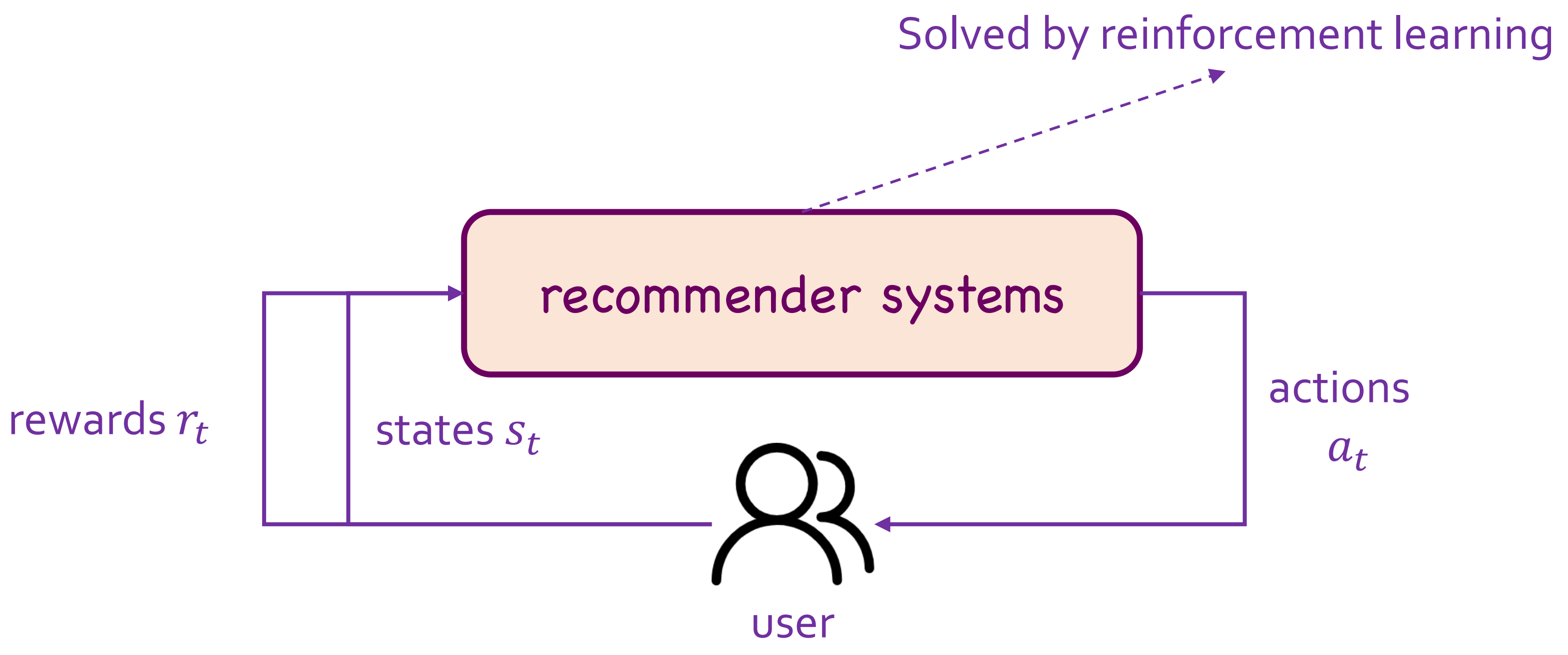
Sim2Rec: A Simulator-based Decision-making Approach to Optimize Real-World Long-term User Engagement in Sequential Recommender Systems

Xiong-Hui Chen[1,3], Yang Yu[1,3,*], Qingyang Li[2], Bowei He[4], Zhiwei Qin[2], Wenjie Shang[2], Jieping Ye[2]
[1] National Key Laboratory of Novel Software Technology, Nanjing University, Nanjing, China
[2] Didi Chuxing, [3] Polixir.ai, [4] City University of Hong Kong
[*] Corresponding author
chenxh@lamda.nju.edu.cn, yuy@nju.edu.cn, qingyangli@didiglobal.com,boweihe2-c@my.cityu.edu.hk

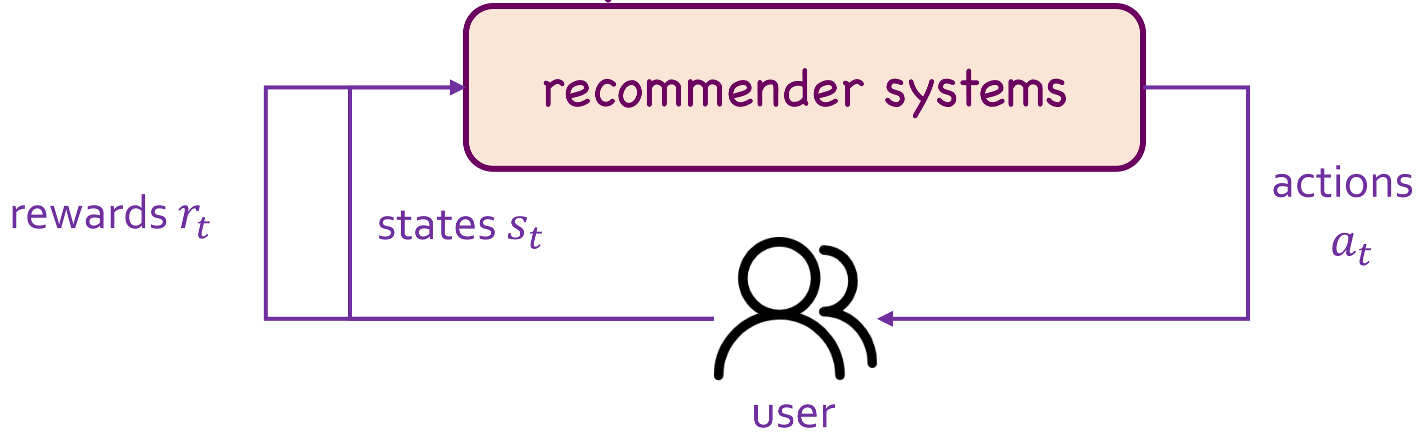
Outline

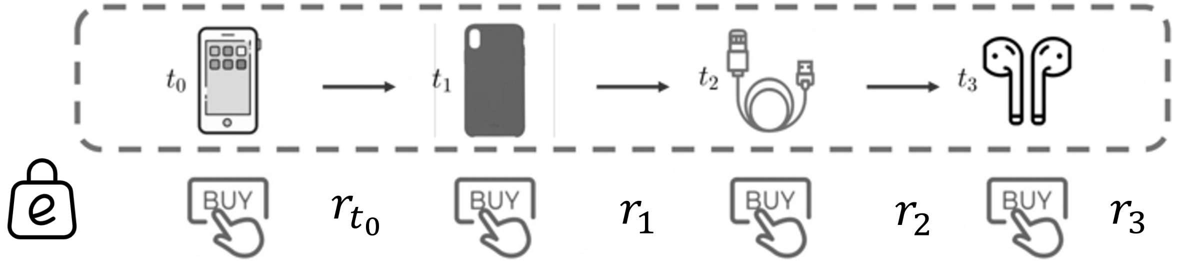
1. **Background and Motivation**
2. Simulation to Recommender Systems (Sim2Rec)
3. Experiment
4. Take-home Messages

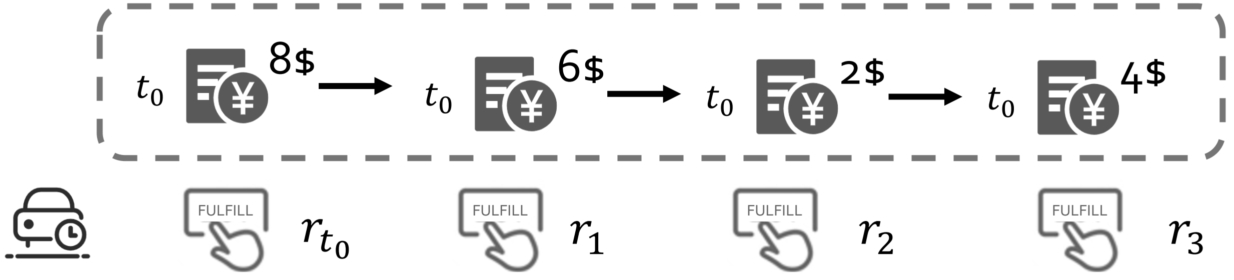
Reinforcement Learning (RL) to Optimize Real-World Long-term User Engagement (LTE) in Sequential Recommender Systems (SRS)



Reinforcement Learning (RL) to Optimize Real-World Long-term User Engagement (LTE) in Sequential Recommender Systems (SRS)







product recommendation in e-commerce platforms

order dispatching in ride-hailing platforms

objective: maximize cumulative user engagement instead of immediate rewards

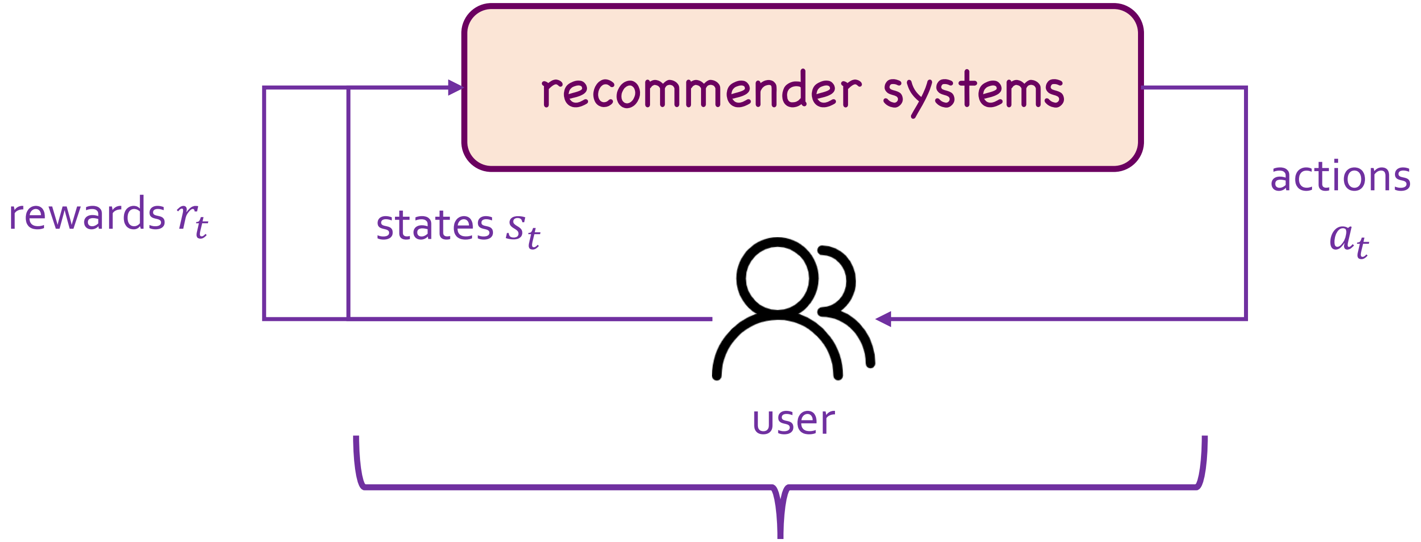
Learning And Mining from DatA

NANJING UNIVERSITY

POLIXIR 南栖仙策

Reinforcement Learning (RL) to Optimize Real-World Long-term User Engagement (LTE) in Sequential Recommender Systems (SRS)

objective of RL for SRS: maximize cumulative rewards instead of immediate rewards



need massive online interactions for exploration to find an optimal recommender policy.

⇩

costly: time-consuming and unsafe (might cause economic losses) in many real-world applications

Learning And Mining from DatA

南京大學 NANJING UNIVERSITY

POLIXIR 南栖仙策

Simulation-based RL to Optimize LTE in SRS

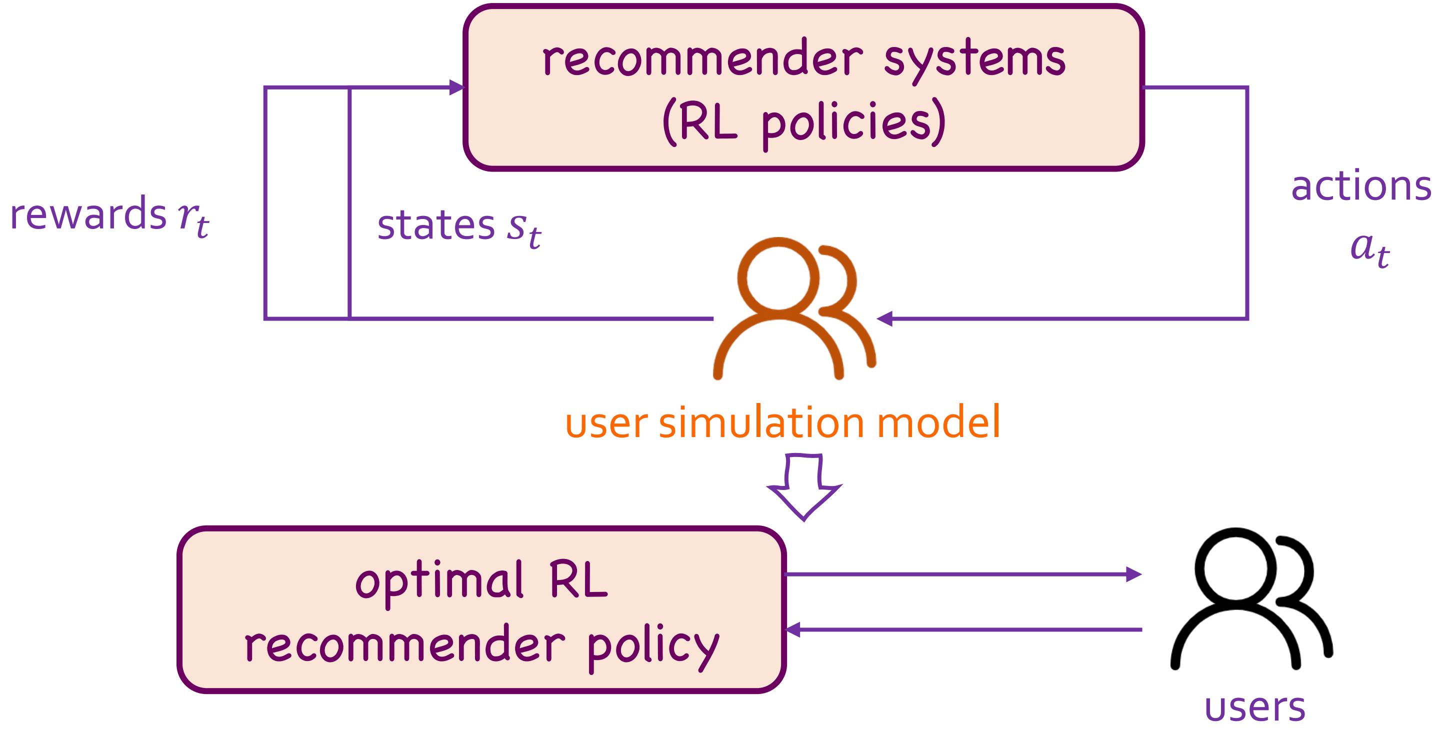
objective of RL for SRS: maximize cumulative rewards instead of immediate rewards

-> Need massive online interactions for exploration to find an optimal recommender policy, which is costly.

✅ ideal solution: simulation-based RL for SRS

1. training RL policies in customer simulation models learned by machine learning techniques.

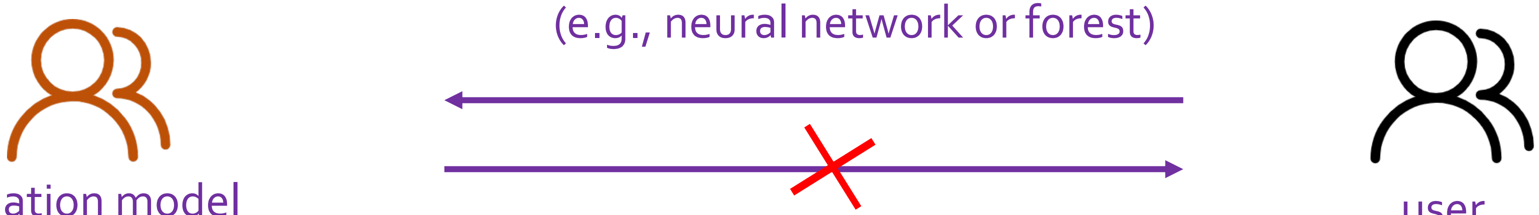
2. deploy after the policy learned to obtain the optimal cumulative rewards in the simulation models.



The Reality-gaps Challenge in Simulation-based RL Workflow

objective of simulation-based RL for SRS: maximize cumulative rewards in the simulation models

Learned by machine learning techniques
(e.g., neural network or forest)



user simulation model

user

inevitably suffer prediction error
to the real world

the discrepancy between simulation and reality, referred to as **the reality-gaps**

The Reality-gaps Challenge in Simulation-based RL Workflow

objective of simulation-based RL for SRS: maximize cumulative rewards in the simulation models

Learned by machine learning techniques
(e.g., neural network or forest)

user simulation model

users

inevitably suffer prediction error
to the real world

the discrepancy between simulation and reality, referred to as **the reality-gaps**

compounding errors in sequential environments.

results in undesired real-world performance degradation of the RL policies learned from the simulators

Learning And Mining from DatA

NANJING UNIVERSITY

POLIXIR
南栖仙策

The Reality-gaps Challenge in Simulation-based RL Workflow

objective of simulation-based RL for SRS: maximize cumulative rewards in the simulation models

Learned by machine learning techniques
(e.g., neural network or forest)

customer simulation models

customers

inevitably suffer prediction error
to the real world

the discrepancy between simulation and reality, referred to as **the reality-gaps**

compounding errors in sequential environments.

results in undesired real-world performance degradation of the RL policies learned from the simulators

rarely been discussed explicitly!

Table of Contents

LAMDA
Learning And Mining from DatA

南京大学
NANJING UNIVERSITY

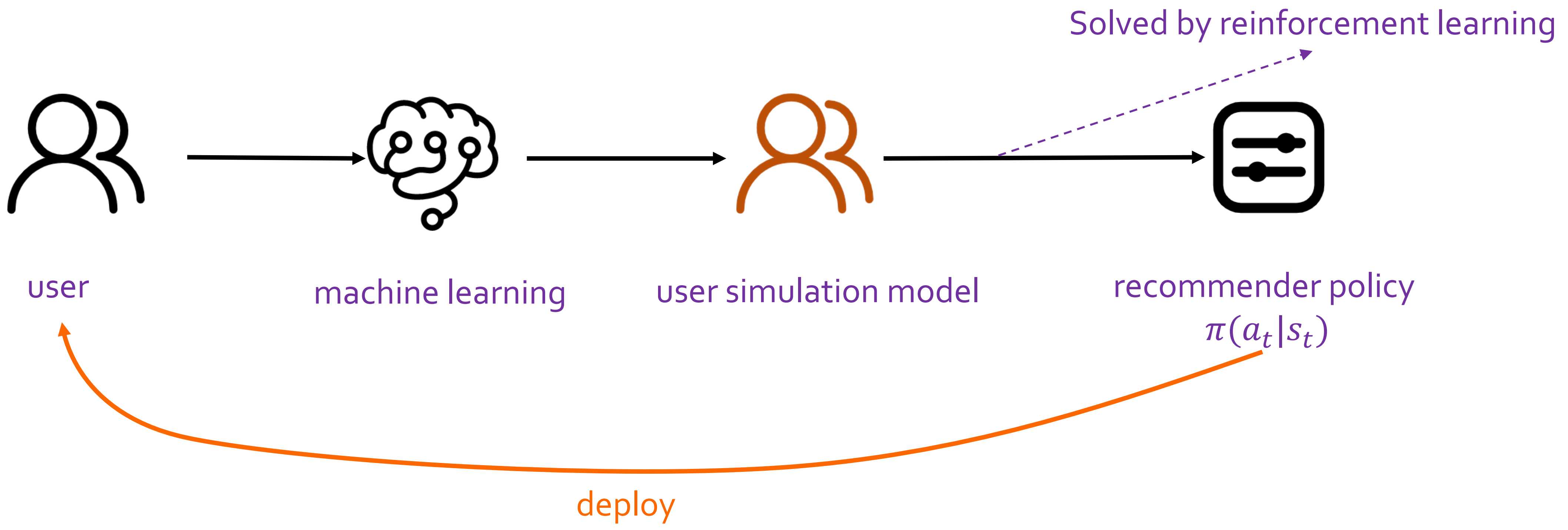
POLIXIR
南栖仙策

Core contributions

1. Introduce **zero-shot policy transfer framework to SRS**, which is a popular solution in Robotics to solve the reality-gap problem between the physical simulator and the real world.

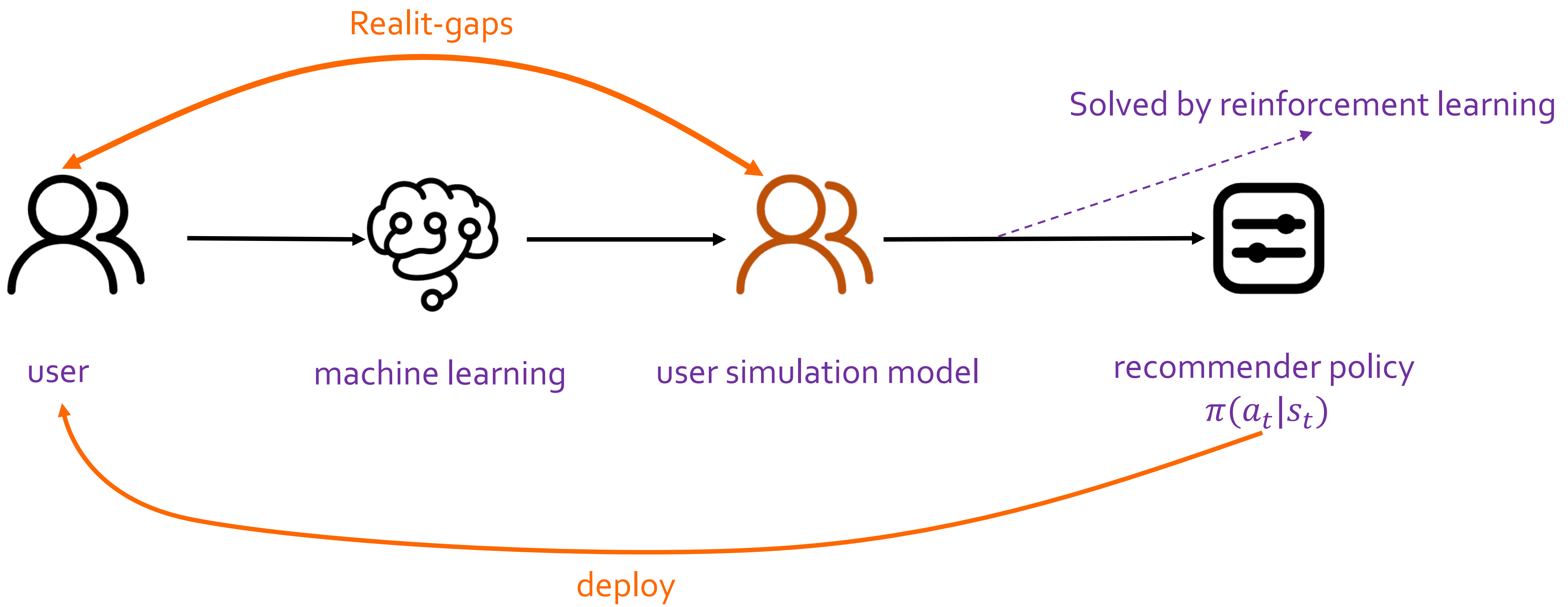
2. We identify and handle several extra challenges when adopting the framework to SRS.

Zero-shot Policy Transfer Framework for SRS: recap

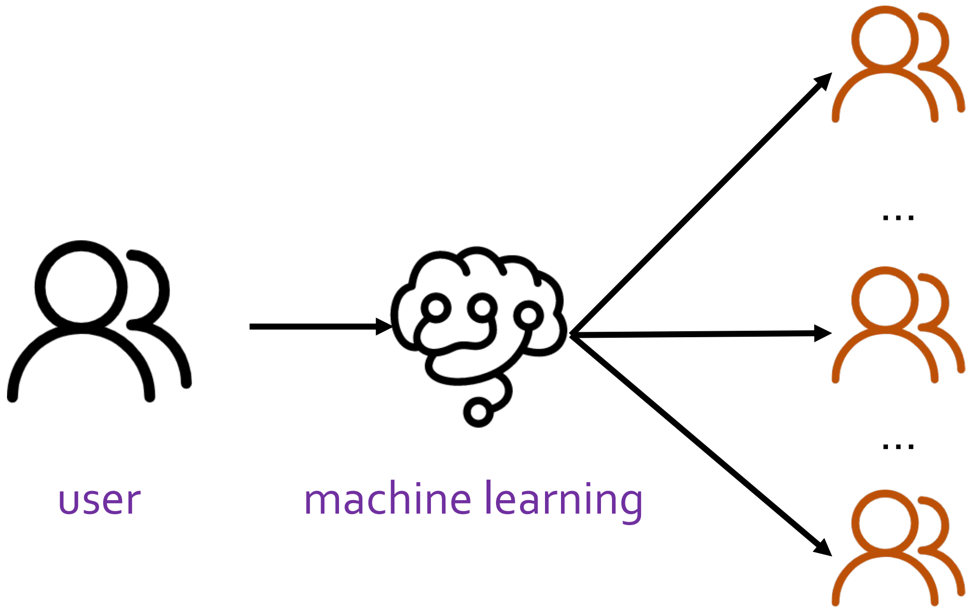


LAMDA Learning And Mining from DatA
南京大學 NANJING UNIVERSITY
POLIXIR 南栖仙策

Zero-shot Policy Transfer Framework for SRS: recap

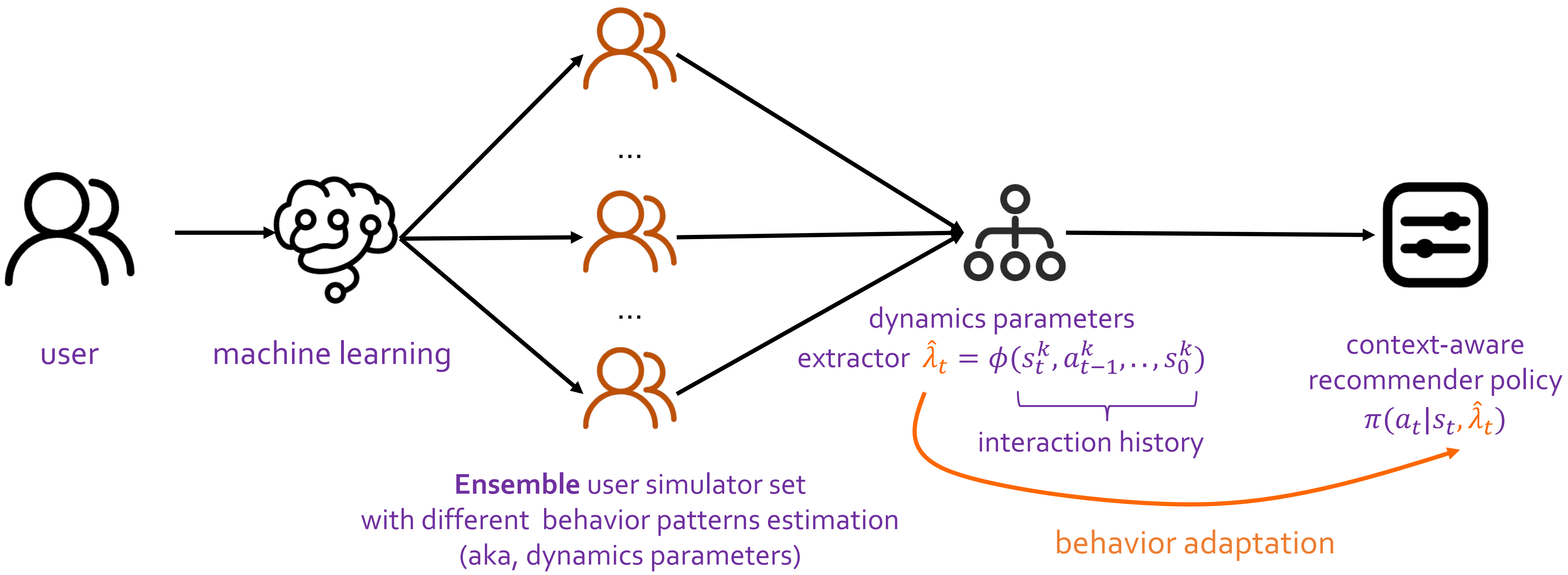


Zero-shot Policy Transfer Framework for SRS

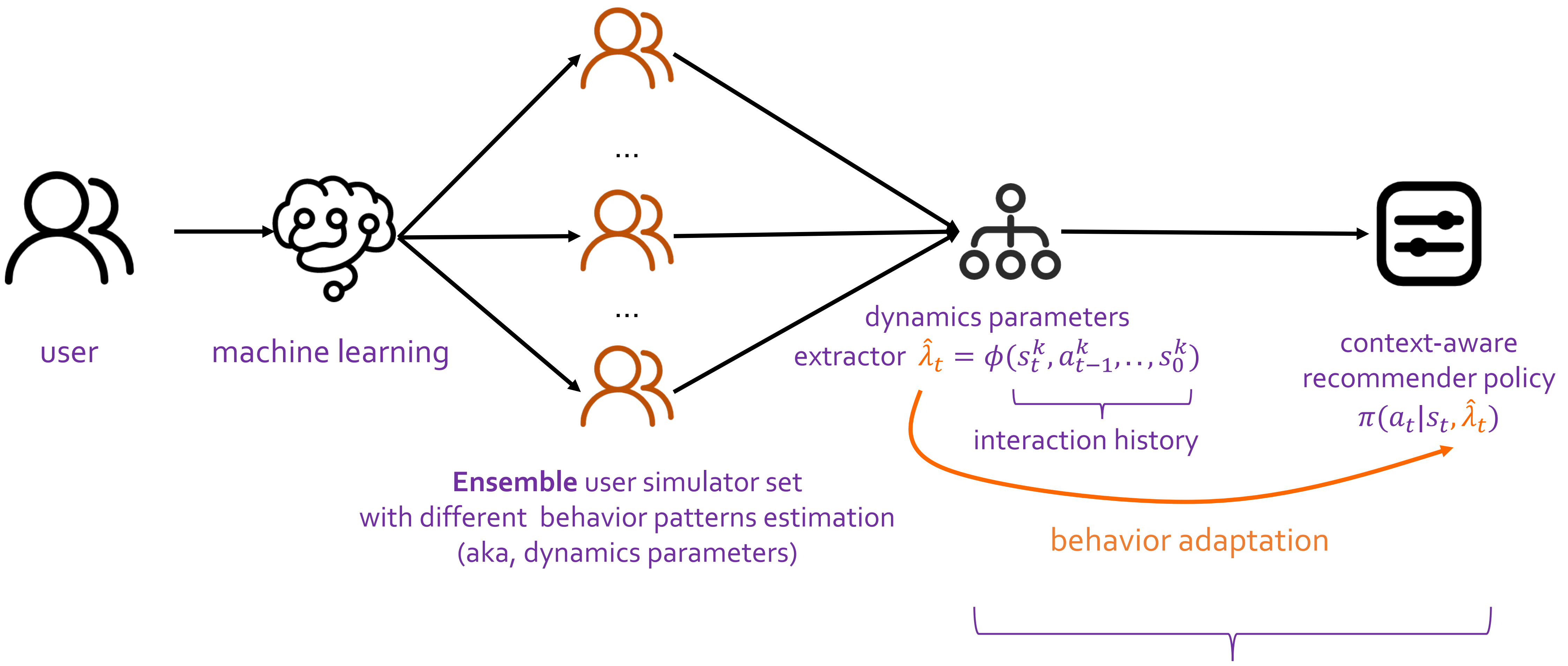


Ensemble user simulator set
with different behavior patterns estimation

Zero-shot Policy Transfer Framework for SRS

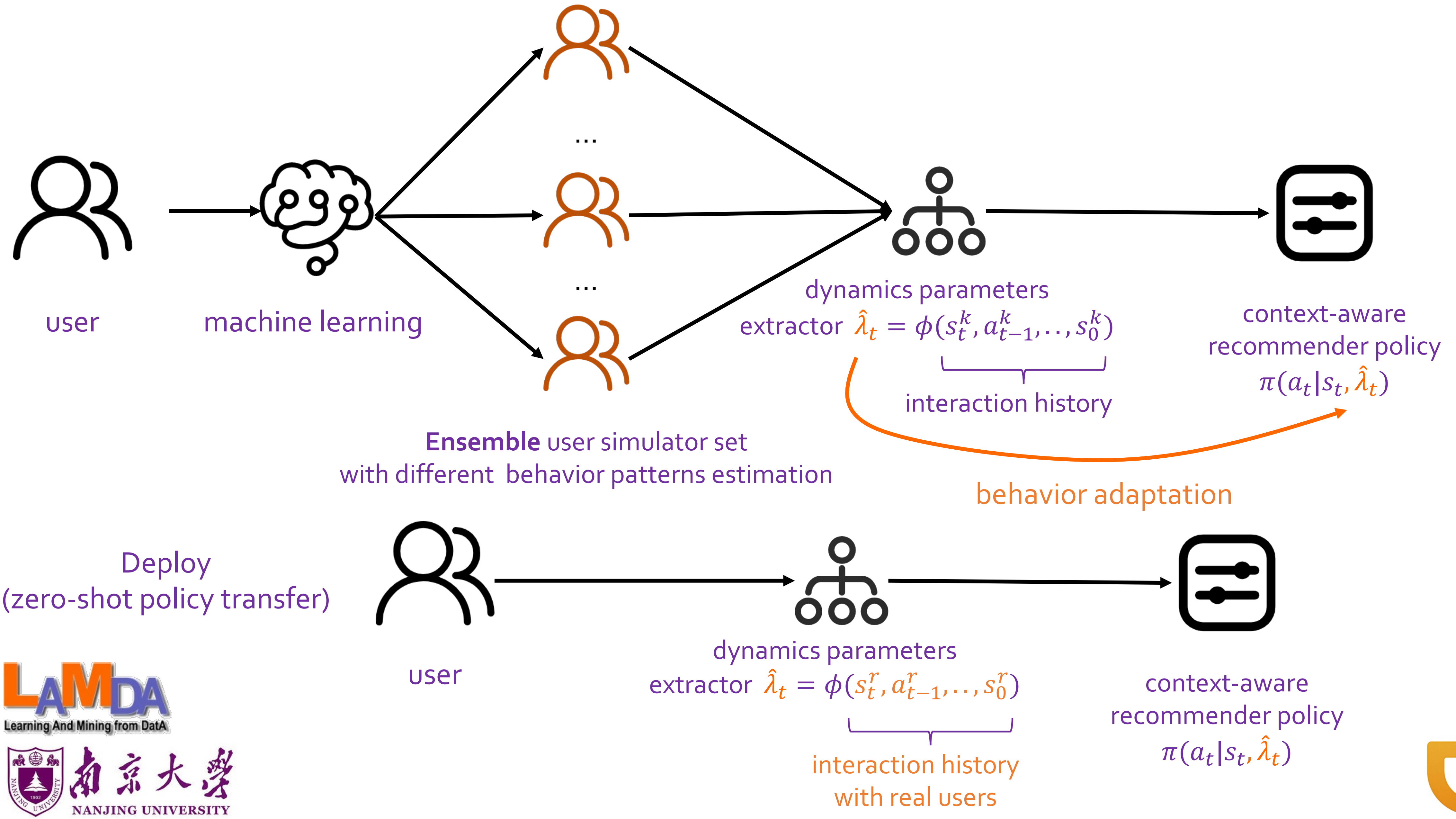


Zero-shot Policy Transfer Framework for SRS



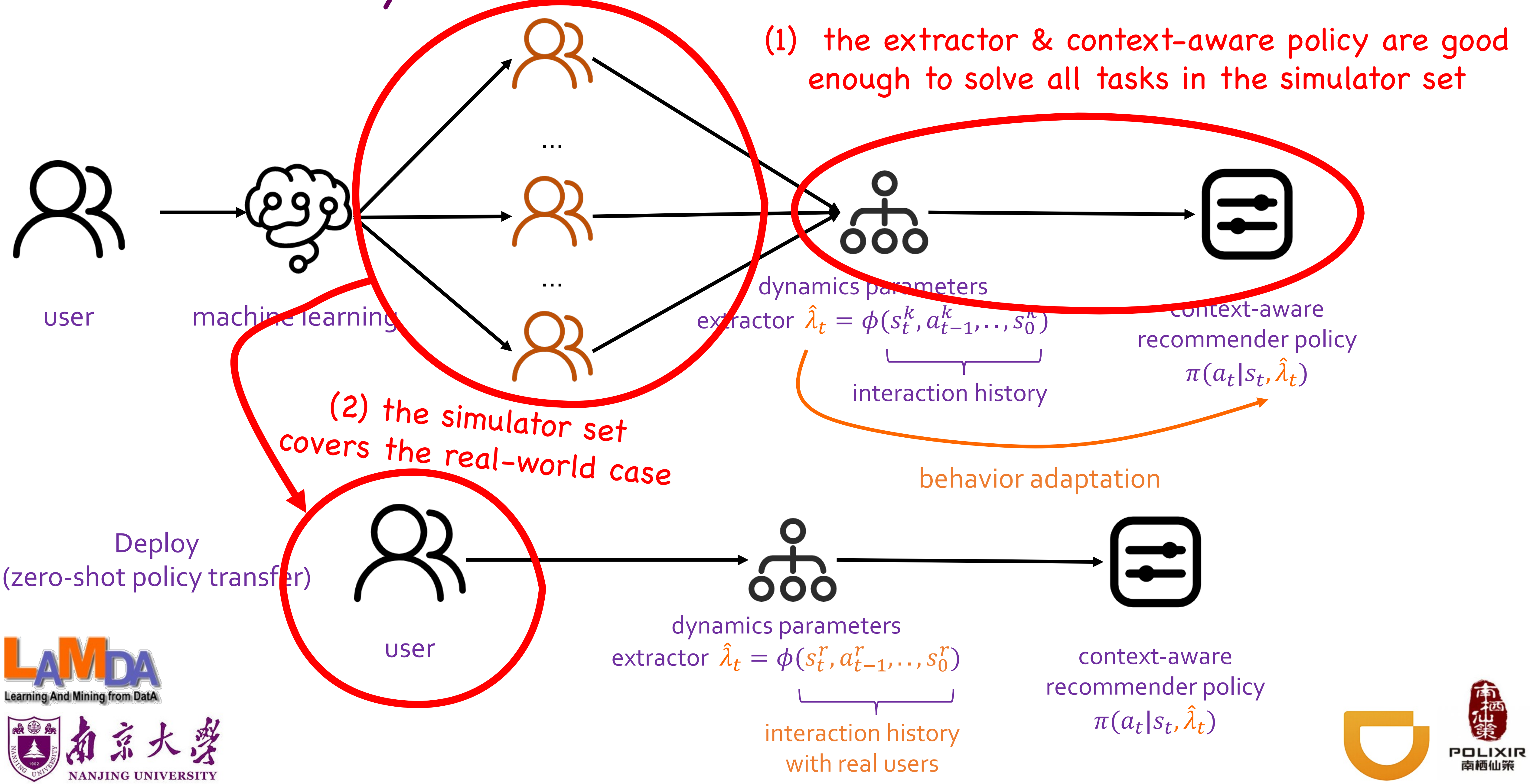
Solved by RL
to maximize the metric in all user simulators

Zero-shot Policy Transfer Framework for SRS



Learning And Mining from DatA
NANJING UNIVERSITY
POLIXIR 南栖仙策

Zero-shot Policy Transfer Framework for SRS



The Challenge of zero-shot policy transfer framework to handle the Reality-gaps in SRS

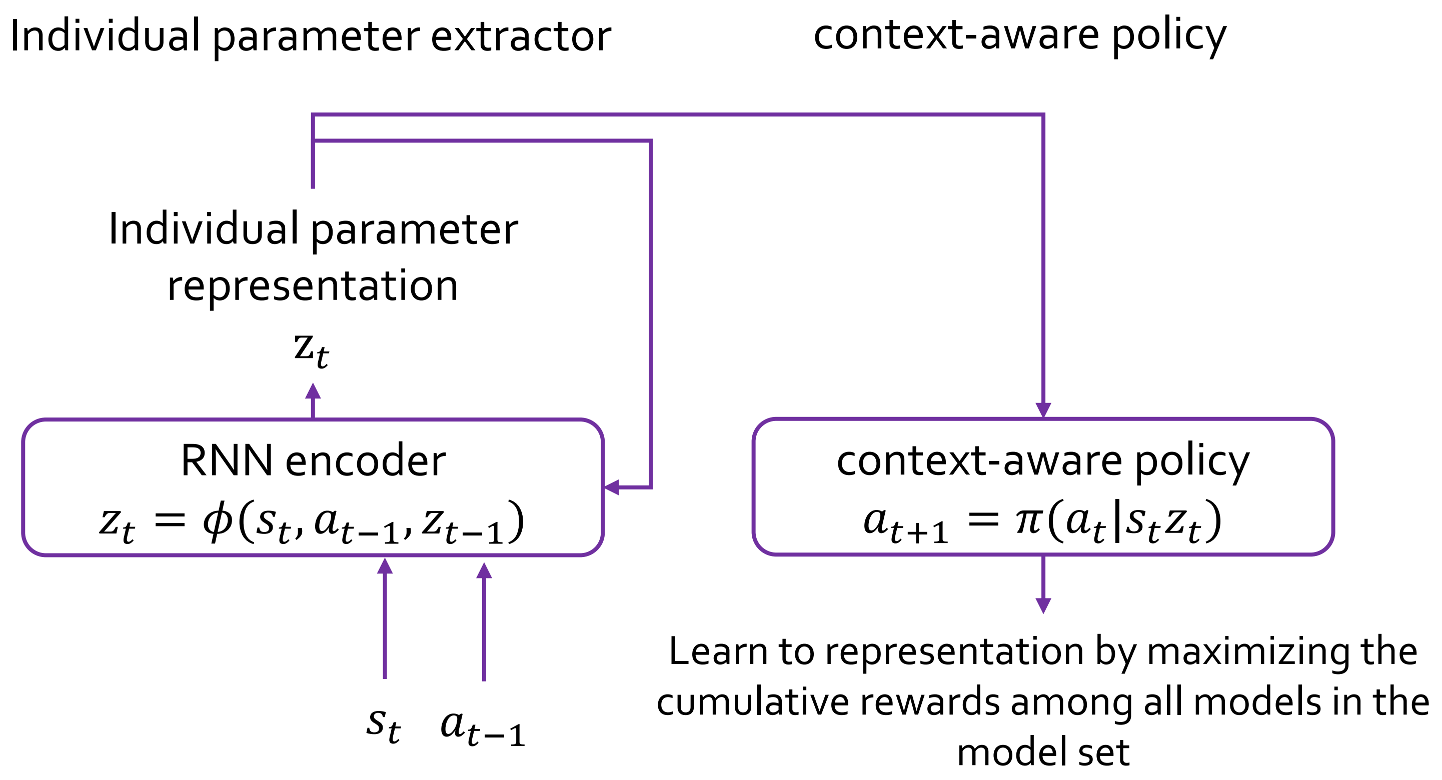
1. The complexity of extractor ϕ to identify correct representations is much larger than in previous applications in robotics.
    - Need to identify **numerous user behavior patterns in different regions** from historical interactions.
        - In robotics, we often only need to infer **a single robot's** dynamics parameter.

    » It is non-trivial to identify the representation of all users in all simulators

2. It is generally impractical to construct a simulator set that can cover all of the real-world user behaviors.
    - Limitation of the presentation capacity of the learning techniques.
    - Or rely on **extremely large size** of ensemble models, which is also impractical.
        - In robotics, simulators are built by physical laws, then we can generate a simulator set with coverability by adjusting the explicitly defined parameters (friction coefficients).

    » Need to build a reliable policy learning paradigm for the simulator set which is uncoverable to reality.

Sim2Rec Solution



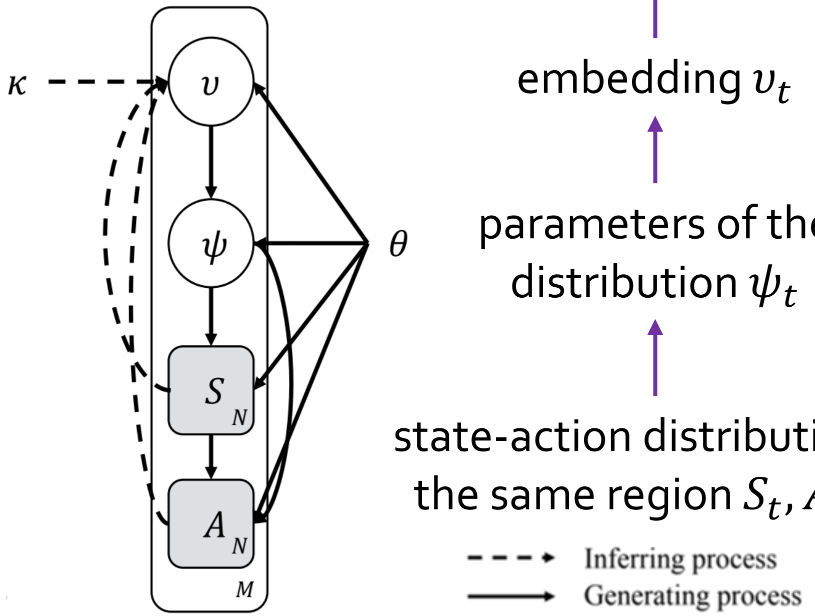
Sim2Rec Solution

(1) efficient representation for extractor learning by utilizing the group information of different users in the same regions

State-Action Distributional variational AutoEncoder (SADAE)

Individual parameter extractor

context-aware policy



embedding υ_t

parameters of the distribution ψ_t

state-action distribution in the same region S_t, A_{t-1}

Individual parameter representation z_t

RNN encoder
$z_t = \phi(s_t, a_{t-1}, \upsilon_t, z_{t-1})$

s_t a_{t-1}

context-aware policy
$a_{t+1} = \pi(a_t | s_t z_t)$

Learn to representation by maximizing the cumulative rewards among all models in the model set

LAMDA Learning And Mining from DatA

南京大學 NANJING UNIVERSITY

POLIXIR 南栖仙策

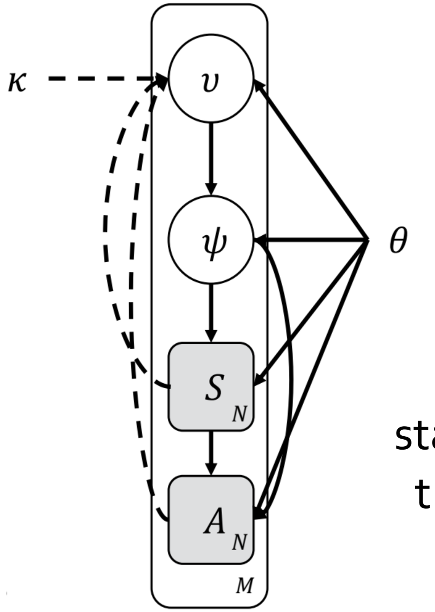
Sim2Rec Solution

(1) efficient representation for extractor learning by utilizing the group information of different users in the same regions

State-Action Distributional variational AutoEncoder (SADAE)

Individual parameter extractor

context-aware policy



embedding v_t

parameters of the distribution ψ_t

state-action distribution in the same region S_t, A_{t-1}

Inferring process

Generating process

Individual parameter representation

z_t

RNN encoder

$z_t = \phi(s_t, a_{t-1}, v_t, z_{t-1})$

s_t a_{t-1}

context-aware policy

$a_{t+1} = \pi(a_t | s_t z_t)$

Learn to representation by maximizing the cumulative rewards among all models in the model set

branching rollout steps to reduce compounding error

reward penalty via inconsistency of ensemble model predictions

pseudo-intervention test to filter out simulation data with severe extrapolation error

(2) reliable policy learning paradigm

Table of Contents

LAMDA Learning And Mining from DatA

南京大學 NANJING UNIVERSITY

POLIXIR 南栖仙策

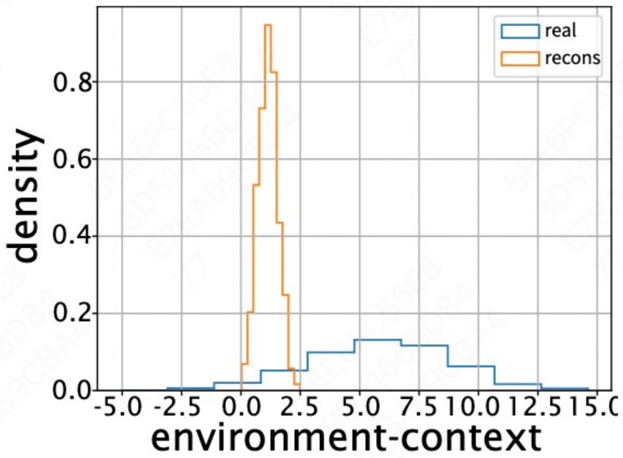
Selected Experiment Results

RecSim

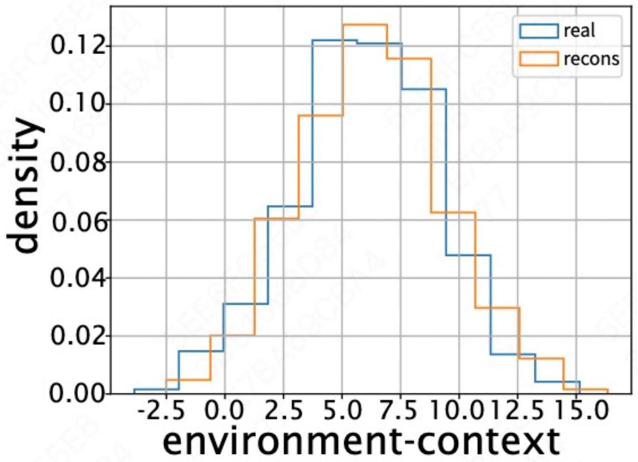
Real-world
(Didichuxing)

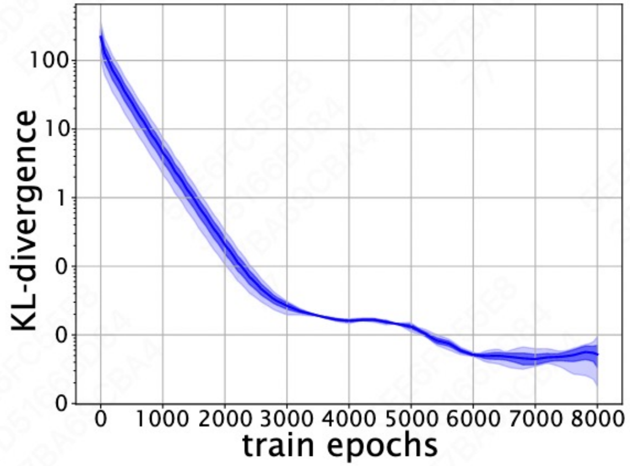
Selected Experiment Results

init



trained

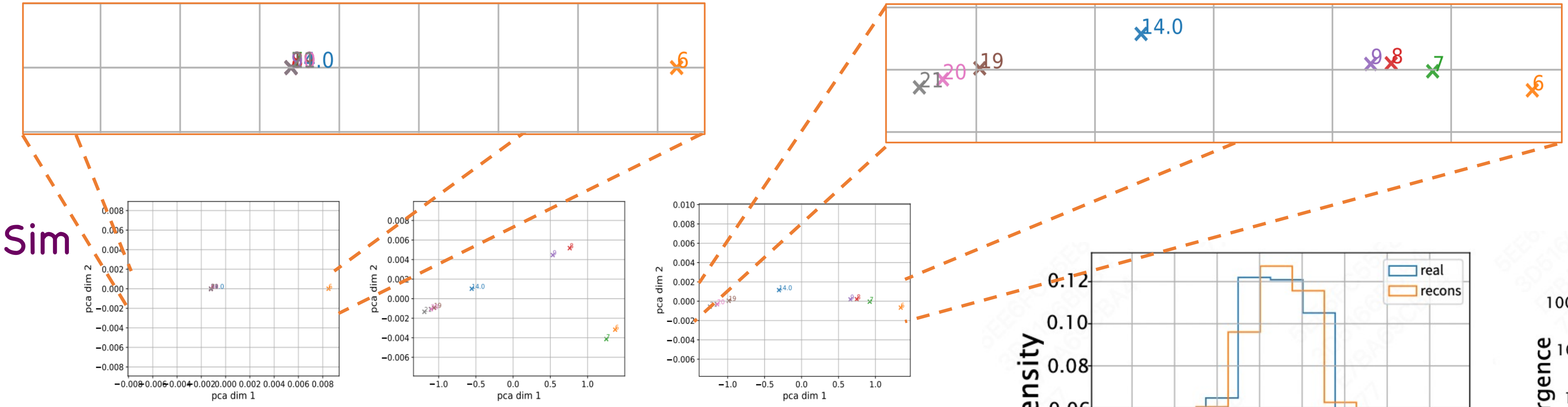




Real-world
(Didichuxing)

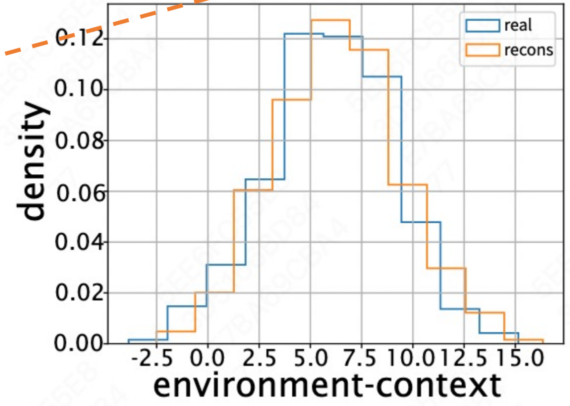
Selected Experiment Results

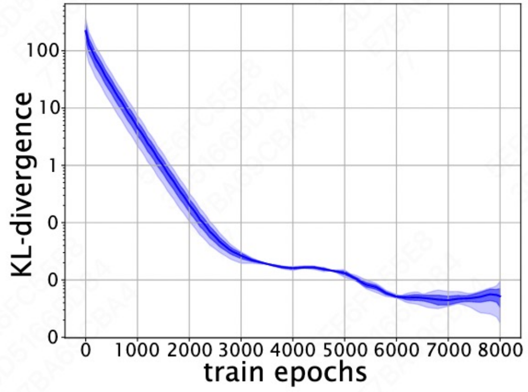
RecSim



(a) epoch 0 (b) epoch 4000 (c) epoch 8000

Figure 1: Illustration of the visualization on v. The X-axis denotes the first principal component, and the Y-axis denotes the second one. Each cross point denotes the projection of the latent code for the state distribution. The numbers with the same color to the point denote the ground-truth environment





Real-world
(Didichuxing)

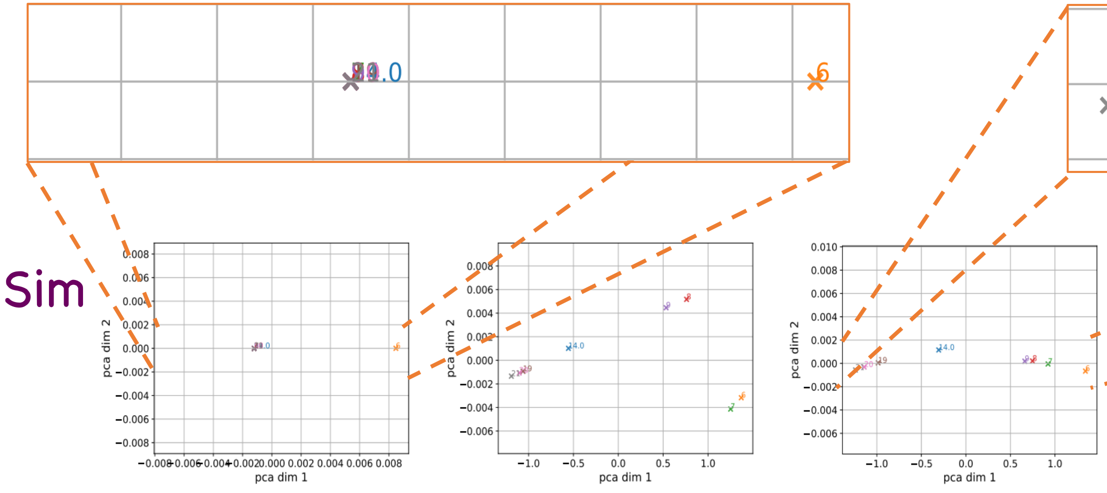
LAMDA Learning And Mining from DatA

南京大学 NANJING UNIVERSITY

POLIXIR 南栖仙策

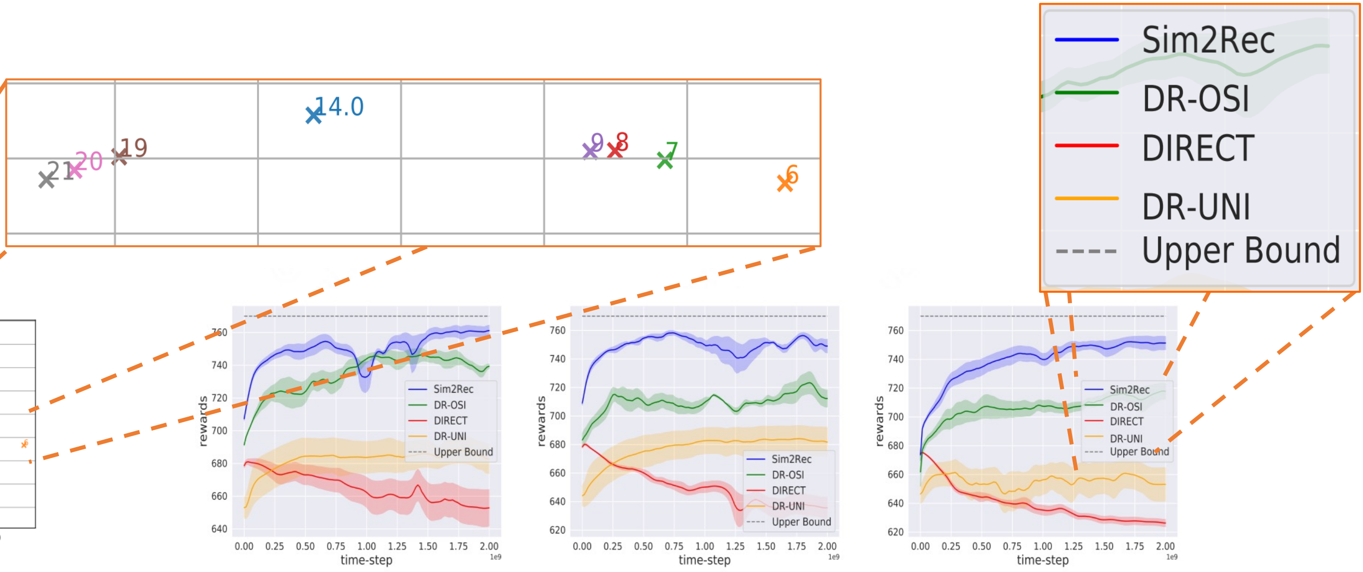
Selected Experiment Results

RecSim



(a) epoch 0 (b) epoch 4000 (c) epoch 8000

Figure 1: Illustration of the visualization on v. The X-axis denotes the first principal component, and the Y-axis denotes the second one. Each cross point denotes the projection of the latent code for the state distribution. The numbers with the same color to the point denote the ground-truth environment



(a) LTS1 (b) LTS2 (c) LTS3

Figure 2: Illustration of the performance in synthetic environments. The solid curves are the mean reward and the shadow is the standard error of three seeds.

Real-world (Didichuxing)

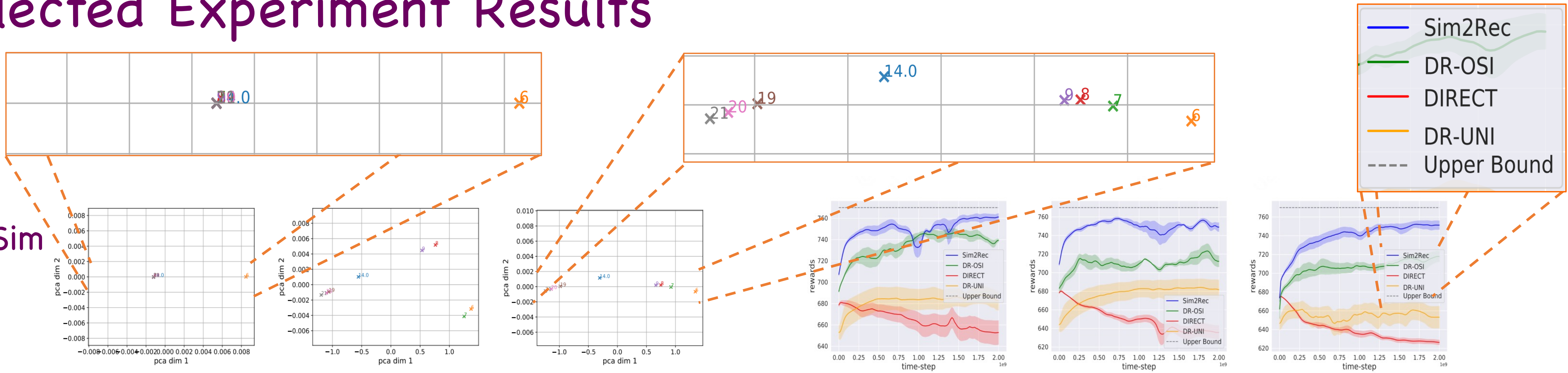
Learning And Mining from DatA

NANJING UNIVERSITY

POLIXIR 南栖仙策

Selected Experiment Results

RecSim



(a) epoch 0 (b) epoch 4000 (c) epoch 8000

Figure 1: Illustration of the visualization on v. The X-axis denotes the first principal component, and the Y-axis denotes the second one. Each cross point denotes the projection of the latent code for the state distribution. The numbers with the same color to the point denote the ground-truth environment

(a) LTS1 (b) LTS2 (c) LTS3

Figure 2: Illustration of the performance in synthetic environments. The solid curves are the mean reward and the shadow is the standard error of three seeds.

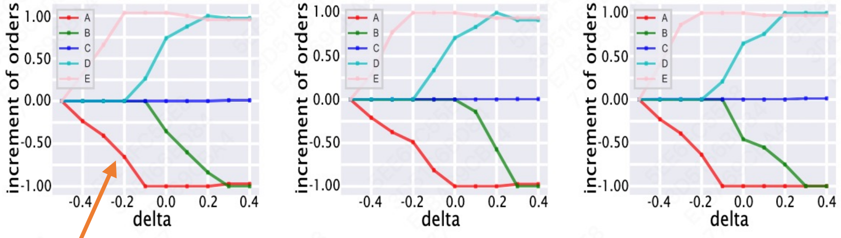
Real-world (Didichuxing)

TABLE III
THE PERFORMANCE OF POLICIES LEARNED WITH DIFFERENT POLICY LEARNING TECHNIQUES. THE PERFORMANCE IS TESTED IN SIMA.

	orders (test)	orders (train)	cost (test)	cost (train)
Sim2Rec	2.0%	1.6%	0.9%	4.5%
Sim2Rec-PE	1.3%	2.3%	-8.0%	-4.0%
Sim2Rec-EE	8.1%	8.2%	-10.0%	-11.1%

performance decline when deployed

exploit the model to reach high performance through unreasonable actions



(a) SimA (b) SimB (c) SimC

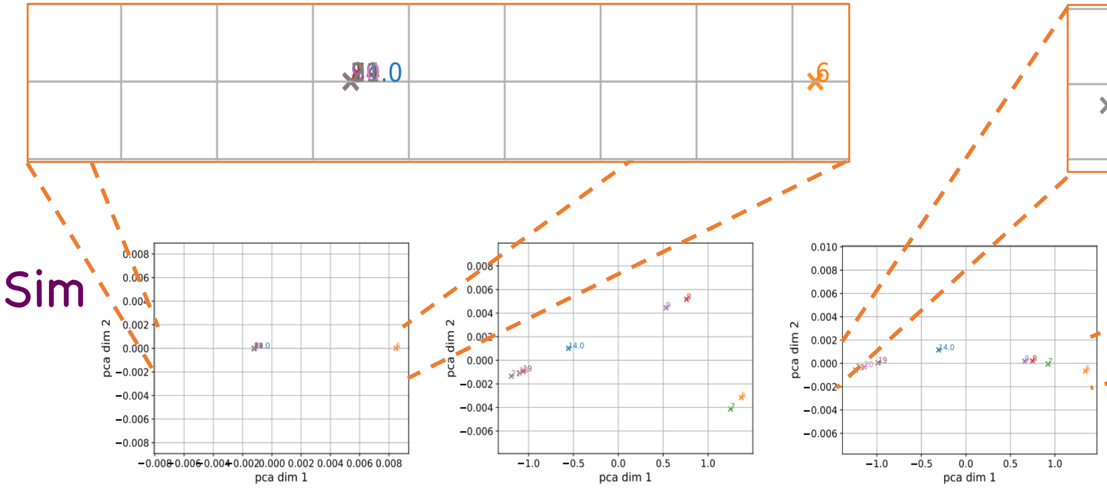
Fig. 10. Illustration of the increment of orders on intervention test. Each figure plots the clustering centers of the drivers' response vectors in a simulator. Each line denotes a cluster center. The X-axis is the value of ΔB. The increment of orders of each point is subtracted to the value in $\Delta B = -0.5$ of the corresponding cluster.

Learning And Mining from DatA

NANJING UNIVERSITY

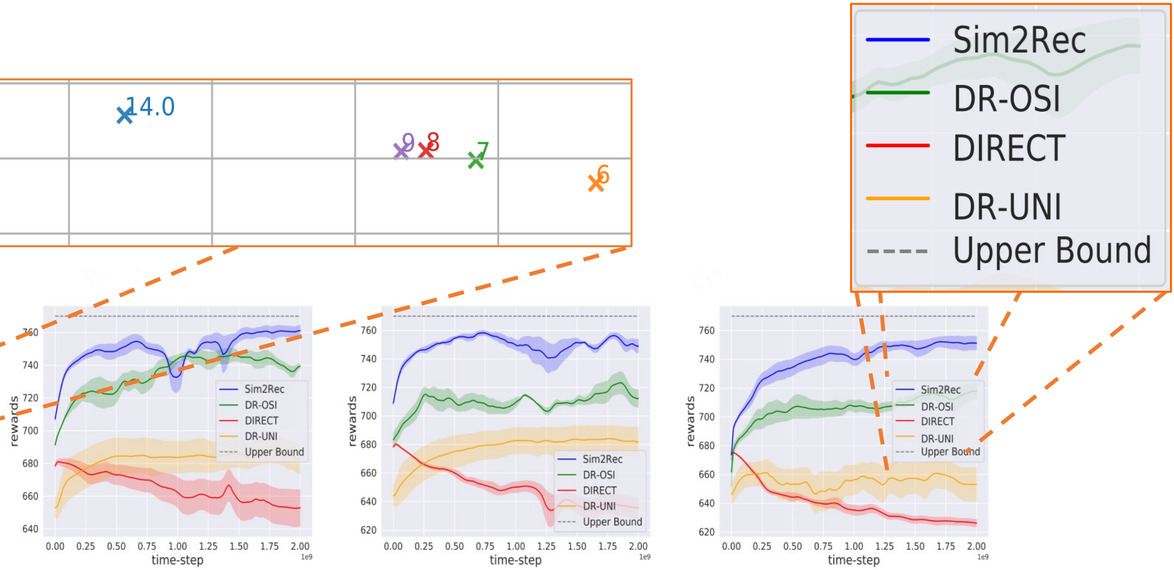
Selected Experiment Results

RecSim



(a) epoch 0 (b) epoch 4000 (c) epoch 8000

Figure 1: Illustration of the visualization on v. The X-axis denotes the first principal component, and the Y-axis denotes the second one. Each cross point denotes the projection of the latent code for the state distribution. The numbers with the same color to the point denote the ground-truth environment



(a) LTS1 (b) LTS2 (c) LTS3

Figure 2: Illustration of the performance in synthetic environments. The solid curves are the mean reward and the shadow is the standard error of three seeds.

Real-world (Didichuxing)

TABLE III
THE PERFORMANCE OF POLICIES LEARNED WITH DIFFERENT POLICY LEARNING TECHNIQUES. THE PERFORMANCE IS TESTED IN SIMA.

	orders (test)	orders (train)	cost (test)	cost (train)
Sim2Rec	2.0%	1.6%	0.9%	4.5%
Sim2Rec-PE	1.3%	2.3%	-8.0%	-4.0%
Sim2Rec-EE	8.1%	8.2%	-10.0%	-11.1%

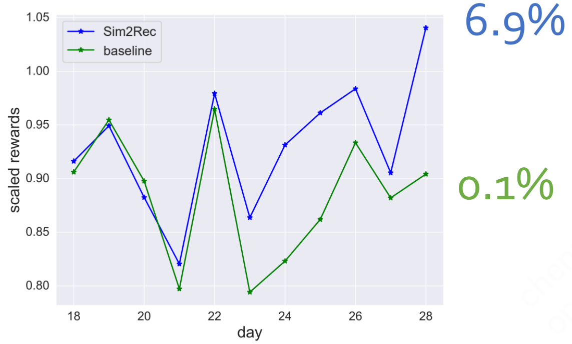


Figure 4: Illustration of the online test. The X-axis is the date. The Y-axis is the average daily reward.

LAMDA Learning And Mining from DatA
南京大学 NANJING UNIVERSITY
POLIXIR 南栖仙策

Take-home Messages & Thanks

KEY points:

1. The reality-gaps problem is important and should be taken into consideration for designing a power RL-based SRS
2. We identify and give a possible way to handle the several extra challenges in adopting standard zero-shot policy transfer solutions to SRS scenarios. We think more powerful and theoretical methods can be developed in future work

LAMDA Learning And Mining from DatA

南京大學 NANJING UNIVERSITY

POLIXIR 南栖仙策

>> Thanks

Learning And Mining from DatA

南京大學

NANJING UNIVERSITY

POLIXIR

南栖仙策