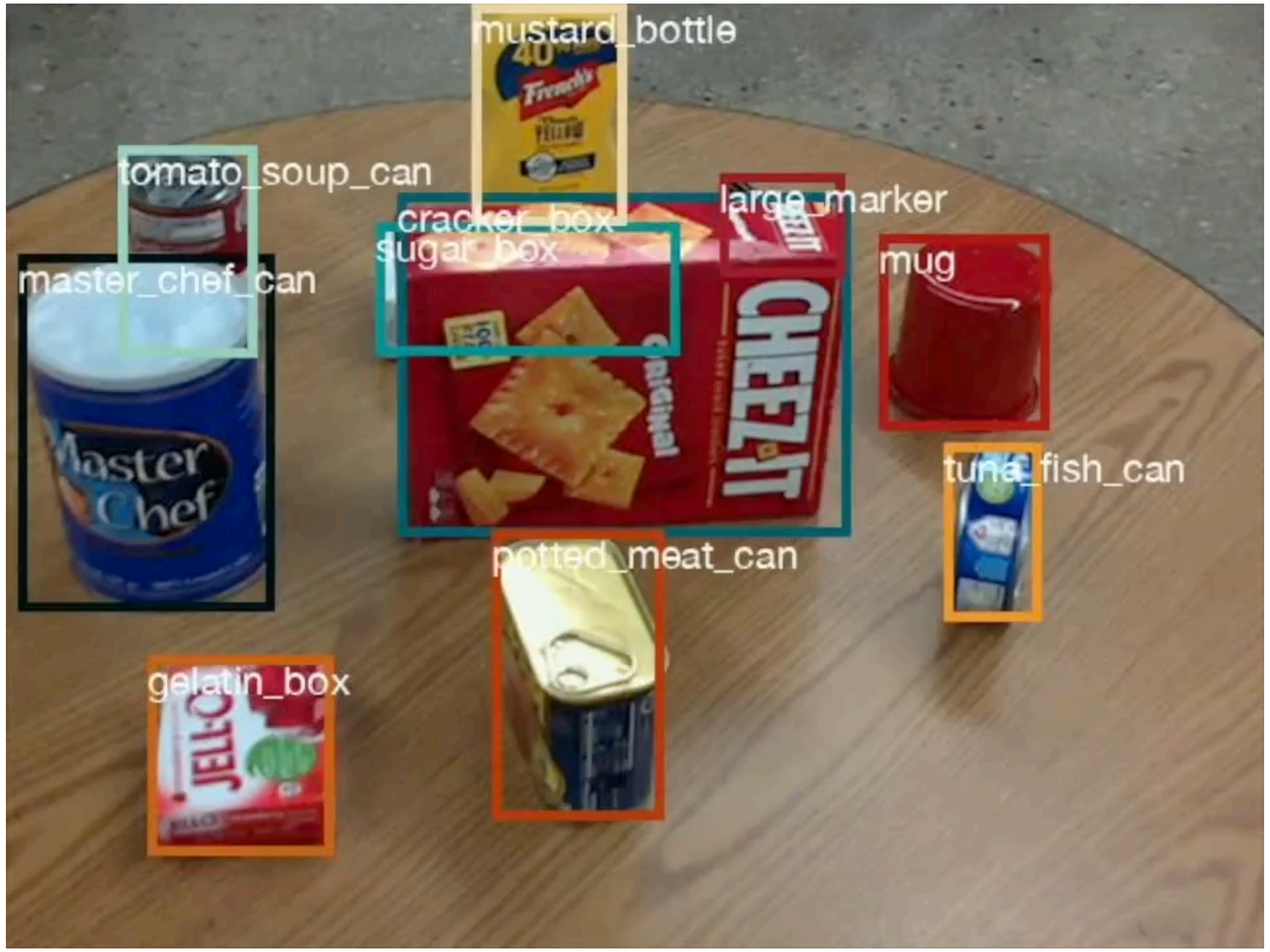
DeepRob

Lecture 15
Imitation Learning - II
University of Minnesota

HOW TO TRAIN YOUR DRAGON THE HIDDEN WORLD Kit Harington Auditions with Toothless

Project 3 — Releases today

- Instructions available on the website
  - Here: https://rpm-lab.github.io/CSCI5980-F24-DeepRob/projects/project3/
  - Uses [PROPS Detection dataset]()
- Implement CNN for classification and Faster R-CNN for detection
- Autograder will be available soon!
- **Due Monday, October 28th 11:59 PM CT**

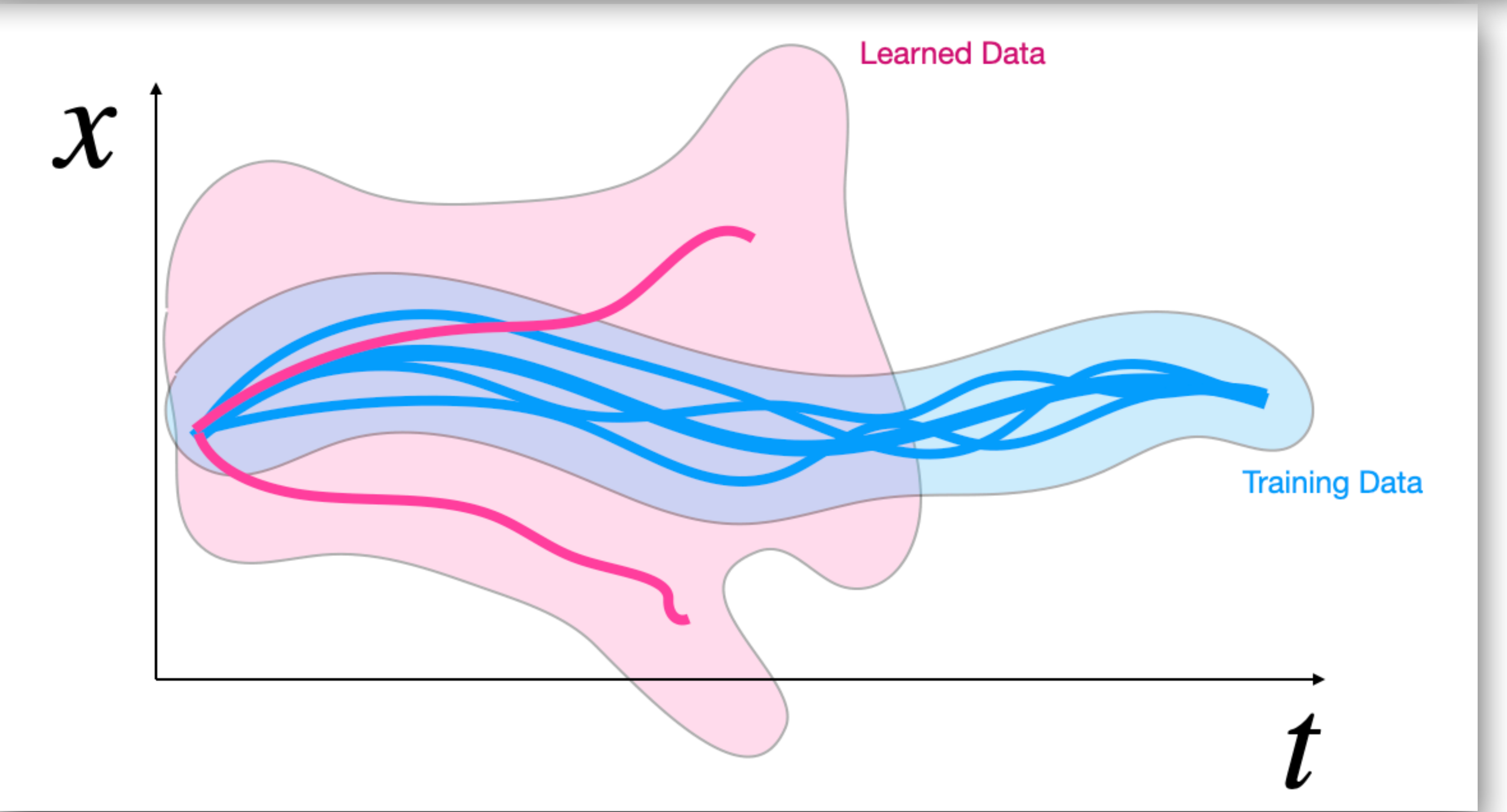


Last lecture

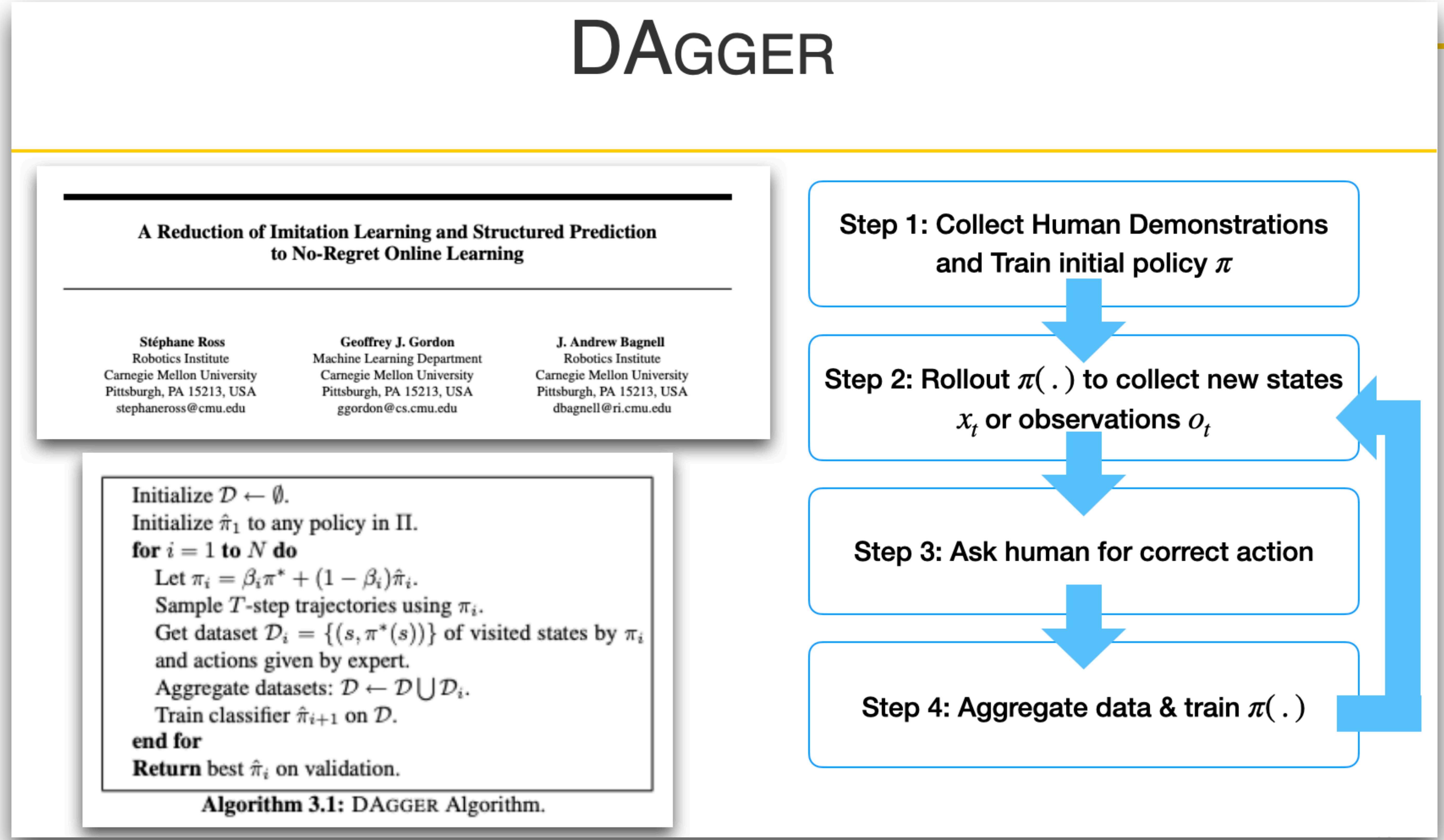
Challenges in going from **Prediction** to **Control**

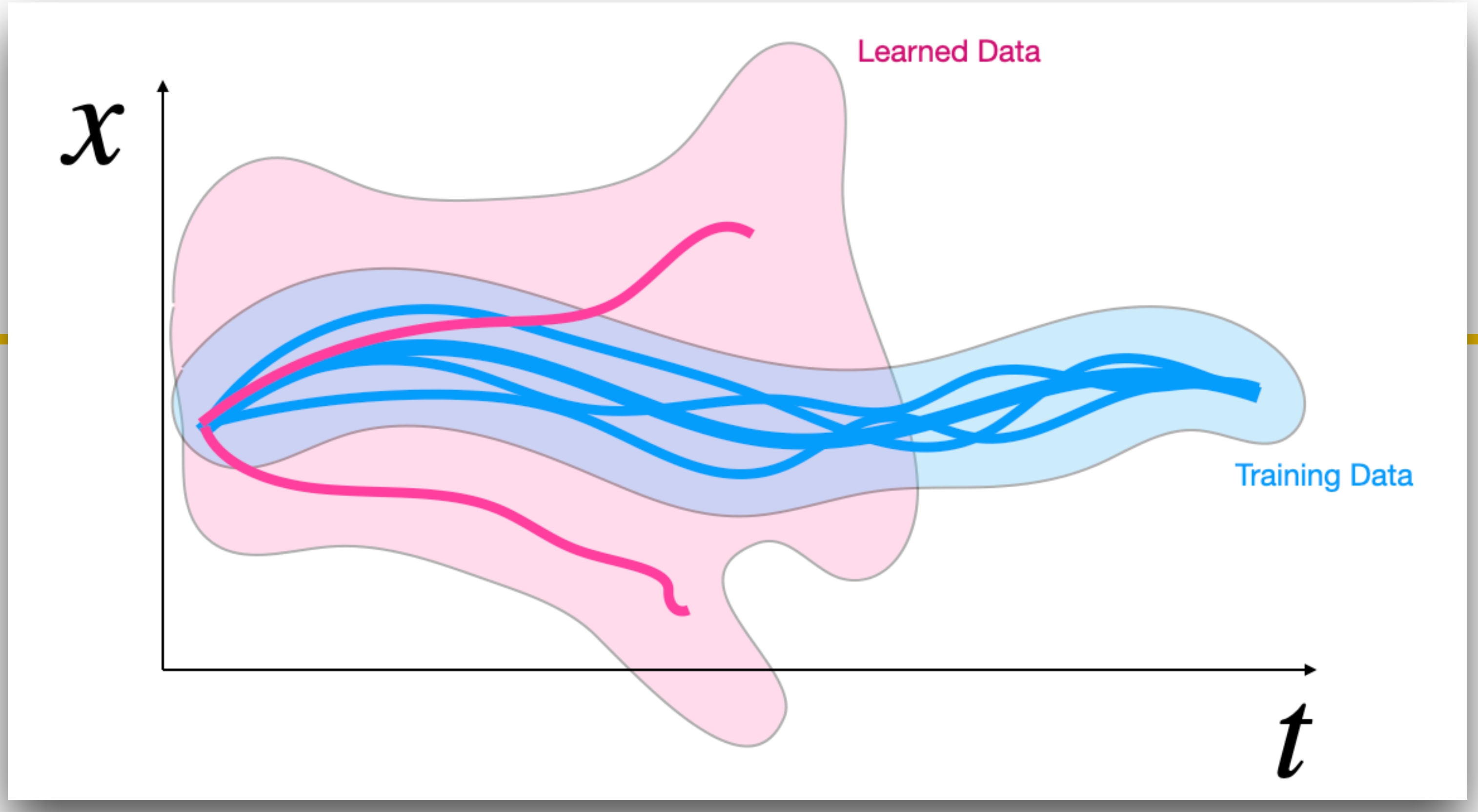
Prediction	Control
• Data is i.i.d distributed	• Data is not independent
• Ground truth supervision for the prediction is available	• Ground truth supervision is limited and mostly high-level
• Objective is to predict the right label or regress a value close to the ground truth	• Objective is to accomplish the task

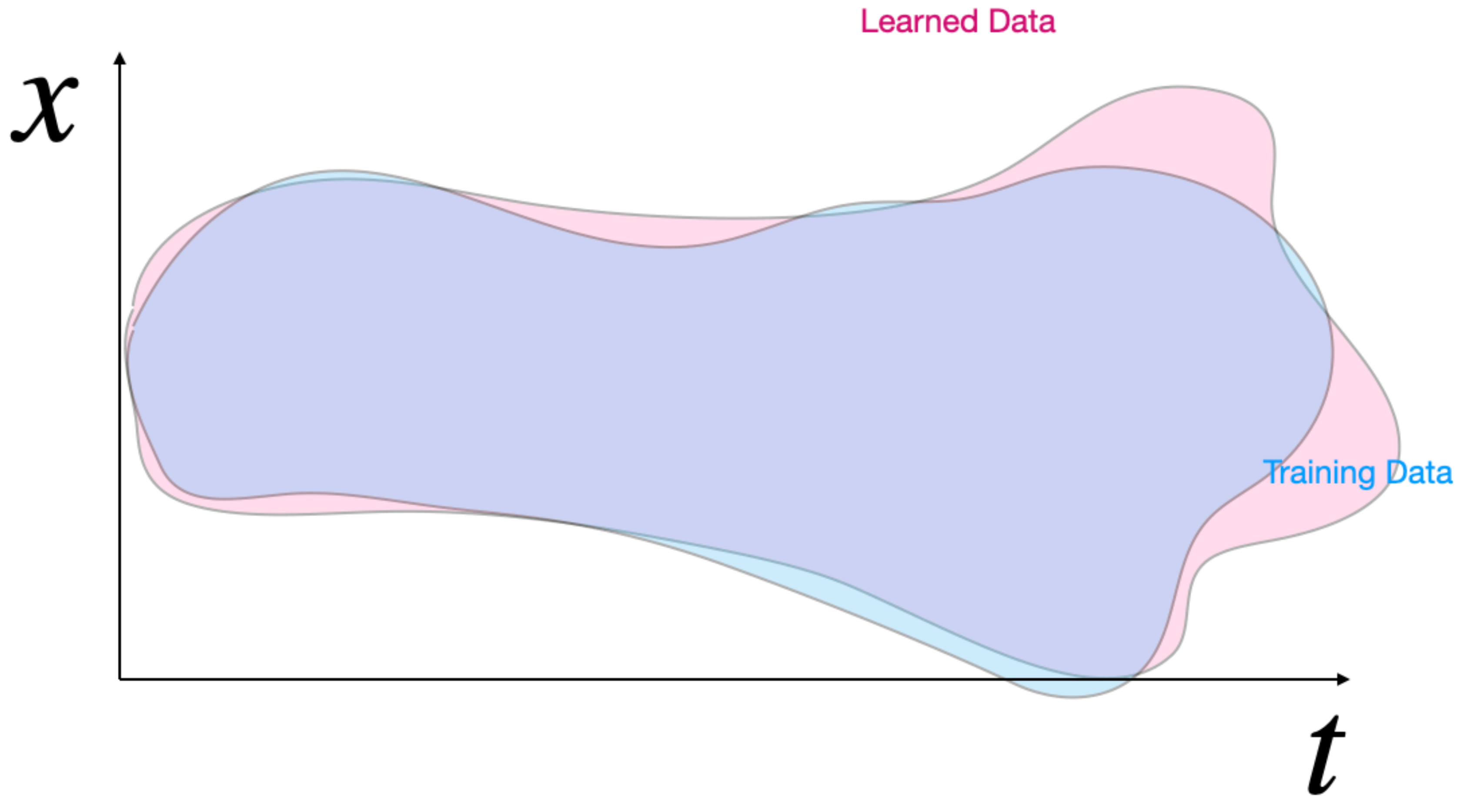
There is feedback and associated issues!



Last lecture







DAgger: Data Aggregation

Problem: the expert is asked to correct almost every time step, which can be expensive or infeasible.

HG-DAgger (Human-Gated DAgger)

2019 International Conference on Robotics and Automation (ICRA)
Palais des congres de Montreal, Montreal, Canada, May 20-24, 2019

HG-DAgger: Interactive Imitation Learning with Human Experts

Michael Kelly, Chelsea Sidrane, Katherine Driggs-Campbell, and Mykel J. Kochenderfer

Algorithm 1 HG-DAGGER

procedure HG-DAGGER($\pi_H, \pi_{N_1}, \mathcal{D}_{BC}$)
  $\mathcal{D} \leftarrow \mathcal{D}_{BC}$
  $\mathcal{I} \leftarrow []$
  for epoch $i = 1 : K$
    for rollout $j = 1 : M$
      for timestep $t \in T$ of rollout j
        if expert has control
          record expert labels into $\mathcal{D}_j$
        if expert is taking control
          record doubt into I_j
      $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_j$
      append $\mathcal{I}_j$ to $\mathcal{I}$
    train $\pi_{N_{i+1}}$ on $\mathcal{D}$
  $\tau \leftarrow f(\mathcal{I})$
  return $\pi_{N_{K+1}}, \tau$

Step 1: Collect Human Demonstrations and Train initial policy π

Step 2: Rollout $\pi(\,.\,)$ to collect new states x_t or observations o_t

Step 3: Ask human - *is this action acceptable?*
If yes, record the state action pair as is
else, record the current state and action by the human expert

Step 4: Aggregate data & train $\pi(\,.\,)$

"As a result, HG-DAGGER is not suitable for application in those realworld domains where the human expert cannot quickly identify and react to unsafe situations."

DAgger vs. HG-DAgger

DAgger

Step 1: Collect Human Demonstrations and Train initial policy π

Step 2: Rollout $\pi(\,.\,)$ to collect new states x_t or observations o_t

Step 3: Ask human for correct action

Step 4: Aggregate data & train $\pi(\,.\,)$

Behavior Cloning (Supervised Learning)

HG-DAgger

Step 1: Collect Human Demonstrations and Train initial policy π

Step 2: Rollout $\pi(\,.\,)$ to collect new states x_t or observations o_t

Step 3: Ask human - *is this action acceptable?*
If yes, record the state action pair as is
else, record the current state and action by the human expert

Step 4: Aggregate data & train $\pi(\,.\,)$

Ross, Stéphane, Geoffrey Gordon, and Drew Bagnell. "A reduction of imitation learning and structured prediction to no-regret online learning." In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627-635. JMLR Workshop and Conference Proceedings, 2011.

M. Kelly, C. Sidrane, K. Driggs-Campbell and M. J. Kochenderfer, "HG-DAgger: Interactive Imitation Learning with Human Experts," *2019 International Conference on Robotics and Automation (ICRA)*, Montreal, QC, Canada, 2019

DAgger vs. HG-DAgger

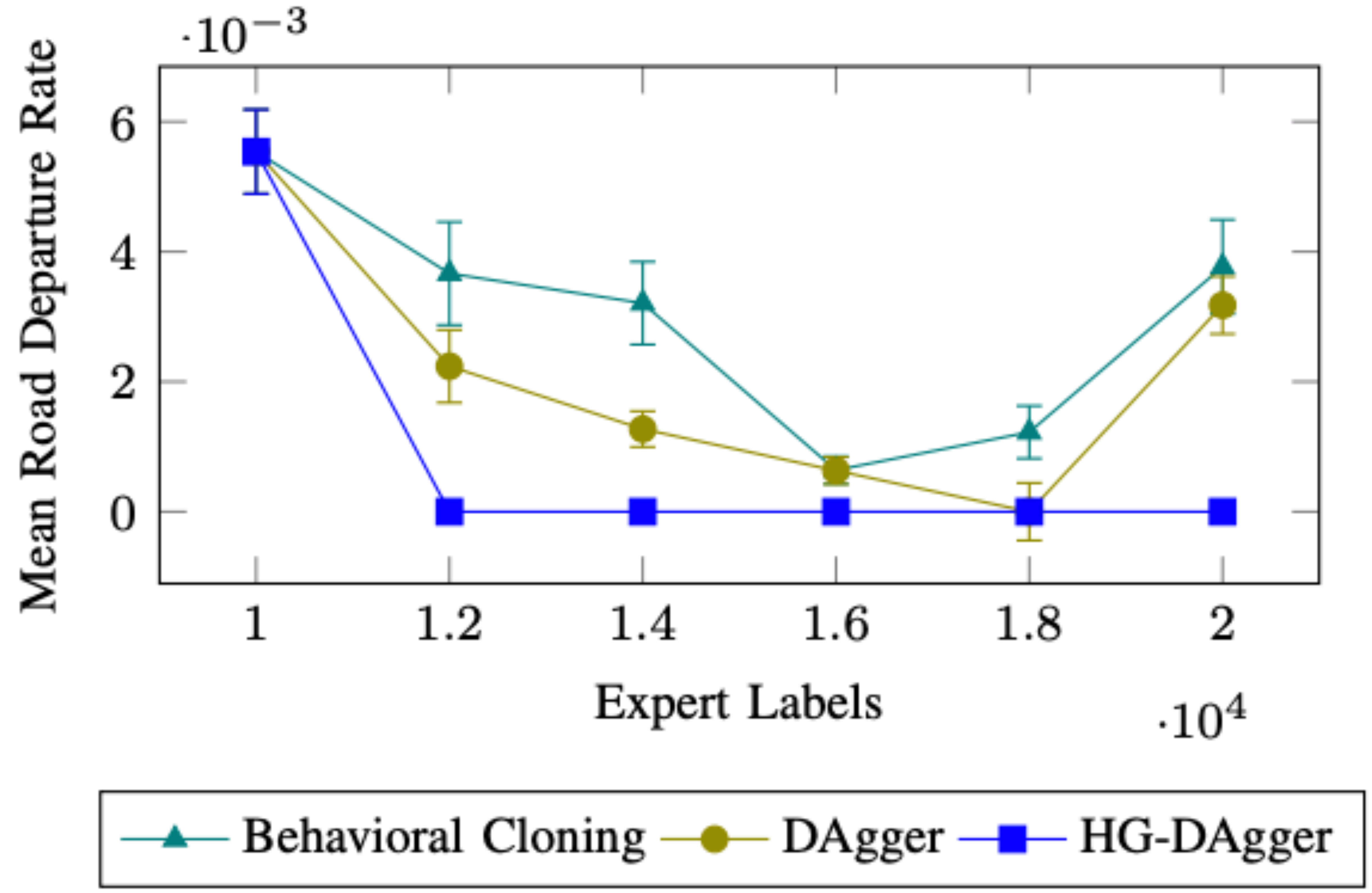


Fig. 3: Mean road departure rate per meter over training epochs. Error bars represent standard deviation.

M. Kelly, C. Sidrane, K. Driggs-Campbell and M. J. Kochenderfer, "HG-DAgger: Interactive Imitation Learning with Human Experts," *2019 International Conference on Robotics and Automation (ICRA)*, Montreal, QC, Canada, 2019

DAgger shows the human interventions are needed to tackle the distributional shift problem we have when we move from *prediction* tasks to *control* tasks

HG-DAgger shows that we reduce the number of human interventions i.e. only when it is needed.

But can we get more information than just corrected action from an expert?

In addition to correcting the actions, can we ask the expert to provide us with cost?

i.e. not all errors are equal, there must be cost associated with it

AGGREVATE: Aggregate Values to Imitate

Reinforcement and Imitation Learning via Interactive No-Regret Learning

Stéphane Ross **J. Andrew Bagnell**
stephaneross@cmu.edu dbagnell@ri.cmu.edu
The Robotics Institute
Carnegie Mellon University,
Pittsburgh, PA, USA

Algorithm 1 AGGREVATE: Imitation Learning with Cost-To-Go

Initialize $\mathcal{D} \leftarrow \emptyset$, $\hat{\pi}_1$ to any policy in Π.
for $i = 1$ **to** N **do**
  Let $\pi_i = \beta_i \pi^* + (1 - \beta_i)\hat{\pi}_i$ #Optionally mix in expert's own behavior.
  Collect m data points as follows:
  for $j = 1$ **to** m **do**
    Sample uniformly $t \in \{1, 2, \ldots, T\}$.
    Start new trajectory in some initial state drawn from initial state distribution
    Execute current policy π_i up to time $t - 1$.
    Execute some exploration action a_t in current state s_t at time t
    Execute expert from time $t + 1$ to T, and observe estimate of cost-to-go $\hat{Q}$ starting at time t
  end for
  Get dataset $\mathcal{D}_i = \{(s, t, a, \hat{Q})\}$ of states, times, actions, with expert's cost-to-go.
  Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \bigcup \mathcal{D}_i$.
  Train cost-sensitive classifier $\hat{\pi}_{i+1}$ on $\mathcal{D}$
    (*Alternately: use any online learner on the data-sets $\mathcal{D}_i$ in sequence to get $\hat{\pi}_{i+1}$*)
end for
Return best $\hat{\pi}_i$ on validation.

AGGREVATE: Aggregate Values to Imitate

Reinforcement and Imitation Learning via Interactive No-Regret Learning

Stéphane Ross **J. Andrew Bagnell**
stephaneross@cmu.edu dbagnell@ri.cmu.edu
The Robotics Institute
Carnegie Mellon University,
Pittsburgh, PA, USA

Algorithm 1 AGGREVATE: Imitation Learning with Cost-To-Go

Initialize $\mathcal{D} \leftarrow \emptyset$, $\hat{\pi}_1$ to any policy in Π.
for $i = 1$ **to** N **do**
  Let $\pi_i = \beta_i \pi^* + (1 - \beta_i)\hat{\pi}_i$ #Optionally mix in expert's own behavior.
  Collect m data points as follows:
  for $j = 1$ **to** m **do**
    Sample uniformly $t \in \{1, 2, \ldots, T\}$.
    Start new trajectory in some initial state drawn from initial state distribution
    Execute current policy π_i up to time $t - 1$.
    Execute some exploration action a_t in current state s_t at time t
    Execute expert from time $t + 1$ to T, and observe estimate of cost-to-go $\hat{Q}$ starting at time t
  end for
  Get dataset $\mathcal{D}_i = \{(s, t, a, \hat{Q})\}$ of states, times, actions, with expert's cost-to-go.
  Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \bigcup \mathcal{D}_i$.
  Train cost-sensitive classifier $\hat{\pi}_{i+1}$ on $\mathcal{D}$
    (Alternately: use any online learner on the data-sets $\mathcal{D}_i$ in sequence to get $\hat{\pi}_{i+1}$)
end for
Return best $\hat{\pi}_i$ on validation.

Step 1: Collect Human Demonstrations and Train initial policy π

Step 2: Rollout $\pi(\,.\,)$ to collect new states x_t or observations o_t

Step 3: Ask human expert:
1. What is the expected future cost (or error) from the current state if the agent were to follow its own policy π
2. What is an optimal action from this current state?

Step 4: Aggregate data (state-action pairs, cost-to-go estimates) & train $\pi(\,.\,)$ to minimize both the immediate cost and the future cost

Cost-to-go in seems familiar to *Reward signal* or *Value function* in Reinforcement learning!

So potentially we can combine this with Reinforcement learning.

Bootstrapping RL via Imitation.

r to Reward signal

rcement learning!

So potentially

L via Imitation.

Published as a conference paper at ICLR 2018

TRUNCATED HORIZON POLICY SEARCH: COMBINING REINFORCEMENT LEARNING & IMITATION LEARNING

Wen Sun
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA, USA
wensun@cs.cmu.edu

J. Andrew Bagnell
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA, USA
dbagnell@cs.cmu.edu

Byron Boots
School of Interactive Computing
Georgia Institute of Technology

Fast Policy Learning through Imitation and Reinforcement

Ching-An Cheng
Georgia Tech
Atlanta, GA 30332

Xinyan Yan
Georgia Tech
Atlanta, GA 30332

Nolan Wagener
Georgia Tech
Atlanta, GA 30332

Byron Boots
Georgia Tech
Atlanta, GA 30332

Dual Policy Iteration

Wen Sun[1], Geoffrey J. Gordon[1], Byron Boots[2], and J. Andrew Bagnell[3]

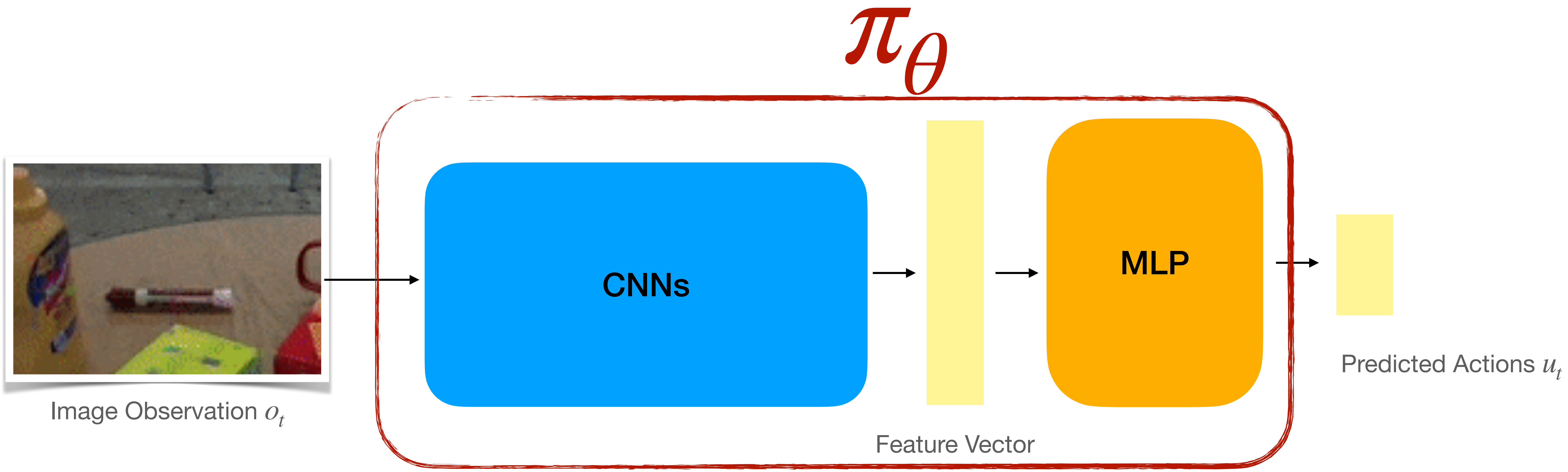
[1]School of Computer Science, Carnegie Mellon University, USA
[2]College of Computing, Georgia Institute of Technology, USA
[3]Aurora Innovation, USA
{wensun, ggordon, dbagnell}@cs.cmu.edu, bboots@cc.gatech.edu

Coming back to Robot Manipulation

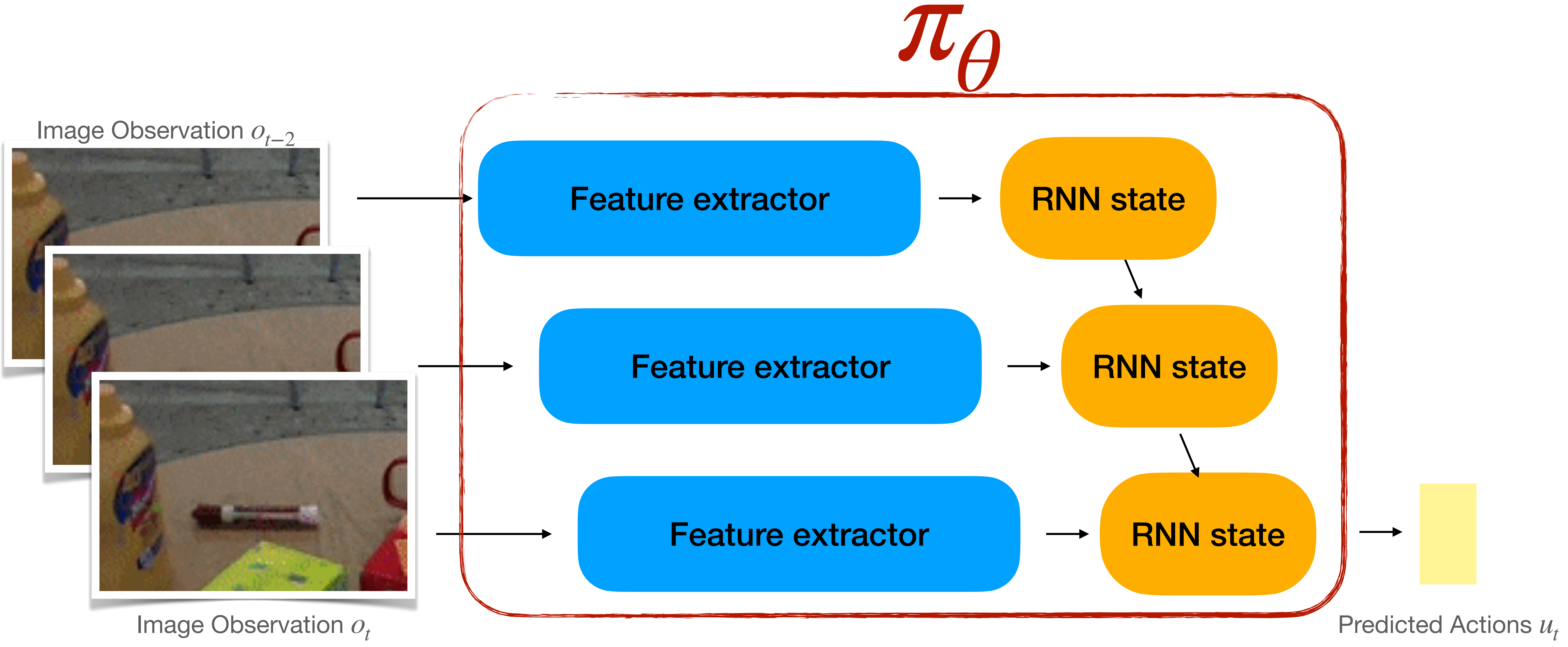
Evolving Policy Learning Methods …

- BC-MLP
- BC-RNN
- BeT- Behavior Transformers
- IBC Implicit behavior cloning
- Diffusion Policy
- Action Chuncking Transformers

BC-MLP (Behavior Cloning with Multi-Layered Perceptron)



BC-RNN (Behavior Cloning with Recurrent Neural Network)



BeT- Behavior Transformers

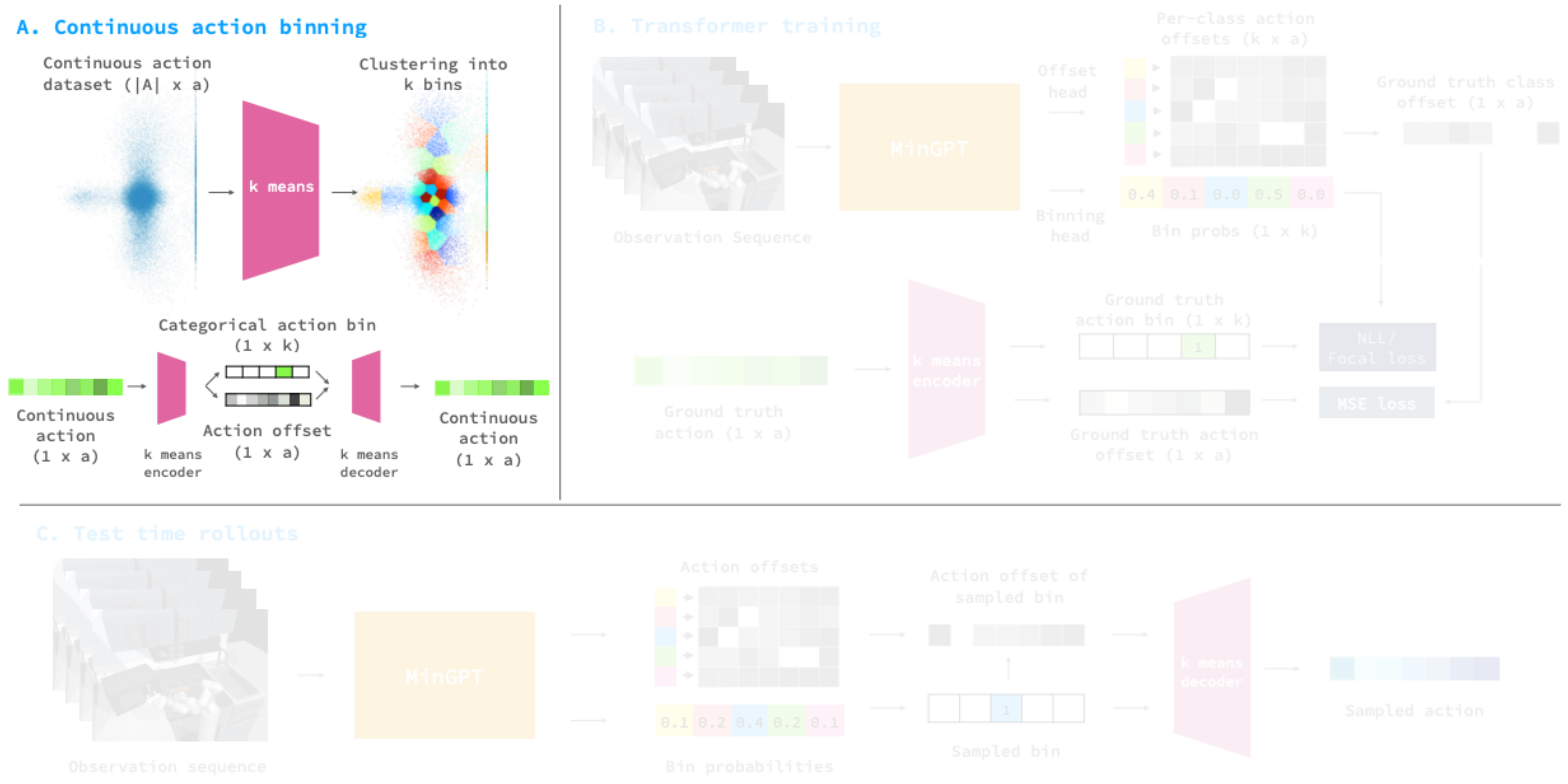


Figure 3: Architecture of Behavior Transformer. (A) The continuous action binning using k-means algorithm that lets BeT split every action into a discrete bin and a continuous offset, and later combine them into one full action. (B) Training BeT using demonstrations offline; each ground truth action provides a ground truth bin and residual action, which is used to train the minGPT trunk with its binning and action offset heads. (C) Rollouts from BeT in test time, where it first chooses a bin and then picks the corresponding offset to reconstruct a continuous action.

Shafiullah, Nur Muhammad, Zichen Cui, Ariuntuya Arty Altanzaya, and Lerrel Pinto. "Behavior transformers: Cloning $ k $ modes with one stone." *Advances in neural information processing systems* 35 (2022): 22955-22968.

IBC: Implicit Behavior Cloning

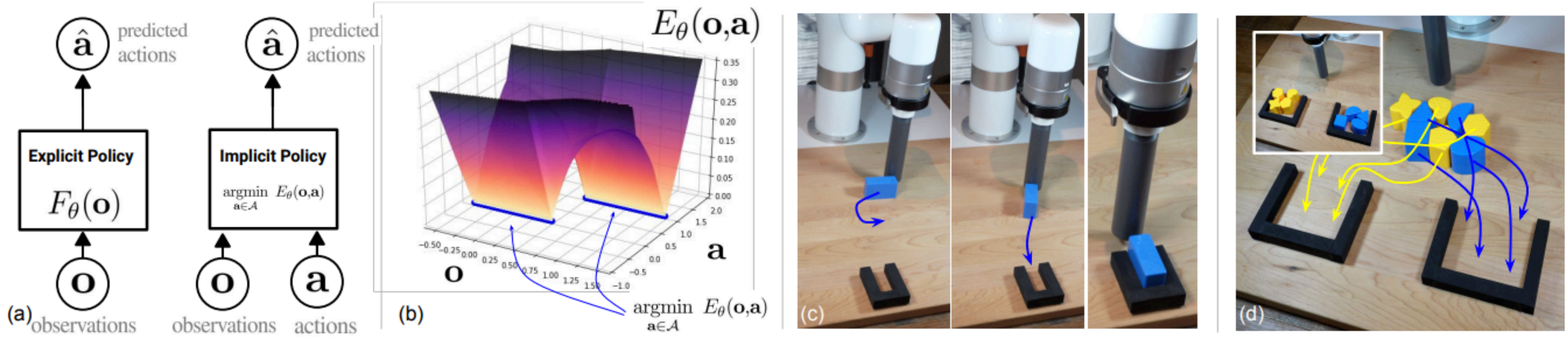


Figure 1. (a) In contrast to explicit policies, implicit policies leverage parameterized energy functions that take both observations (e.g. images) and actions as inputs, and optimize for actions that minimize the energy landscape (b). For learning complex, closed-loop, multimodal visuomotor tasks such as precise block insertion (c) and sorting (d) from human demonstrations, implicit policies perform substantially better than explicit ones.

Florence, Pete, Corey Lynch, Andy Zeng, Oscar A. Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. "Implicit behavioral cloning." In *Conference on Robot Learning*, pp. 158-168. PMLR, 2022.

Diffusion Policy

Chi, Cheng, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. "Diffusion policy: Visuomotor policy learning via action diffusion." *The International Journal of Robotics Research* (2023): 02783649241273668.

ACT: Action Chunking with Transformers

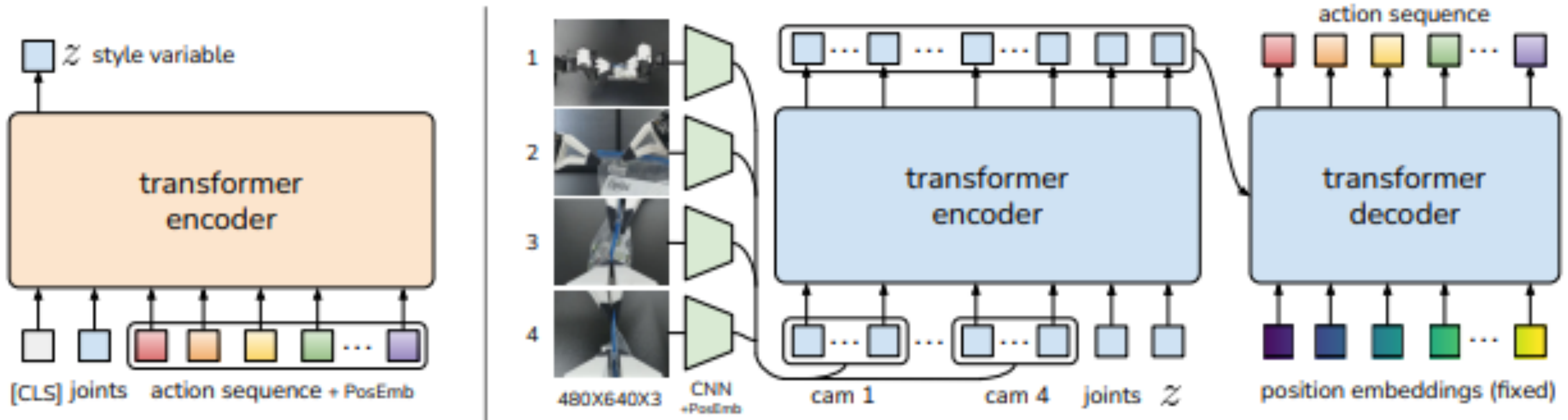


Fig. 4: *Architecture of Action Chunking with Transformers (ACT).* We train ACT as a Conditional VAE (CVAE), which has an encoder and a decoder. *Left:* The encoder of the CVAE compresses action sequence and joint observation into z, the style variable. The encoder is discarded at test time. *Right:* The decoder or policy of ACT synthesizes images from multiple viewpoints, joint positions, and z with a transformer encoder, and predicts a sequence of actions with a transformer decoder. z is simply set to the mean of the prior (i.e. zero) at test time.

Zhao, Tony Z., Vikash Kumar, Sergey Levine, and Chelsea Finn. "Learning fine-grained bimanual manipulation with low-cost hardware." *arXiv preprint arXiv:2304.13705* (2023).

ACT: Action Chunking with Transformers

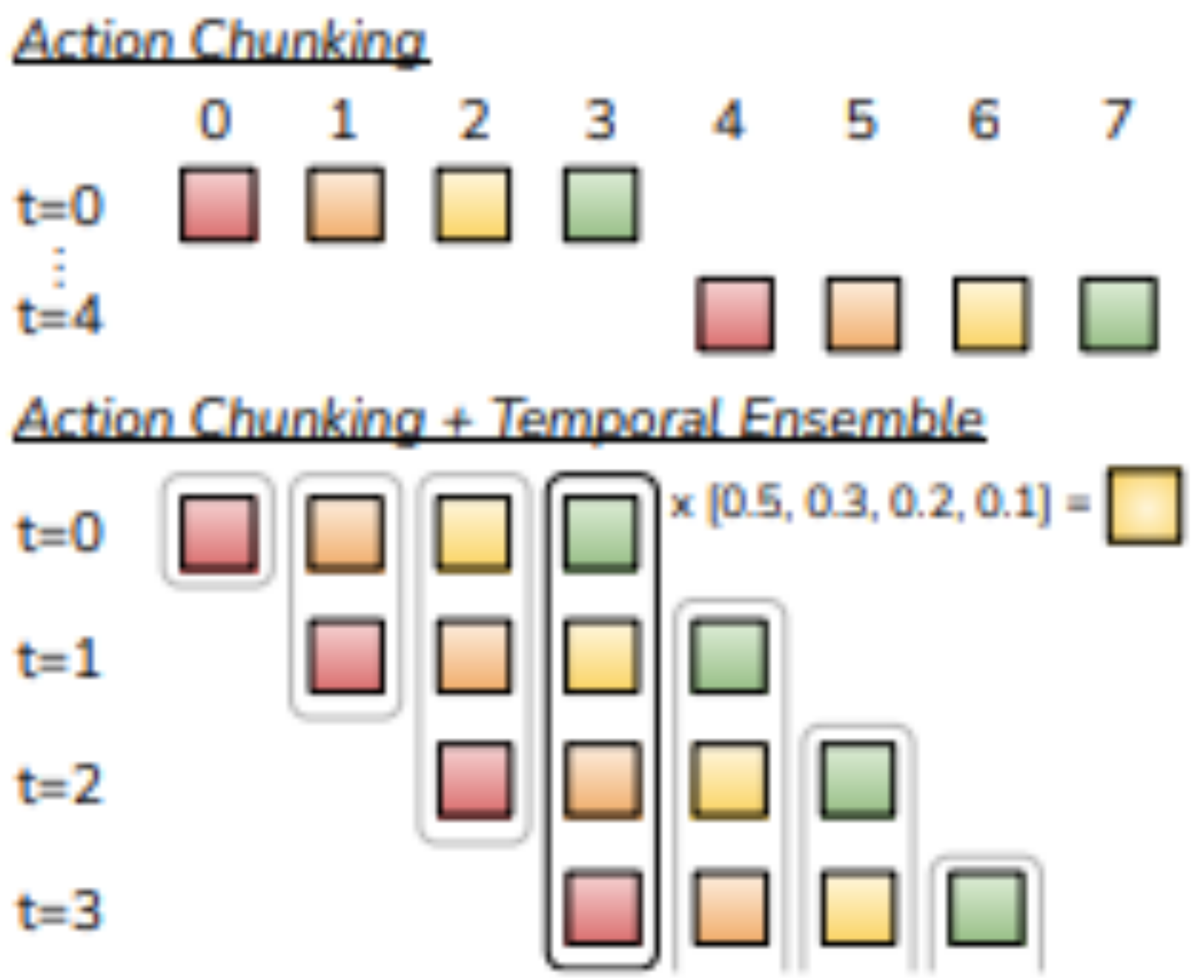
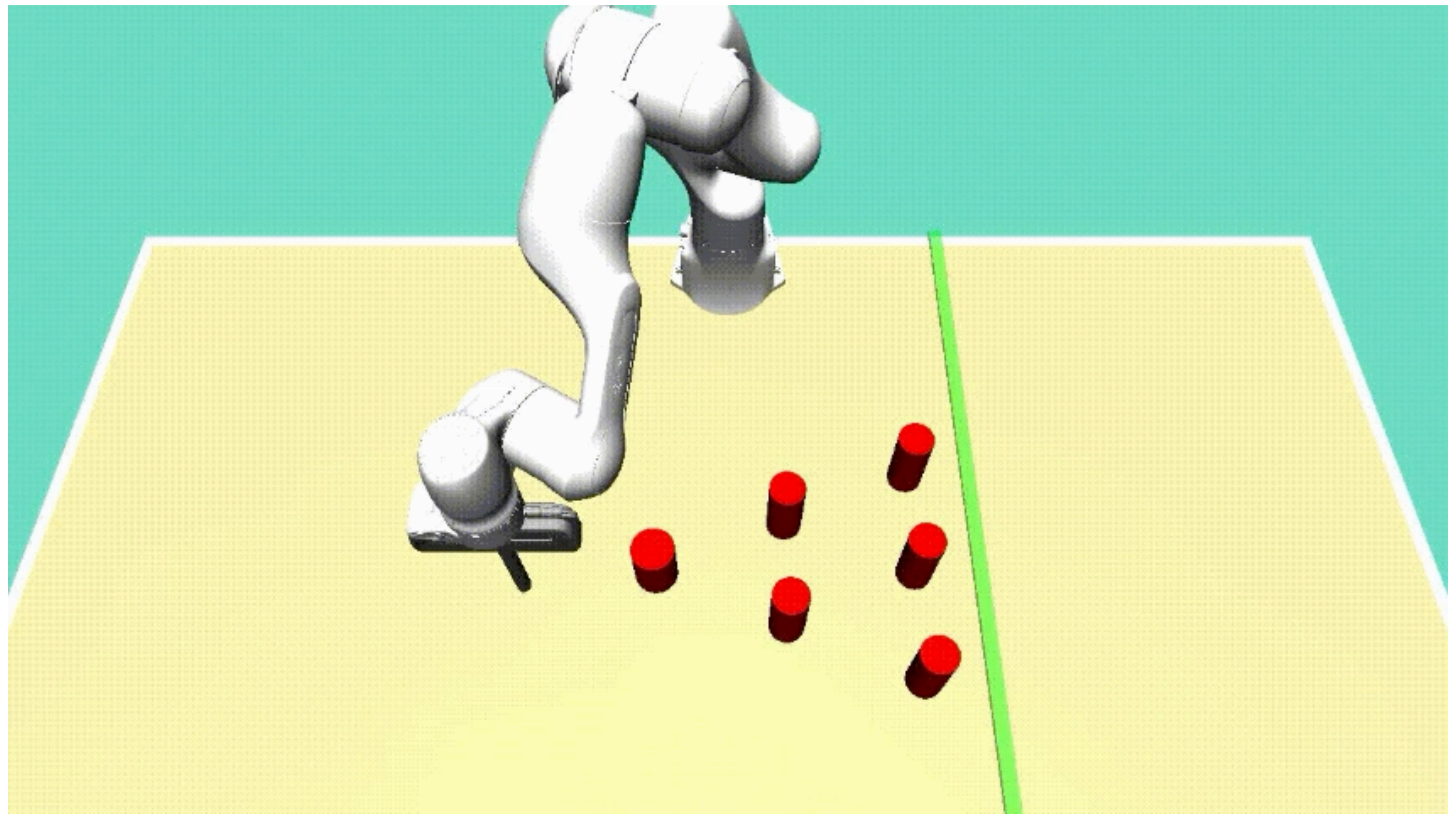
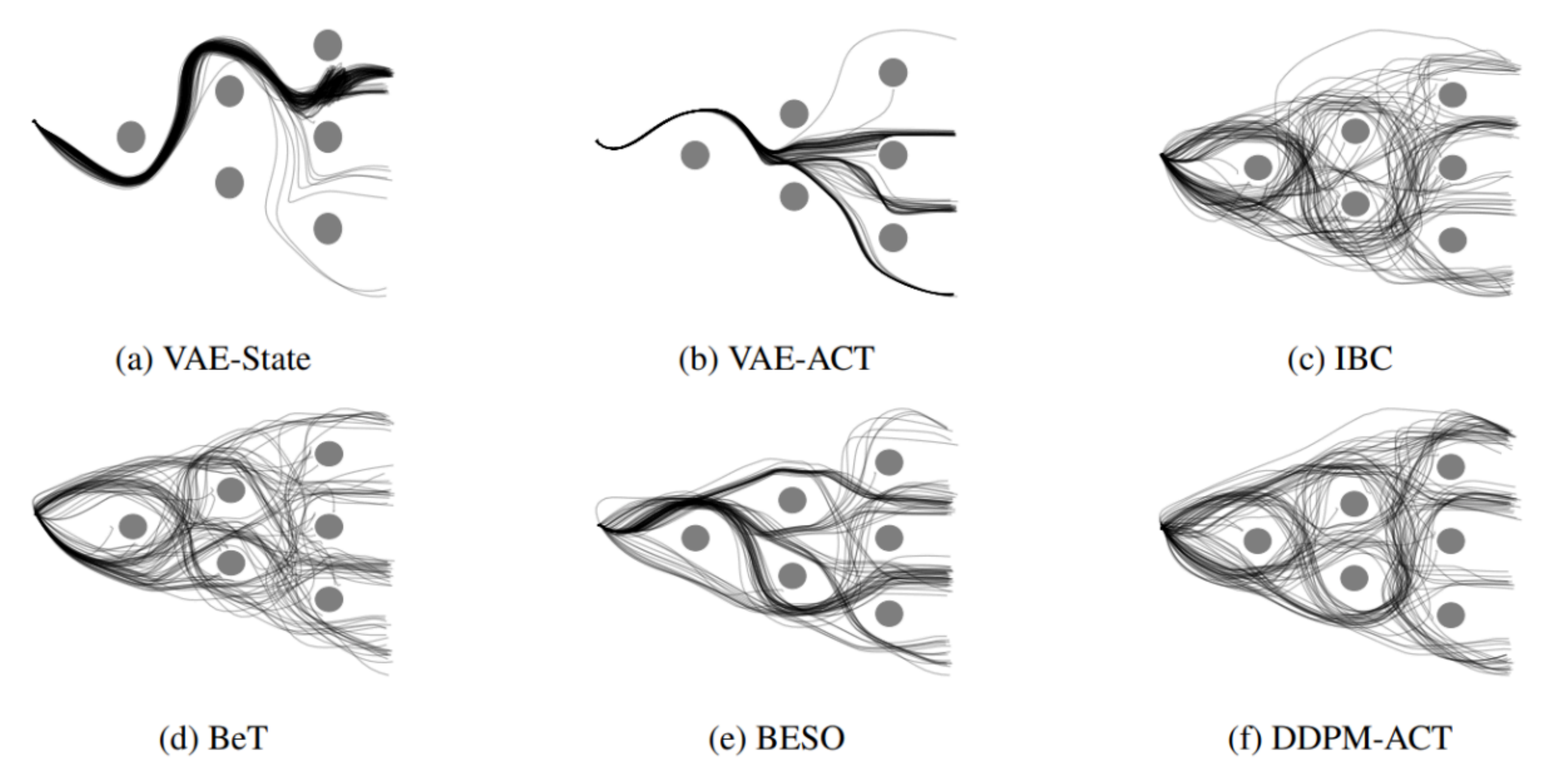


Fig. 5: We employ both Action Chunking and Temporal Ensembling when applying actions, instead of interleaving observing and executing.

Zhao, Tony Z., Vikash Kumar, Sergey Levine, and Chelsea Finn. "Learning fine-grained bimanual manipulation with low-cost hardware." *arXiv preprint arXiv:2304.13705* (2023).

Handling Diverse Behaviors





Jia, Xiaogang, Denis Blessing, Xinkai Jiang, Moritz Reuss, Atalay Donat, Rudolf Lioutikov, and Gerhard Neumann. "Towards diverse behaviors: A benchmark for imitation learning with human demonstrations." *arXiv preprint arXiv:2402.14606* (2024).

Next Lecture: Transformers

DeepRob

Lecture 15
Imitation Learning - II
University of Minnesota

HOW TO TRAIN YOUR DRAGON THE HIDDEN WORLD Kit Harington Auditions with Toothless