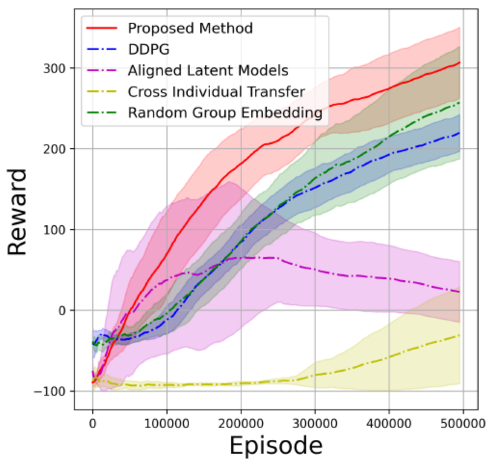
- ☐ *Pendulum*: Proposed method achieves the highest rewards and faster convergence across episodes
- ☐ *HeartPole*: Consistently outperforms other methods with higher rewards
- ☐ *Half Cheetah*: Significant reward improvements and rapid convergence over time




Pendulum




HeartPole




Half Cheetah

# Summary

- Our work
  - New approach for individualized reinforcement learning with theoretical guarantees
  - Successfully estimates latent factors, supporting personalized policy optimization

- Limitations
  - Does not address instantaneous causal influences within states
  - Lacks nonparametric proof for continuous latent factors
  - Does not account for time-varying latent factors

# Thank you for your listening!