On Trajectory Augmentations for Off-Policy Evaluation

Ge Gao*, Qitong Gao†, Xi Yang‡, Song Ju*, Miroslav Pajic†, Min Chi*

*North Carolina State University

†Duke University

‡IBM Research

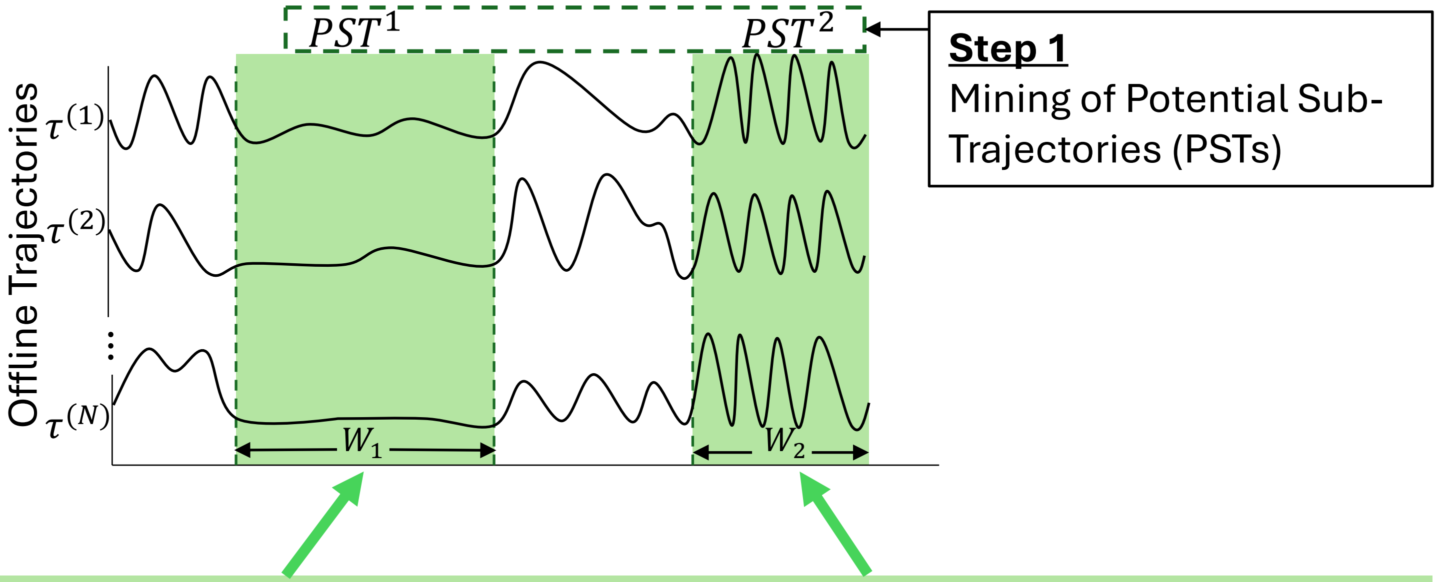
Background

- Off-policy evaluation (OPE)
  - Estimates the performance of a *target* policy using historical data collected over a (different) *behavior* policy.
  - Safety: evaluate diverse interventions/policies

Background

- Offline trajectories: limited coverage of the entire state-action space
  - Hinder OPE methods to evaluate diverse policies
- Data augmentation: powerful for data enrichment; present effectiveness in varied tasks (e.g., supervised learning) [Deng09, Yoon19, Kamycki19, Iwana21, Xie20]
  - General supervised and unsupervised learning: ignore Markovian nature
  - RL policy optimization: different goals. Learning from high-reward regions of the state-action space (policy optimization) vs both high- and low-reward regions (OPE)

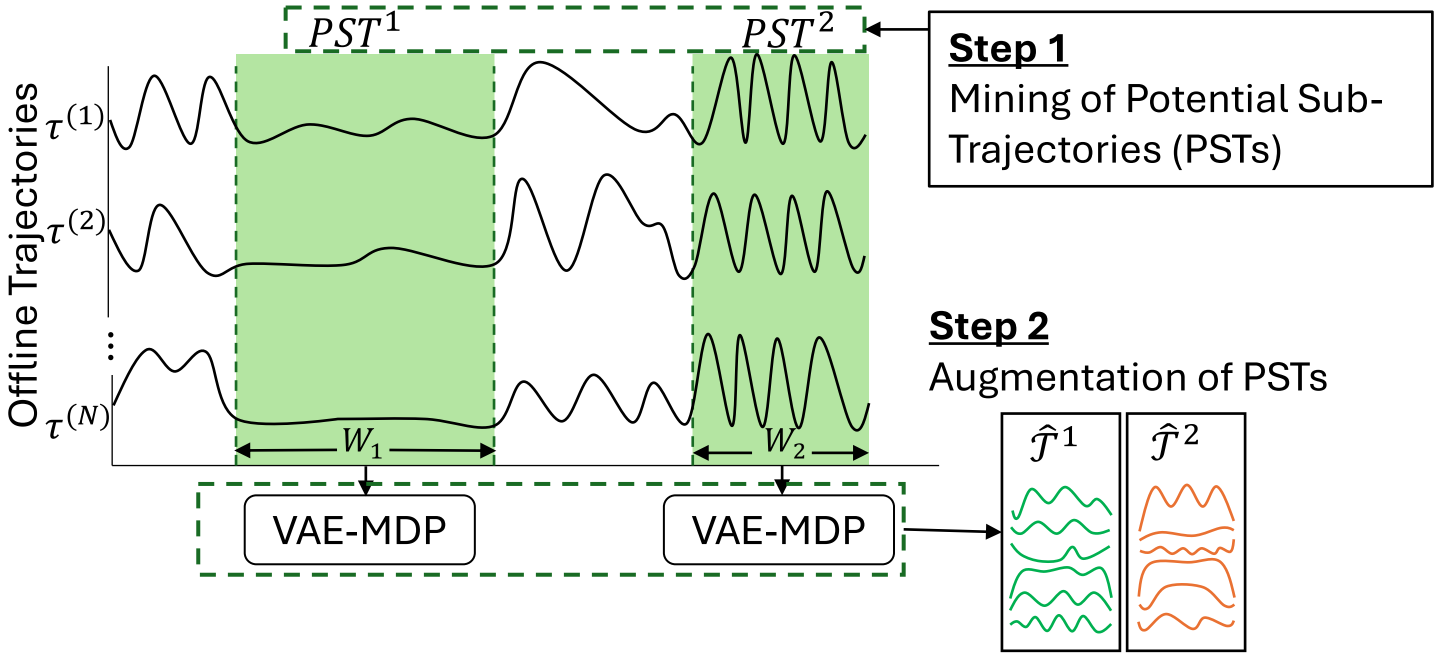
Methods



Share similar behaviors while may have more potential to behave diversely under heterogenous target policies

Facilitate OPE with Augmented Trajectories (OAT)

Methods

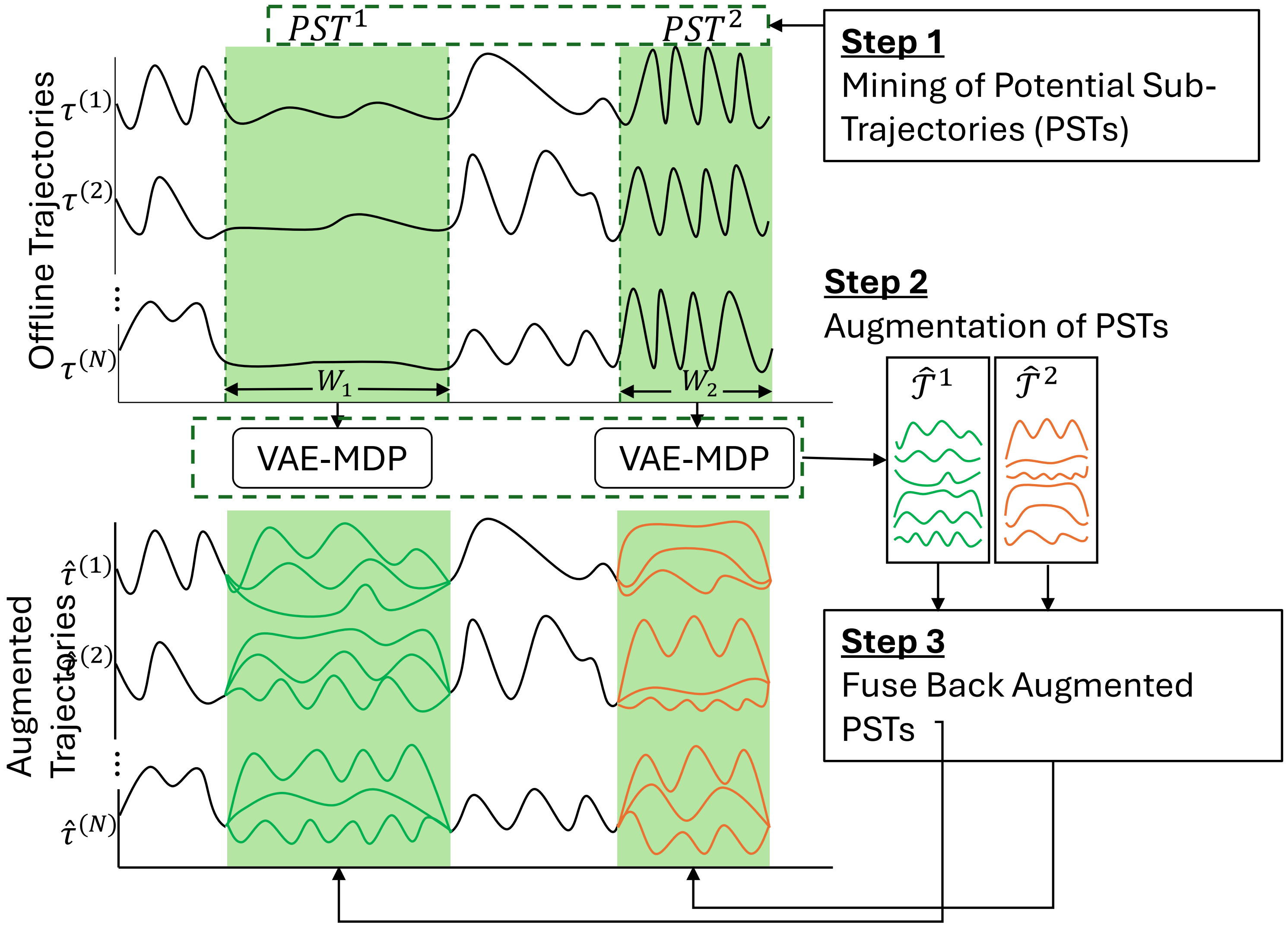


<u>Enrich state-action coverage of PSTs</u>

(i) The latent prior: represents distributions of initial latent space over PSTs
(ii) The encoder: encodes MDP transitions into the latent space
(iii) The decoder: reconstructs new PST samples

Objective: maximize the evidence lower bound (ELBO)

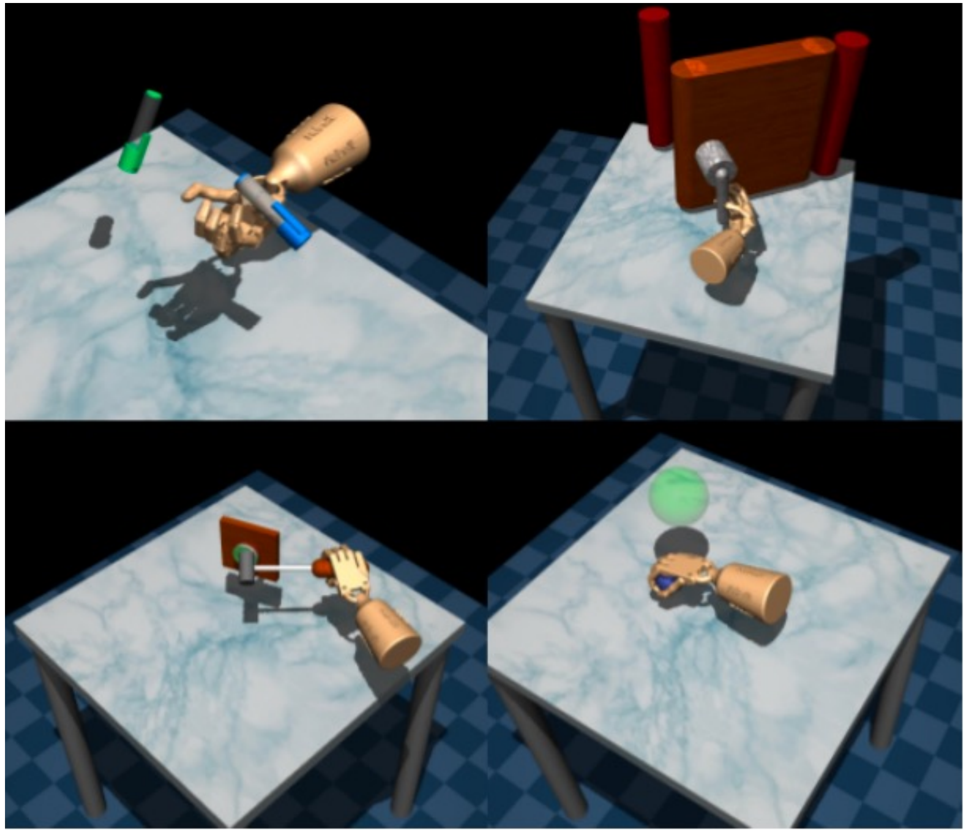
Methods


PST^1
PST^2
Offline Trajectories
$\tau^{(1)}$
$\tau^{(2)}$
$\tau^{(N)}$
W_1
W_2
VAE-MDP
VAE-MDP
Augmented Trajectories
$\hat{\tau}^{(1)}$
$\hat{\tau}^{(2)}$
$\hat{\tau}^{(N)}$
$\hat{\mathcal{T}}^1$
$\hat{\mathcal{T}}^2$
Step 1
Mining of Potential Sub-Trajectories (PSTs)
Step 2
Augmentation of PSTs
Step 3
Fuse Back Augmented PSTs

Experiment 1: Adroit

Adroit [Rajeswaran18]:

- 4 tasks: a simulated Shadow Hand robot is asked to hammer a nail (***hammer***), open a door (***door***), twirl a pen (***pen***), or pick up and move a ball (***relocate***)
- Deep OPE settings [Fu20]
- Behavior policy: behavior clone
- Target policies: 11 DAPG-based policies ranging from random to expert performance



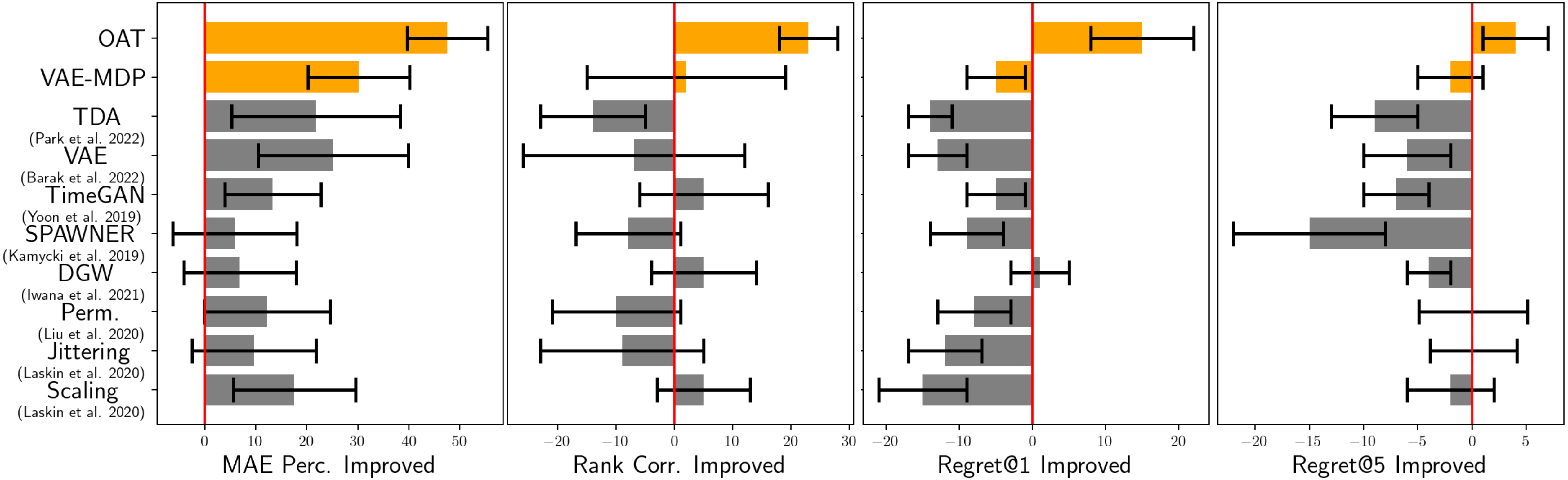
Baselines and Evaluation Metrics

- Baselines:
  - RL-oriented: TDA [Park22], permutation, jittering, scaling [Laskin20, Liu20, Raileanu21]
  - Generative methods: TimeGAN [Yoon19], VAE [Barak22]
  - Time series-oriented: SPAWNER [Kamycki19], DGW [Iwana21]
  - VAE-MDP

- OPE methods considered:
  - WIS, FQE [Le19], DualDICE [Yang20], DR [Thomas16], MB [Zhang20]

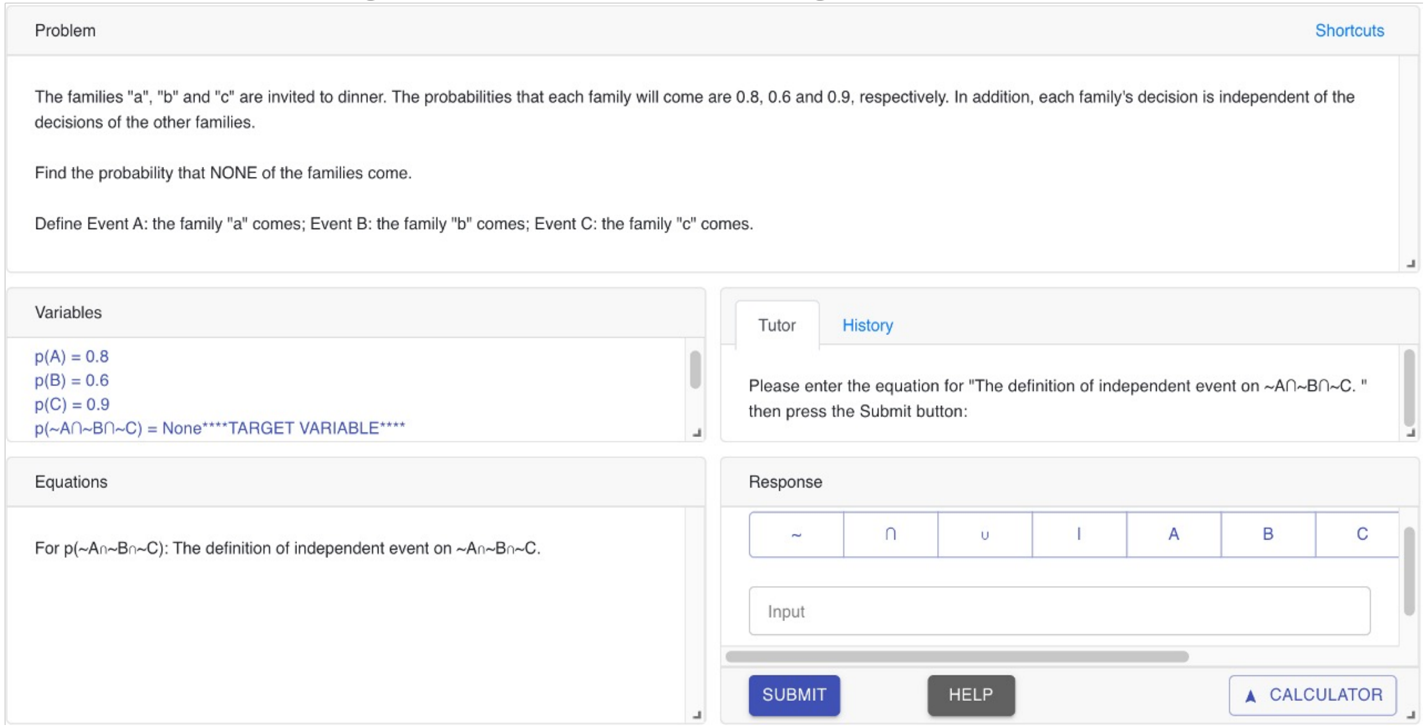
- Evaluation Metrics: Absolute error, Regret@1, Regret@5, Spearman's rank correlation

Experiment 1: Results (100% human-involving datasets)



Averaging across 5 OPE methods and 4 tasks
Results from each dataset averaging over 3 random seeds

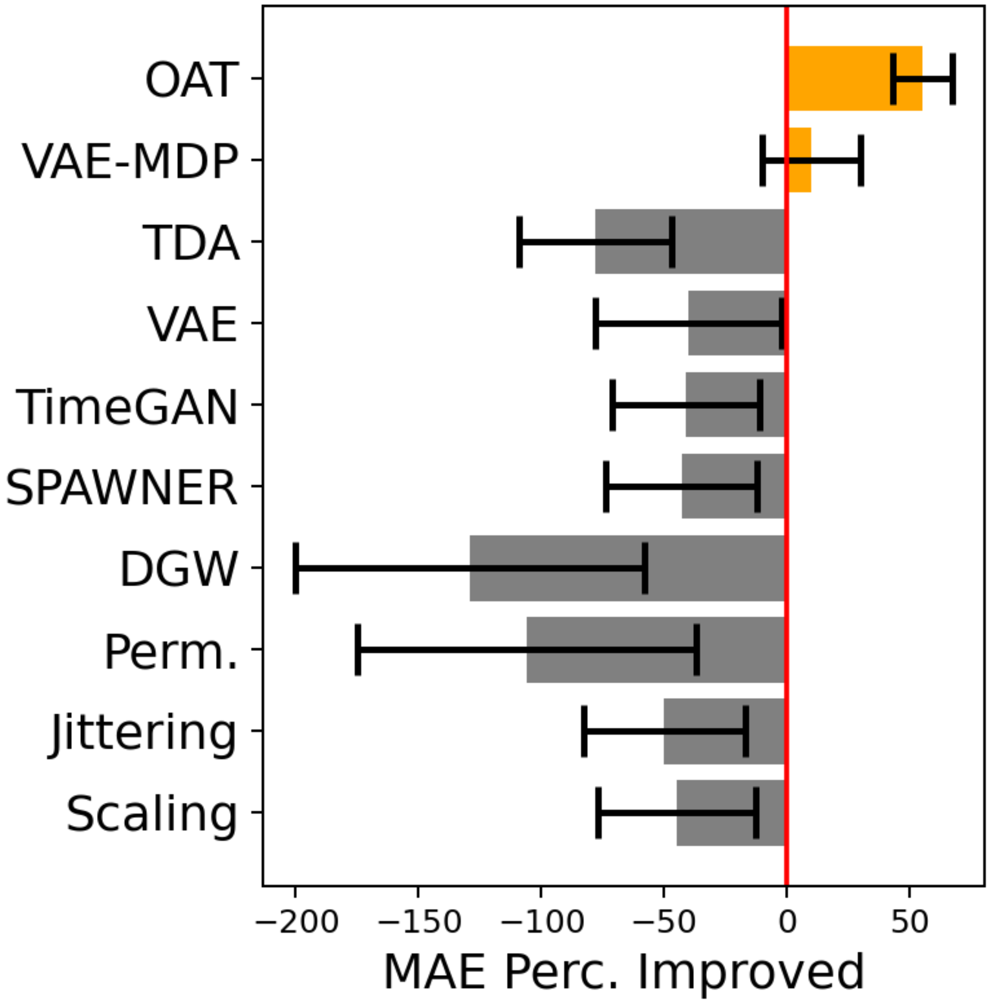
Experiment 2: Intelligent Tutoring



1,307 students over seven semesters (<u>Prior 6 for training and evaluating, the following one for testing</u>)

- States: 142 attributes
- Actions: 3 types of next problem
- Rewards: students' normalized learning gain
- Behavior policy: behavior clone.
- Target policy: 3 DQN-based policies and 1 instructor hand-designed policy.

Experiment 2: Results



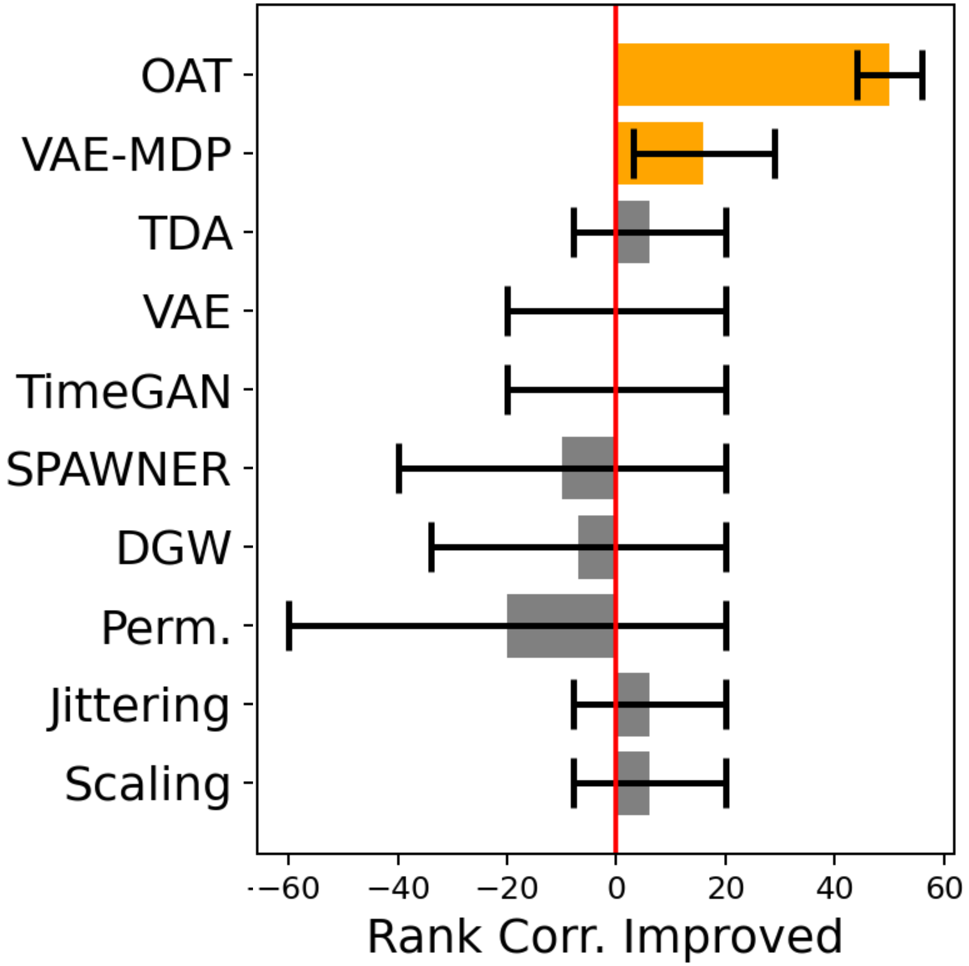
Experiment 3: Sepsis Treatment

Sepsis treatment: challenging problem; fully offline evaluation

Our data: 221,700 visits of patients over two years. (80-20 split for training and testing)

- States: 15 continuous attributes (e.g., heart rate)
- Actions: 4. Binary options over antibiotic administration & oxygen assistance.
- Rewards: Obtain on the four stages of sepsis (infection (±5), inflammation (±10), organ failure (±20), and septic shock (±50)).
- Behavior policy: behavior clone
- Target policy: 5 DQN-policies

Experiment 3: Results



Summary

- OAT:
    - Improve the state-action coverage of offline trajectories
    - Potential-sub-trajectory mining; VAE-MDP
    - Superior performance across domains, including robotic control, education, and healthcare

Thank you!