



# Paper Sharing

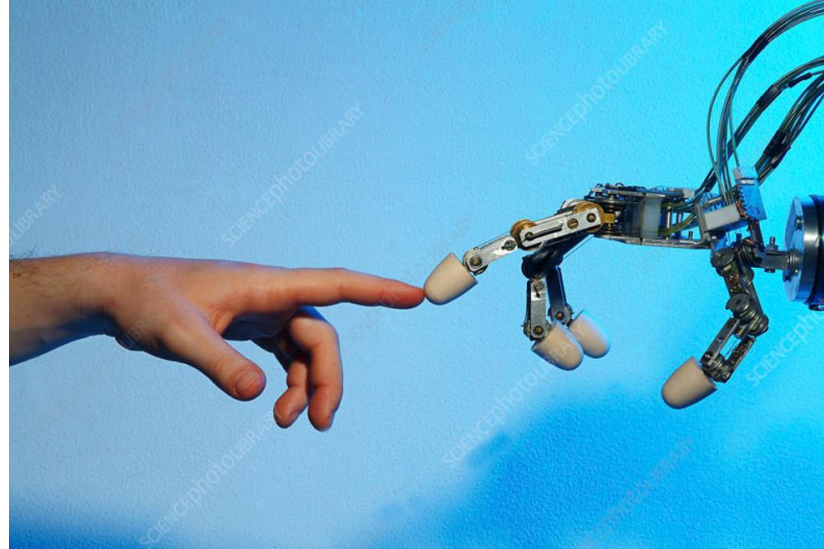
**Watch and Match: Supercharging Imitation with  
Regularized Optimal Transport**

**Lecturer: Yuxin Wu**

**2023.2.23**

# Motivation

- How should we teach our robots?



- Watch and Match

★ Imitation Learning

# Motivation

- Imitation Learning-BC (Behavior Cloning)



- BC (Behavior Cloning) -disadvantage:

- Time-consuming
- Distribution mismatch
- Lack of diversity
- Error accumulation

# Motivation

- BC (Behavior Cloning) -disadvantage:
  - tend to require large amounts of data, which takes a long time to collect
- Time-consuming
- Distribution mismatch
- Lack of diversity
- Error accumulation



# Motivation

- BC (Behavior Cloning) -disadvantage:

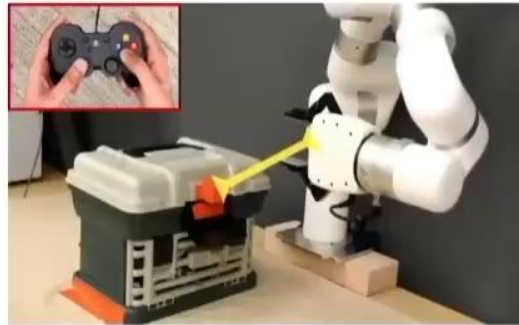
- Time - consuming

- Distributional mismatch

- Lack of diversity

- Error accumulation

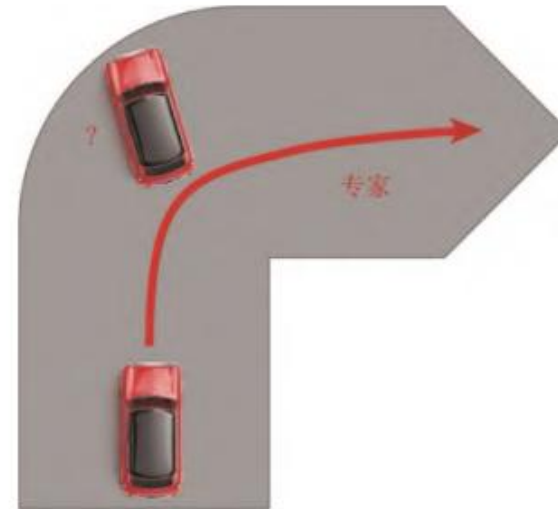
Demonstration



**Task:** Opening a box

# Motivation

- BC (Behavior Cloning) -disadvantage:
- Time-consuming
- Distribution mismatch
- Lack of diversity
- Error accumulation



# Motivation

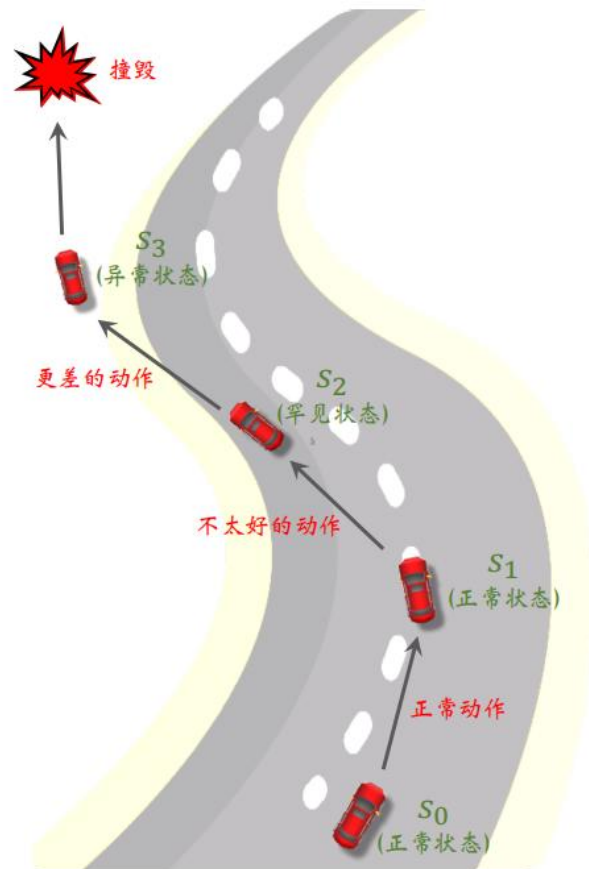
- BC (Behavior Cloning) -disadvantage:

- Time-consuming

- Distribution mismatch

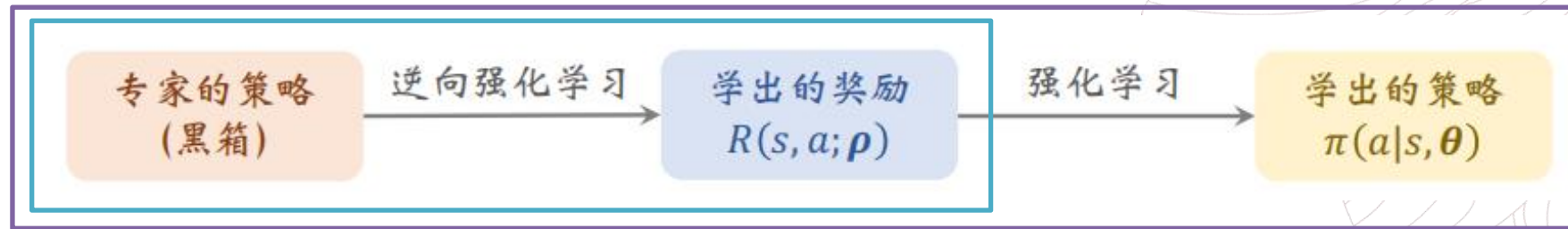
- Lack of diversity

- Error accumulation



# Motivation

- Imitation Learning-IRL (Inverse Reinforcement Learning)



Inverse Reinforcement Learning

apprenticeship learning

Inverse Reinforcement Learning(if called loosely)

- IRL (Inverse Reinforcement Learning) -disadvantage:
  - require numerous expensive online interactions with the environment
  - inferred reward function-highly non-stationary
  - require effective exploration to maximize rewards
  - strong priors applied-cause a distribution shift



# Motivation

- BC (Behavior Cloning)



Ability to imitate demonstration behavior



How to obtain a balance between the both?



Online learning to fine-tune a pre-trained policy

- IRL (Inverse Reinforcement Learning)

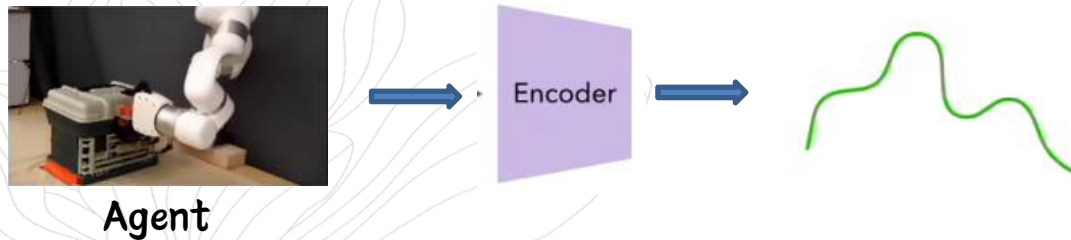
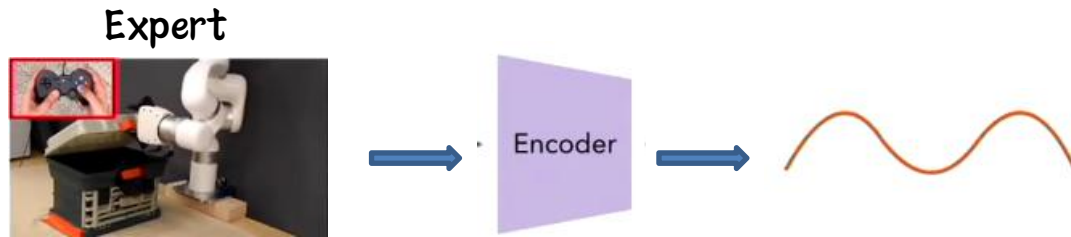


Recovery ability from OOD



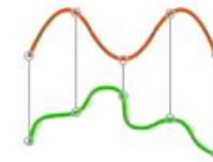
# Motivation

- Online learning to match demonstrations



## Variants trajectory matching

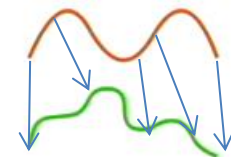
1. Compare each time step



2. Discriminator as reward functions

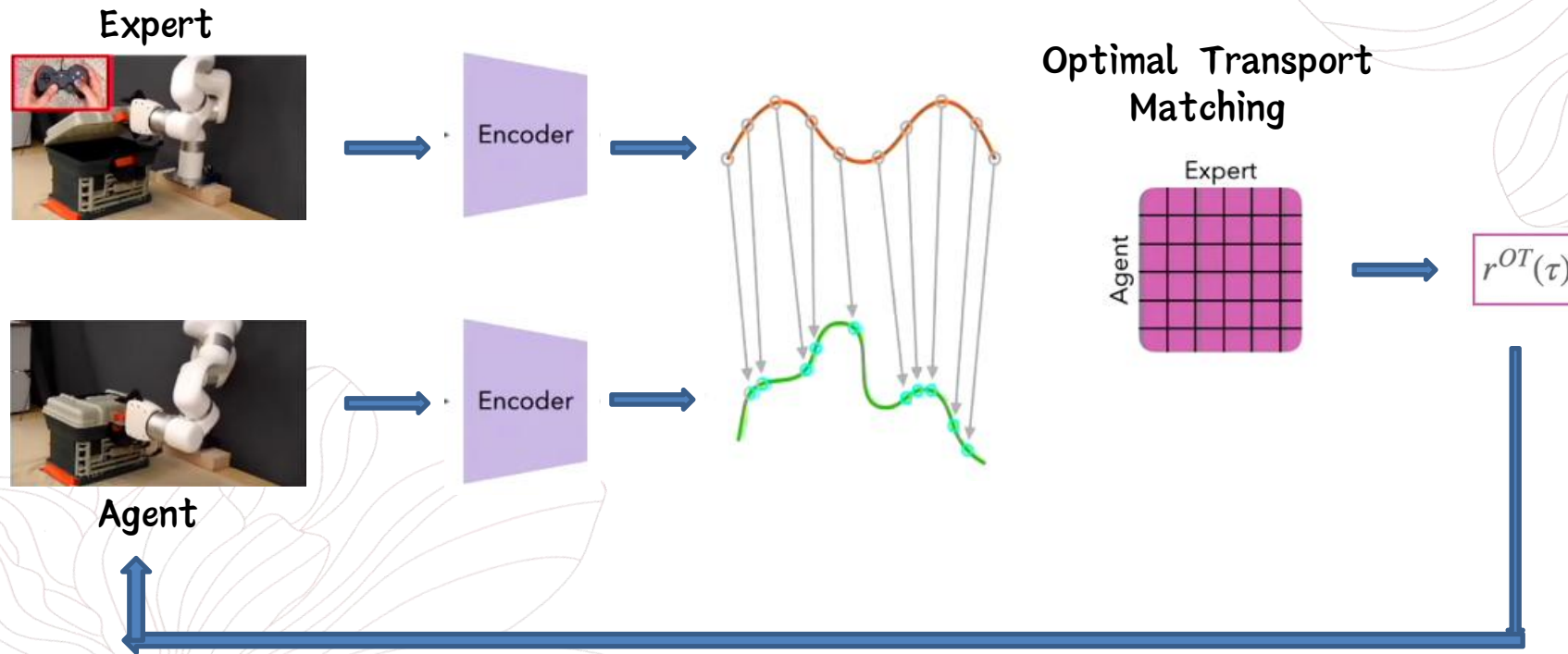


3. OT-based matching



# Motivation

- Online learning to match demonstrations

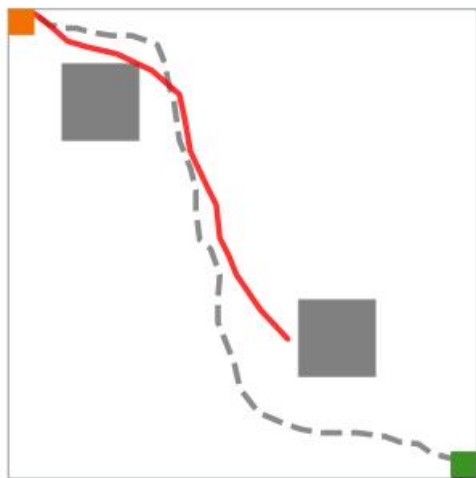


# Motivation

- Issue with Online Finetuning

- Naïvely combining offline pre-training with online fine tuning does not work well

(a) Task: Particle Reach

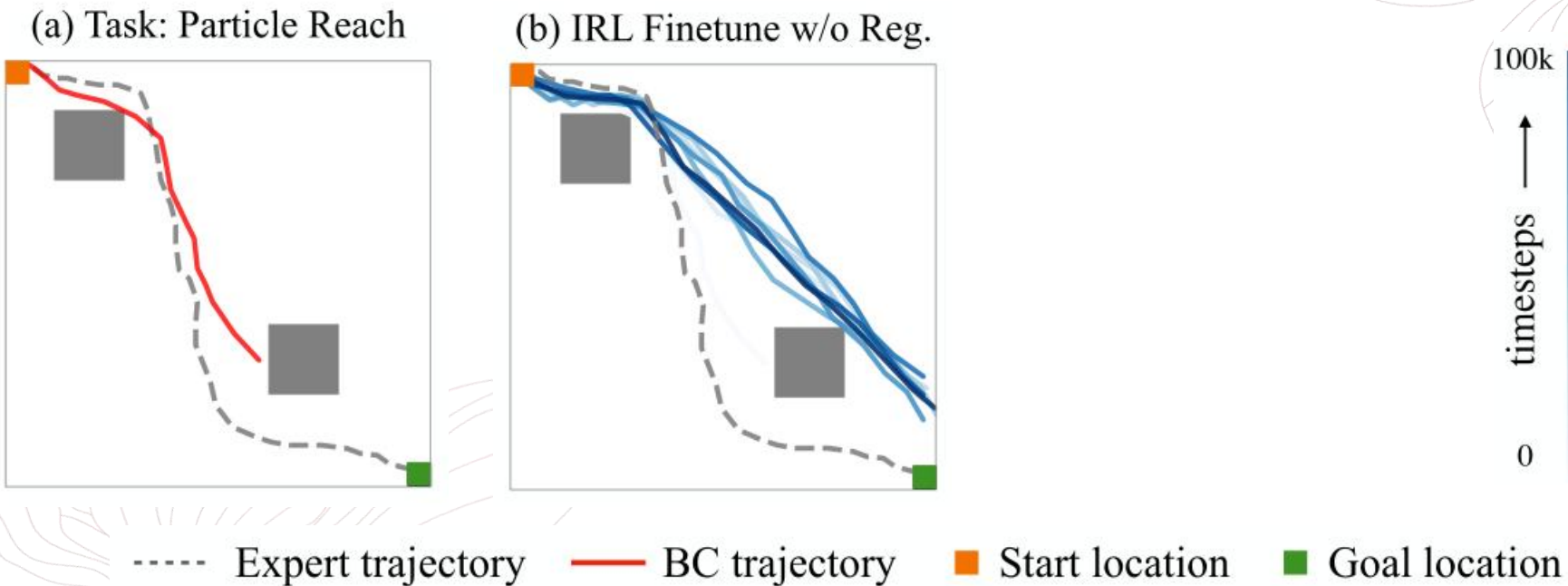


----- Expert trajectory    — BC trajectory    ■ Start location    ■ Goal location

# Motivation

- Issue with Online Finetuning

- Naïvely combining offline pre-training with online fine tuning does not work well





# Method

- Regularized Optimal Transport (ROT)

- Online Fine tuning with BC Regularization

$$\pi^{ROT} = \operatorname{argmax}_{\pi} \left[ (1 - \underbrace{\lambda(\pi)}_{\text{Adaptive Weight}}) \underbrace{\mathbb{E}_{(s,a) \sim \mathcal{D}_{\beta}} [Q(s,a)]}_{\text{RL Loss}} - \alpha \underbrace{\lambda(\pi)}_{\text{Adaptive Weight}} \underbrace{\mathbb{E}_{(s^e, a^e) \sim \mathcal{T}^e} \|a^e - \pi^{BC}(s^e)\|^2}_{\text{BC regularization}} \right]$$

$$\text{policy loss} = (1 - \underbrace{\text{Adaptive Weight}}_{\lambda(\pi)}) * \text{RL Loss} + \underbrace{\text{Adaptive Weight}}_{\lambda(\pi)} * \text{BC regularization}$$

- Key idea:** Keep  $\pi$  close to the Expert Data Distribution when  $\pi^{BC}$  performs better than  $\pi$

# Method

- Regularized Optimal Transport (ROT)

- Online Fine tuning with BC Regularization

$$\text{policy loss} = (1 - \text{Adaptive Weight}) * \text{RL Loss} + \text{Adaptive Weight} * \text{BC regularization}$$

- Key idea:** Keep  $\pi$  close to the Expert Data Distribution when  $\pi^{\text{BC}}$  performs better than  $\pi$

$$\lambda(\pi^{\text{ROT}}) = \mathbb{E}_{(s, \cdot) \sim \mathcal{D}_e} \left[ \mathbb{1}_{\boxed{Q(s, \pi^{\text{BC}}(s))} > \boxed{Q(s, \pi^{\text{ROT}}(s))}} \right]$$

$\pi^{\text{BC}}$  Performance

$\pi^{\text{ROT}}$  Performance

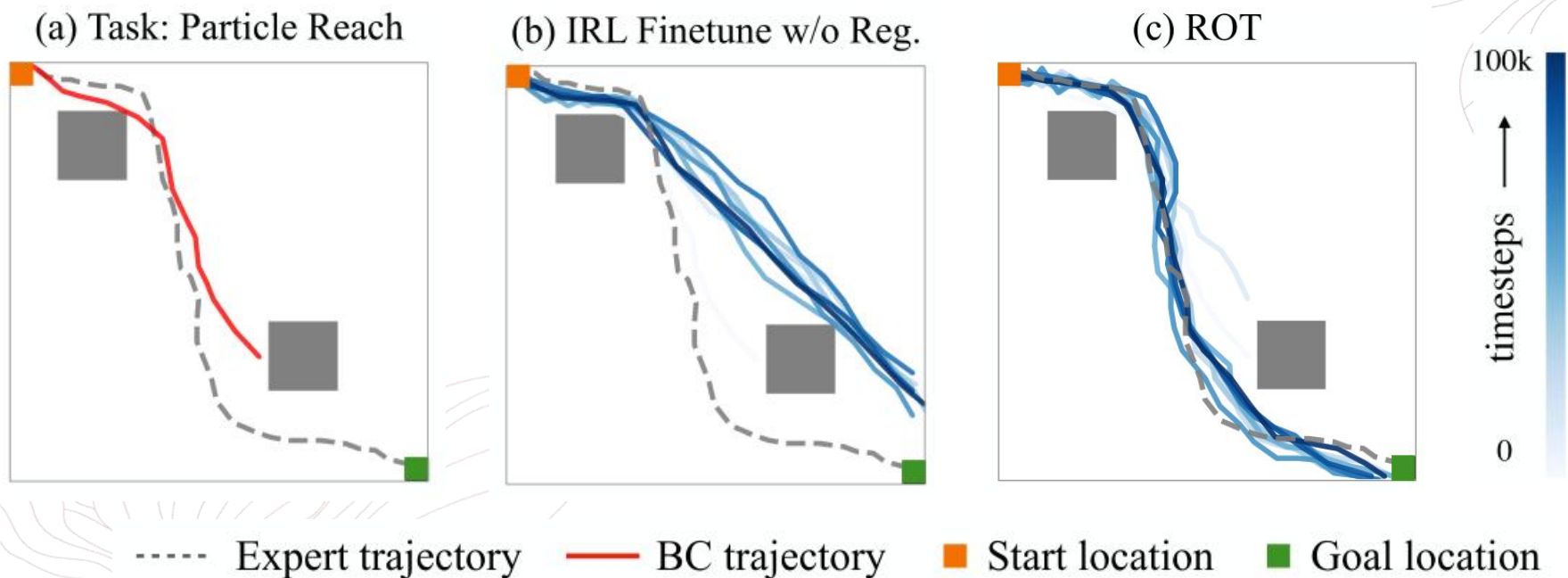
↓  
Proportion of states where  
BC performs better

- Key idea:** Increase BC regularization when  $\pi^{\text{BC}}$  performs better (on average) than  $\pi^{\text{ROT}}$

# Method

- Issue with Online Finetuning

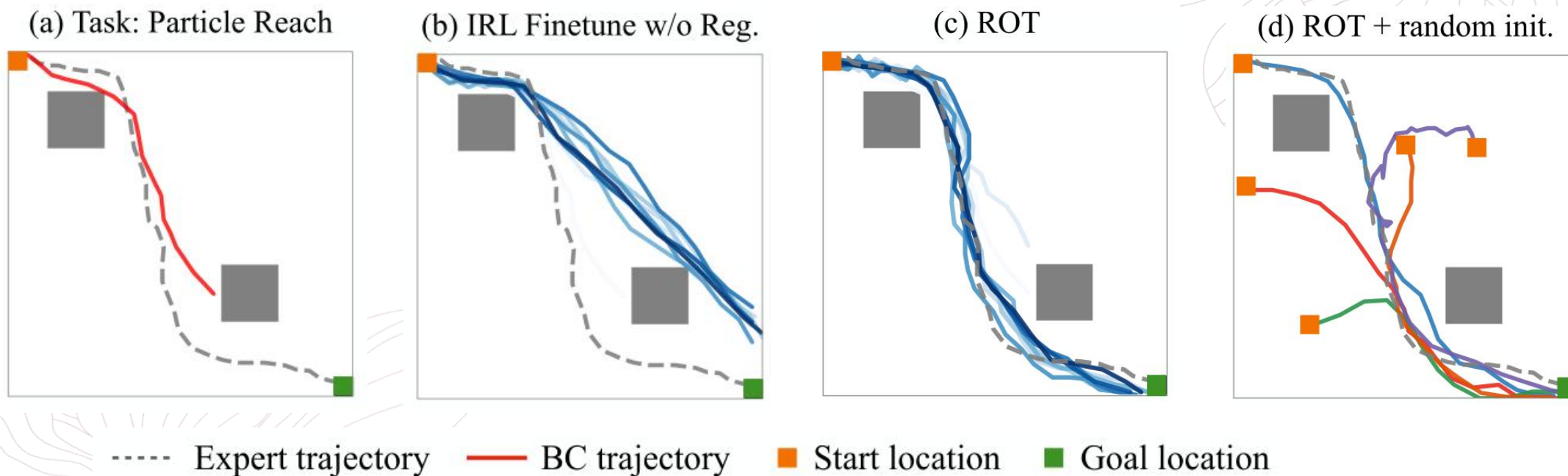
- Naïvely combining offline pre-training with online fine tuning does not work well



# Method

- Issue with Online Finetuning

- Naïvely combining offline pre-training with online fine tuning does not work well



# Experiments

- Experiments are designed to answer the following questions:
  - (a) How efficient is ROT for imitation learning?
  - (b) How does ROT perform on real-world tasks?
  - (c) How important is the choice of IRL method in ROT?
  - (d) Does soft Q-filtering improve imitation?
  - (e) How does ROT compare to standard reward-based RL?



# Experiments

- Experiments are designed to answer the following questions:
  - (a) How efficient is ROT for imitation learning?

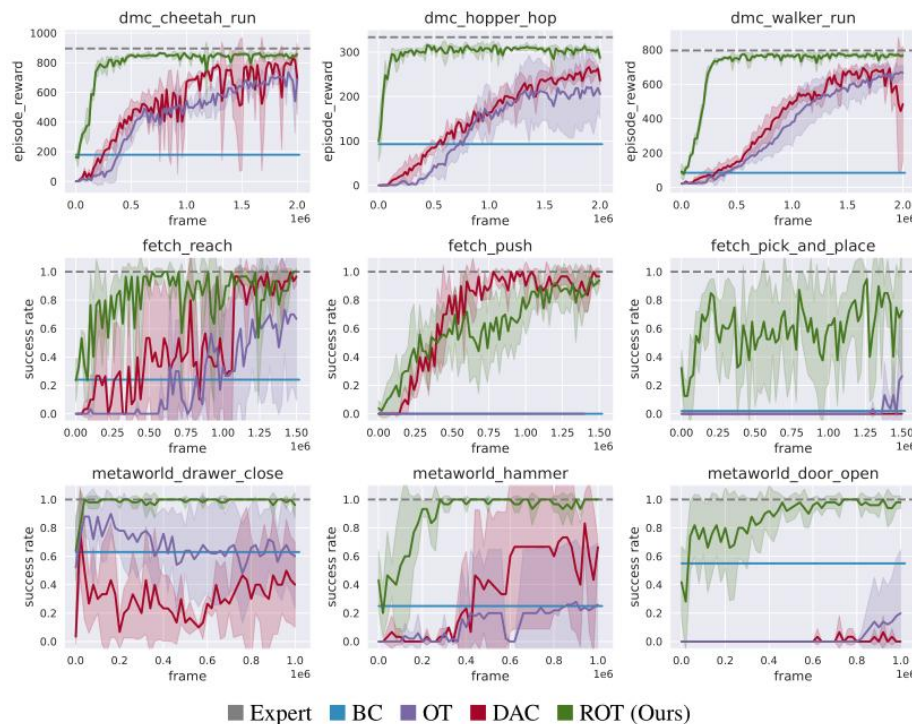


Figure 3: Pixel-based continuous control learning on 9 selected environments. Shaded region represents  $\pm 1$  standard deviation across 5 seeds. We notice that ROT is significantly more sample efficient compared to prior work.

# Experiments

- Experiments are designed to answer the following questions:
  - (b) How does ROT perform on real-world tasks?

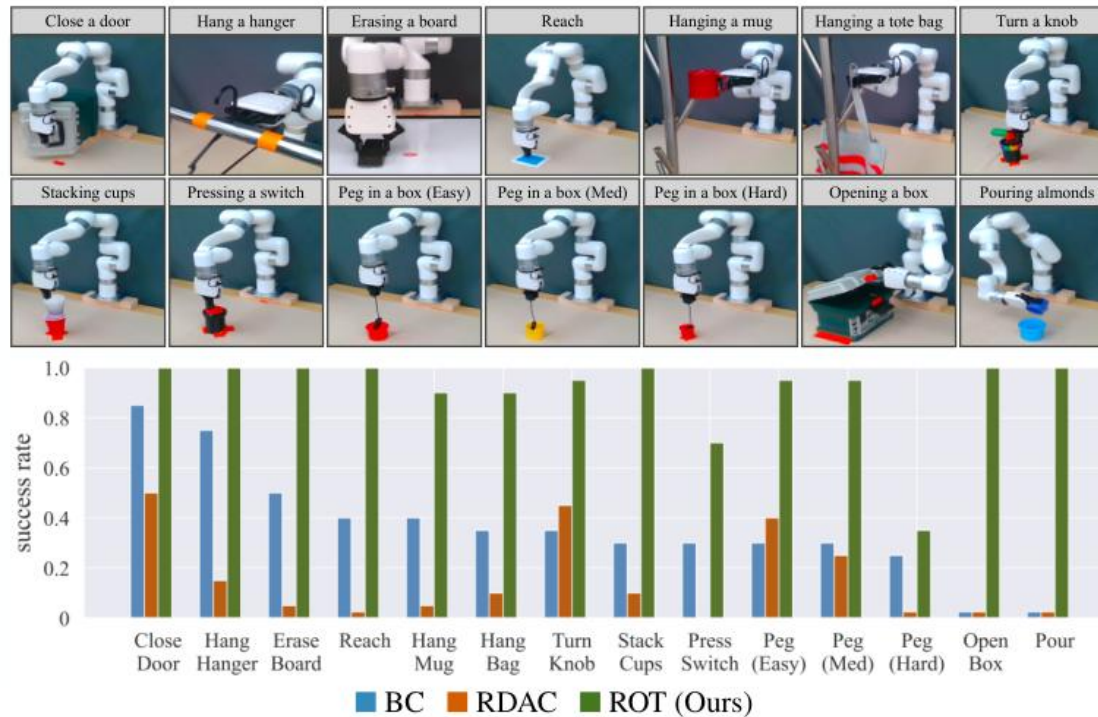


Figure 4: **(Top)** ROT is evaluated on a set of 14 robotic manipulation tasks. **(Bottom)** Success rates for each task is computed by running 20 trajectories from varying initial conditions on the robot.

# Experiments

- Experiments are designed to answer the following questions:
  - (c) How important is the choice of IRL method in ROT?

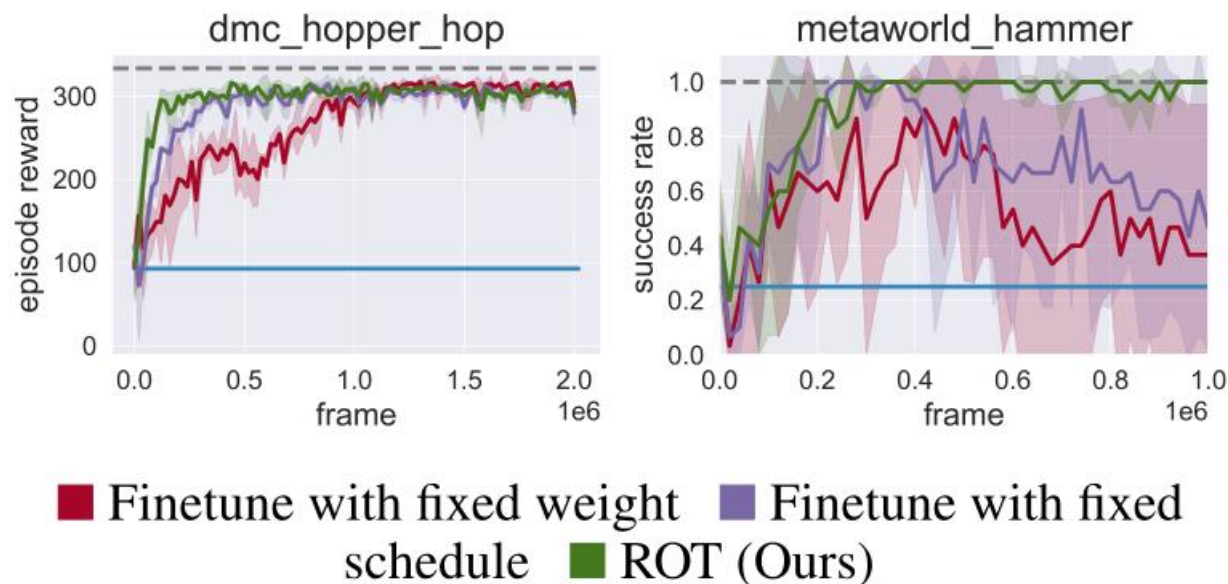


Figure 5: Effect of various BC regularization schemes compared with our adaptive soft-Q filtering regularization.



# Experiments

- Experiments are designed to answer the following questions:
  - (d) Does soft Q-filtering improve imitation?

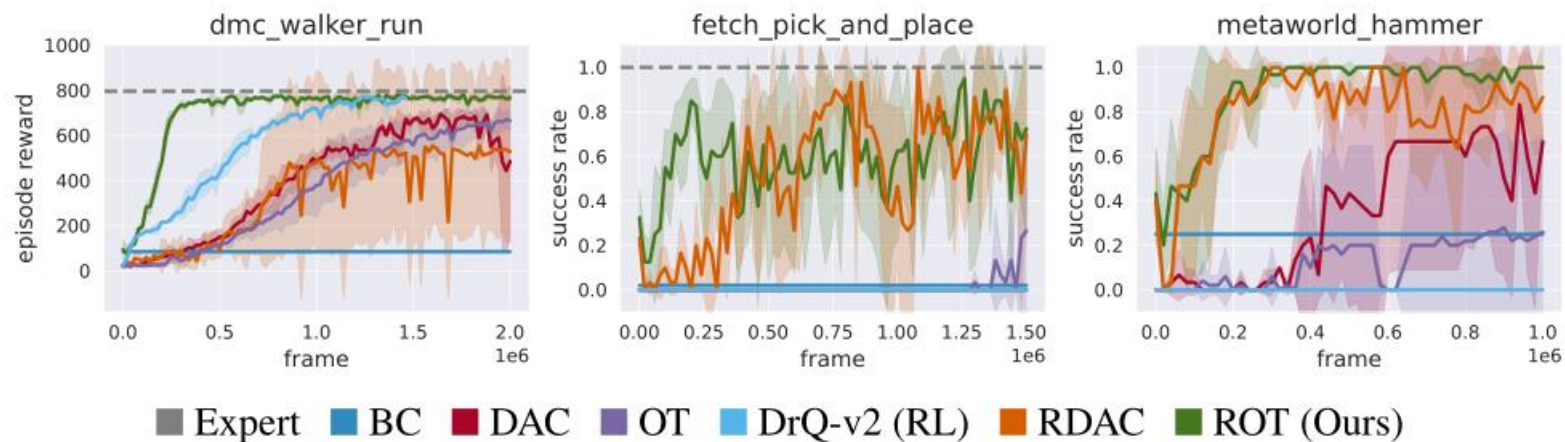


Figure 6: Ablation analysis on the choice of base IRL method. We find that although adversarial methods benefit from regularized BC, the gains seen are smaller compared to ROT. Here, we also see that ROT can outperform plain RL that requires explicit task-rewards.

# Experiments

- Experiments are designed to answer the following questions:
  - (e) How does ROT compare to standard reward-based RL?

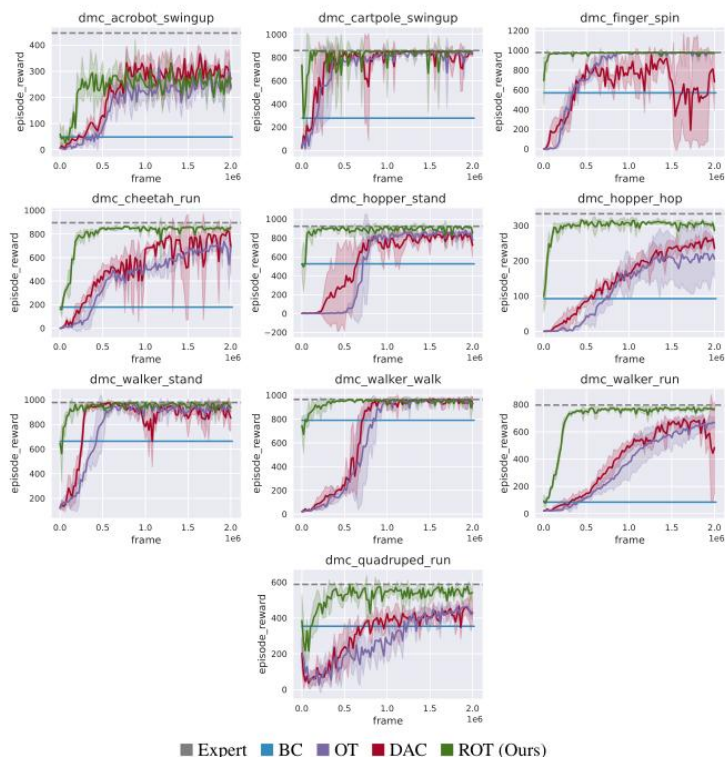
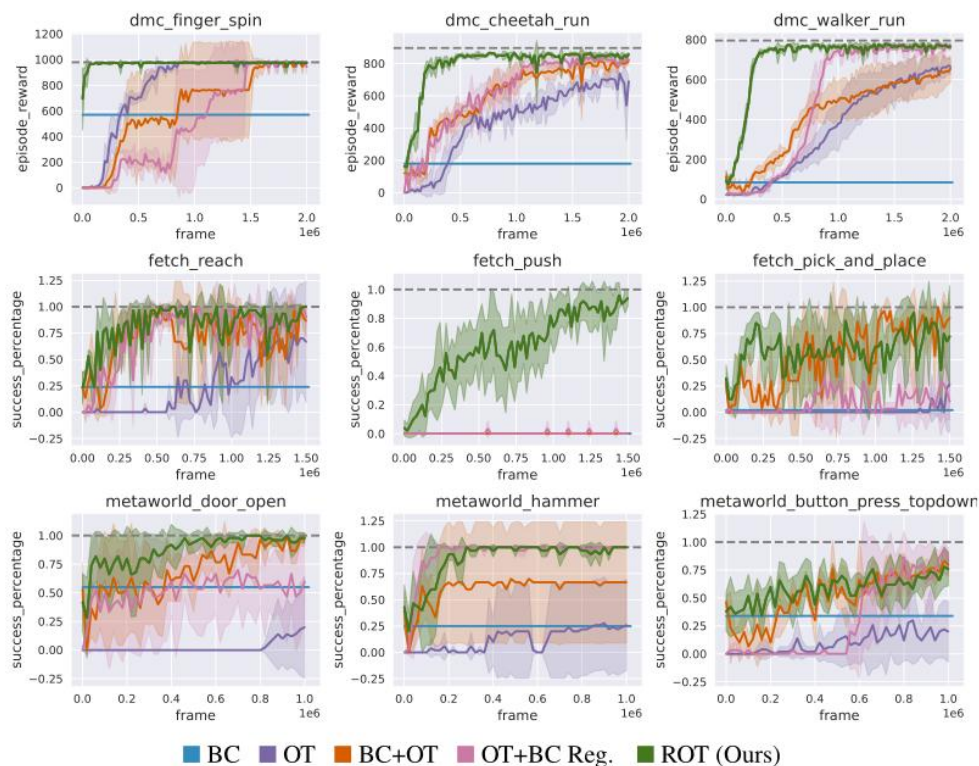


Figure 9: Pixel-based continuous control learning on 10 DMC environments. Shaded region represents  $\pm 1$  standard deviation across 5 seeds. We notice that ROT is significantly more sample efficient compared to prior work.



# Experiments

- Experiments are designed to answer the following questions:
- (f) How important are the design choices in ROT?

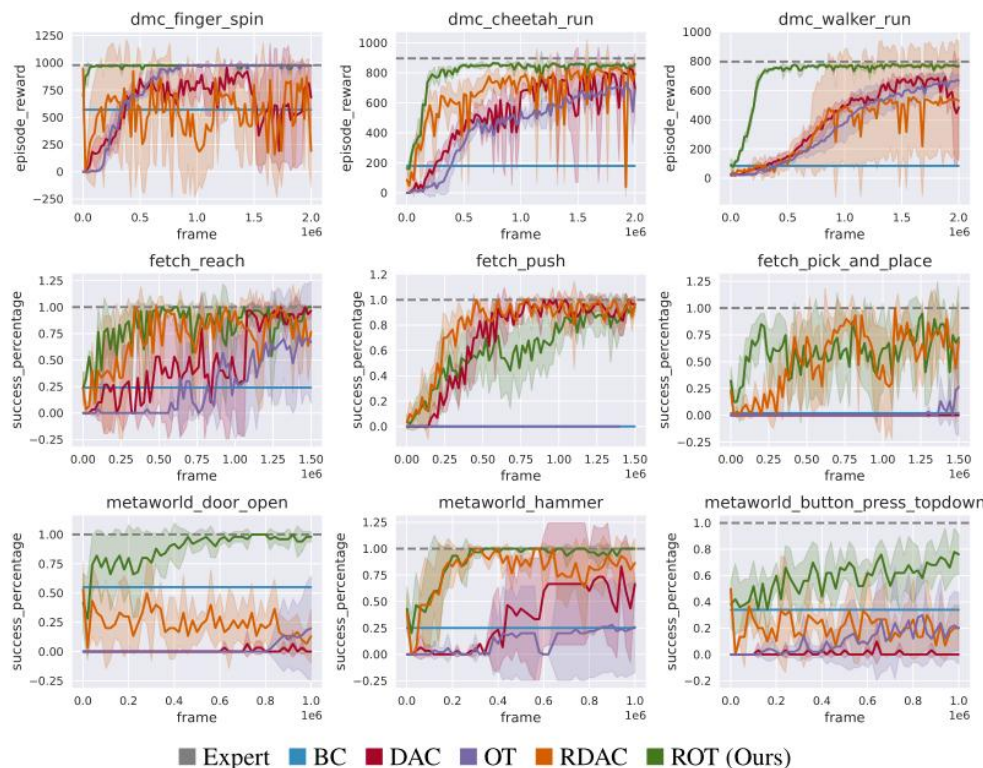


Importance of pretraining and regularizing the IRL policy

Figure 14: Pixel-based ablation analysis on the importance of pretraining and regularizing the IRL policy. The key takeaway from these experiments is that both pretraining and BC regularization are required to obtain sample-efficient imitation learning.

# Experiments

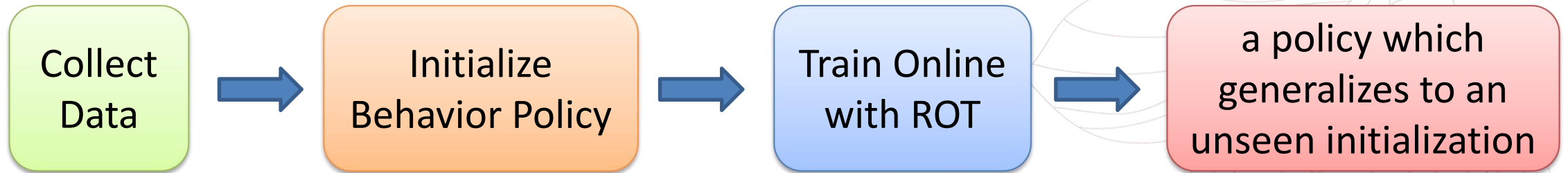
- Experiments are designed to answer the following questions:
  - (f) How important are the design choices in ROT?



Choice of IRL method

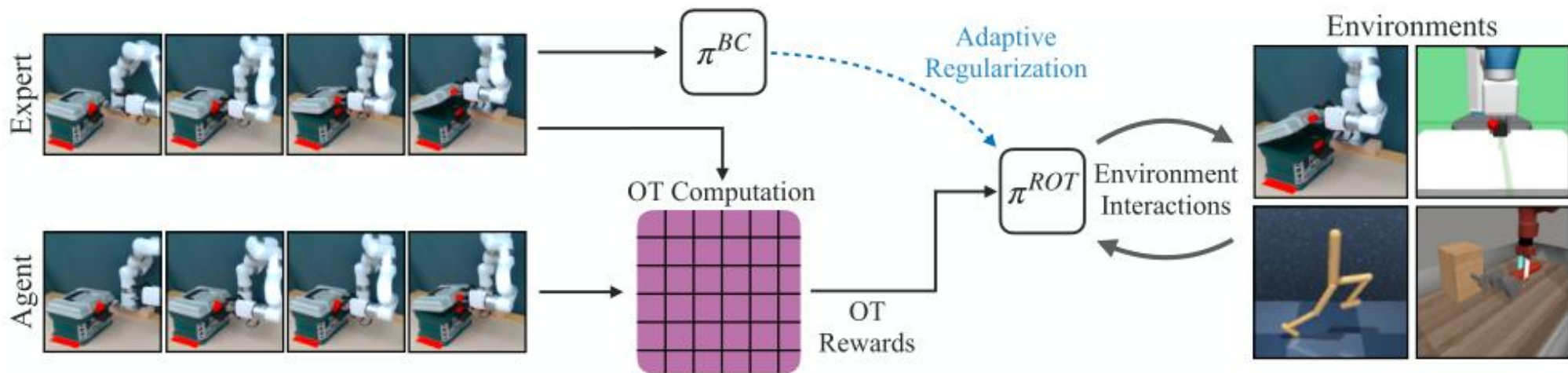
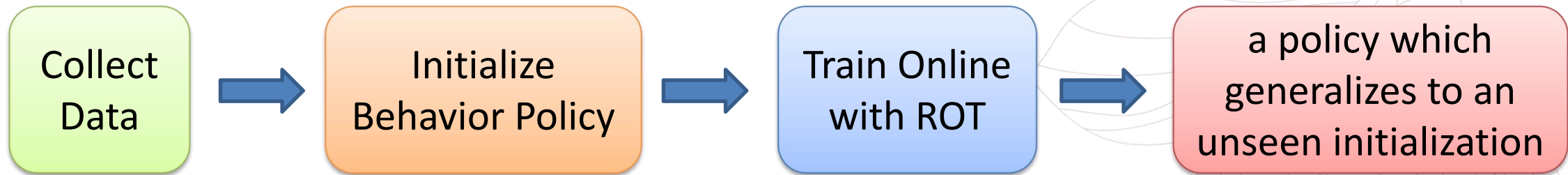
Figure 15: Pixel-based ablation analysis on the choice of base IRL method. We find that although adversarial methods benefit from regularized BC, the gains seen are smaller compared to ROT.

# Summary



- This work starts with collecting a single expert demonstration.
- This work initializes the behavior policy using behavior cloning.
- This work fine-tunes this behavior policy online with ROT.
- This online fine-tuning continues for an hour to obtain a policy which generalizes to an unseen initialization.

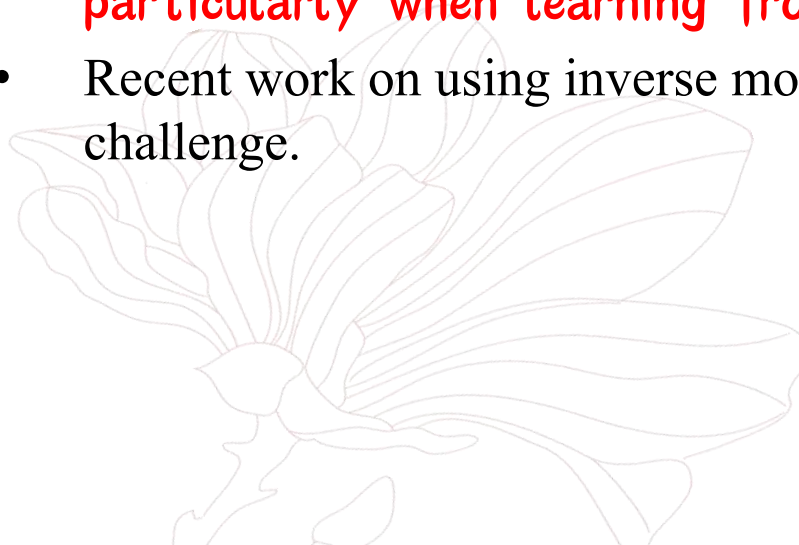
# Summary





# Limitations

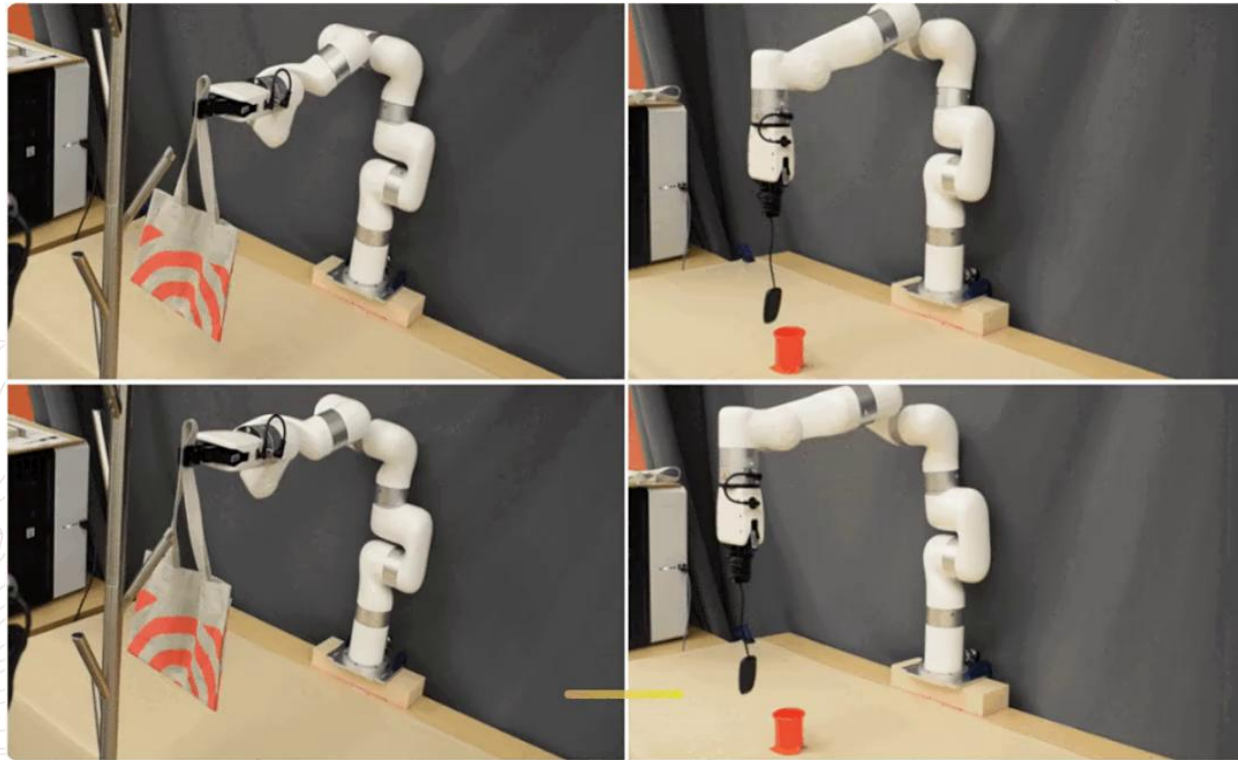


- The method relies on the demonstrator being an ‘expert’
  - Extending ROT to suboptimal, noisy and multimodal demonstrations would be an exciting problem to tackle.
  - Sometimes access to expert actions may not be present in some real-world scenarios, particularly when learning from humans.
  - Recent work on using inverse models to infer actions given observational data could alleviate this challenge.
- 



# Limitations

- The model struggles in cases where the task features aren't visually distinct.
- This limitation can be addressed by integrating more sensory modalities such as additional cameras, and tactile sensors in the observation space.





# Thank you for listening

主講人：吳雨欣

2023.2.23