Imitation Learning

Problem Statement

- Can we learn a policy from expert demonstrations?
  - RL typically learns from scratch.
  - But we often learn by imitating others: why not for RL agents?

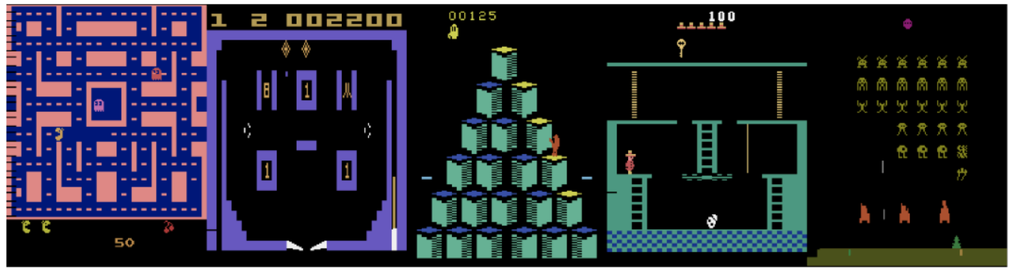
Early Imitation Learning

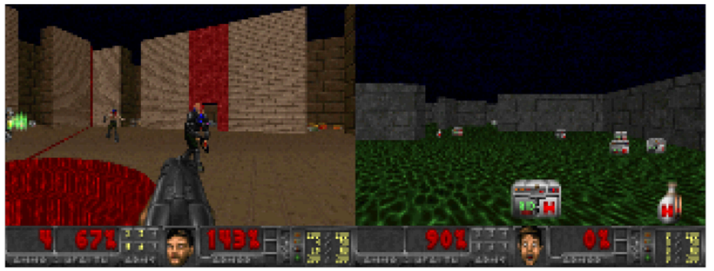
Pomerleau. “ALVINN: AN AUTONOMOUS LAND VEHICLE IN A NEURAL NETWORK”, 1989

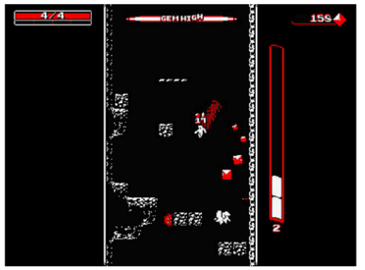
Use cases

Samak et al., 2020

Zhang et al., 2017

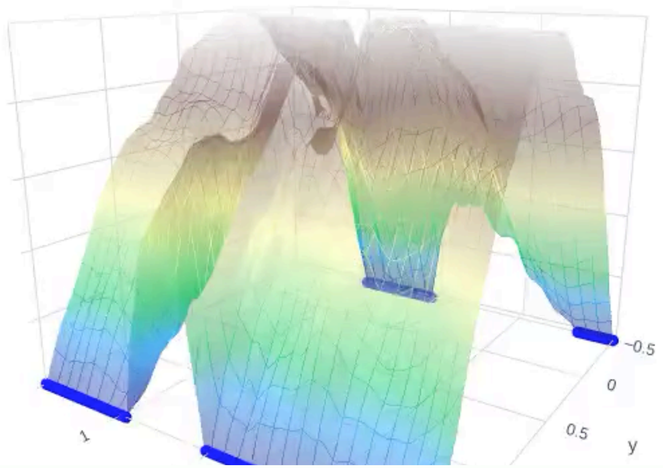


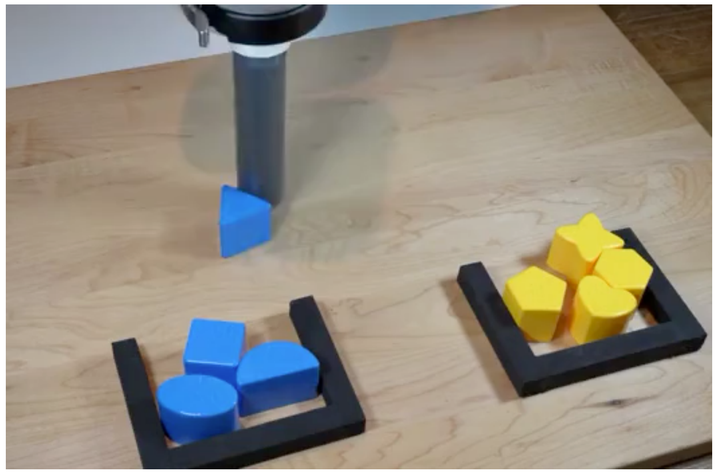


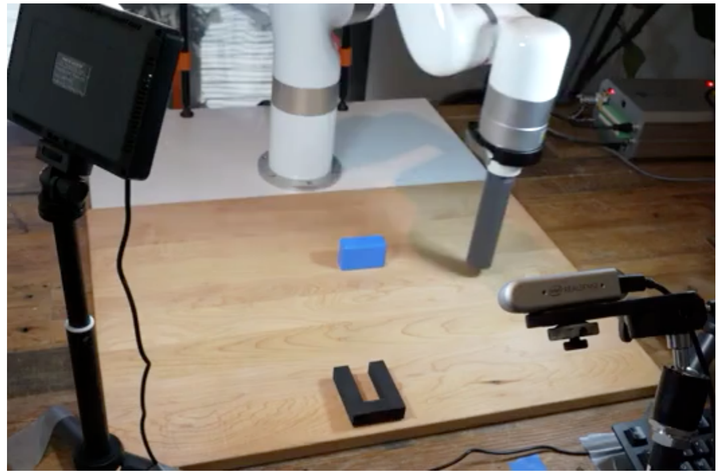


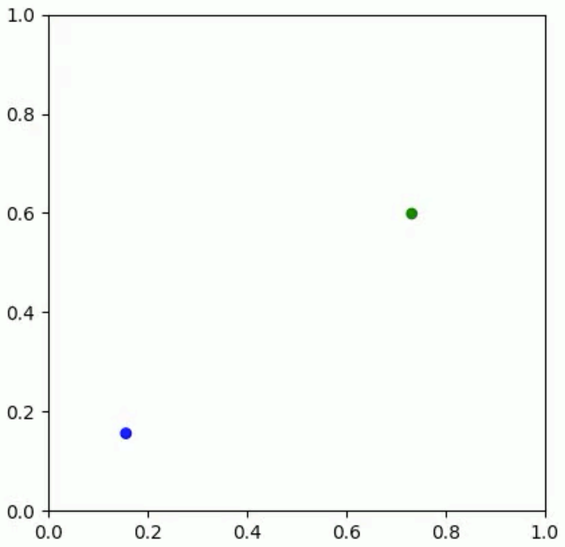
Kanervisto et al. “Benchmarking End-to-End Behavioural Cloning on Video Games”, 2020

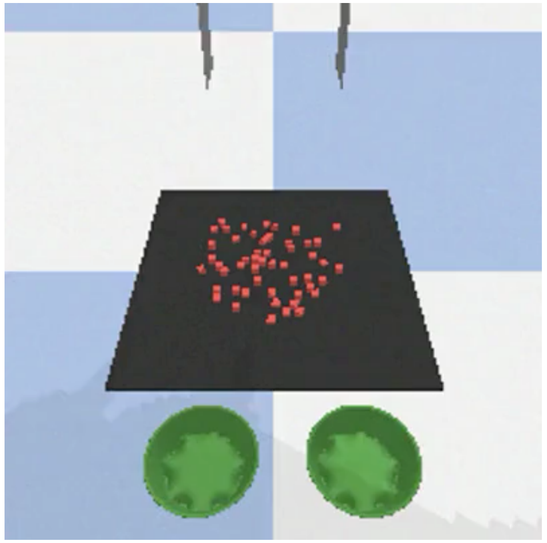
Use Cases











Florence et al. “Implicit behavioral cloning”, 2022

Vanilla Behavior Cloning

- We can formulate the first version of the behavior cloning as follows:
  - Let's assume that the data $D = \{(s_1, a_1), (s_2, a_2), \cdots, (s_n, a_n)\}$.
  - $L(\theta) = \sum_{D=\{s_i,a_i\}} |\blacksquare - \blacksquare|^2$
  - or $L(\theta) = \sum_{D=\{s_i,a_i\}} - log\pi_\theta(a \mid s)$
  - Quiz: is this reinforcement learning or supervised learning?

Vanilla Behavior Cloning

- Collect expert data D
- Optimize the policy parameters θ w.r.t. the loss $L(\theta) = \sum_{D=\{s_i,a_i\}} |\pi_\theta(s_i) - a_i|^2$

Will it work?

- Yes. We just saw the demos.
- No. Training and testing data distributions can be mismatched.

Pomerleau. ALVINN. 1989

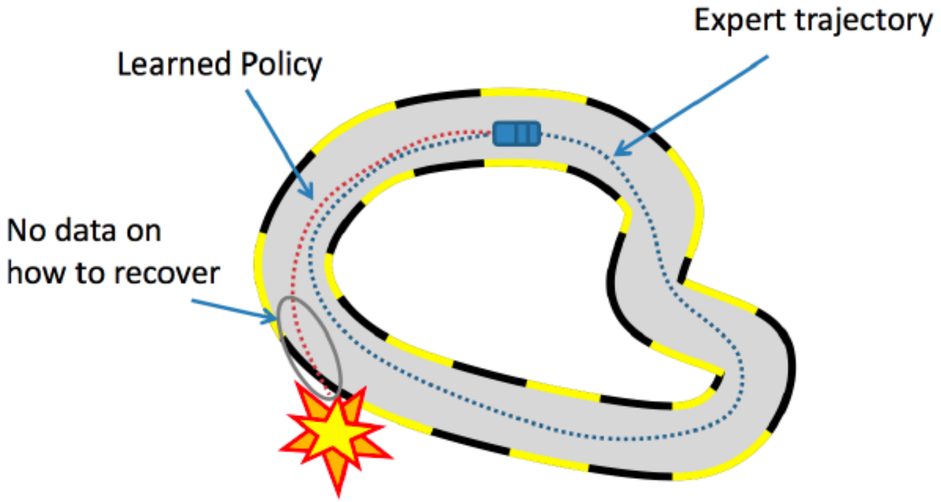
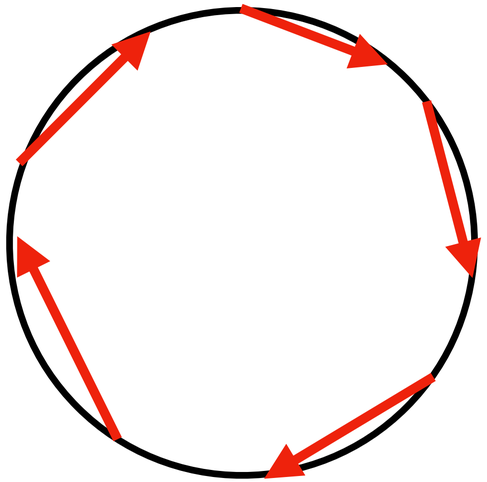


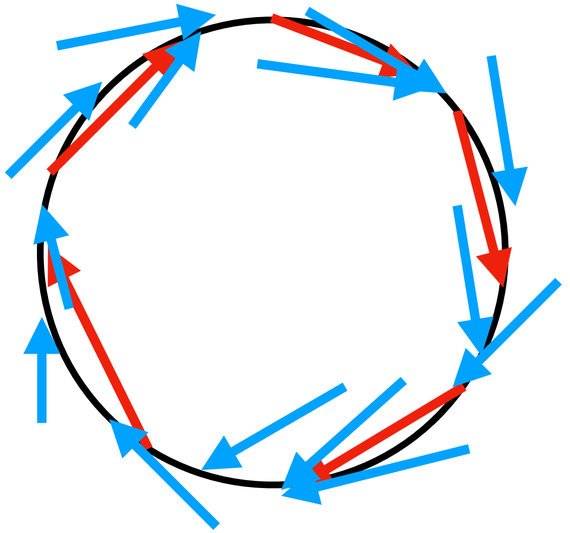
Image from Stanford CS 234

Noise Injection

- We can inject noise to obtain more richer expert demonstrations.
  - We can inject noises to expert's control.
  - Quiz: why not inject noises to the state space?



Expert data without noises



Expert data without noises

Noise Injection

- Then how much we can inject noises? We can iteratively optimize it.
- While not converge:
  - Collect the expert data with the noise level ψ
  - Update the noise level to match the robot and expert's distribution

$$\hat{\psi}_{k+1} = \underset{\psi}{\operatorname{argmin}} E_{p(\xi|\pi_{\theta^*},\psi_k)} - \sum_{t=0}^{T-1} \log\left[\pi_{\theta^*}(\pi_{\hat{\theta}}(\mathbf{x}_t)|\mathbf{x}_t,\psi)\right]$$

  - Optimize the policy parameter w.r.t. the BC loss.

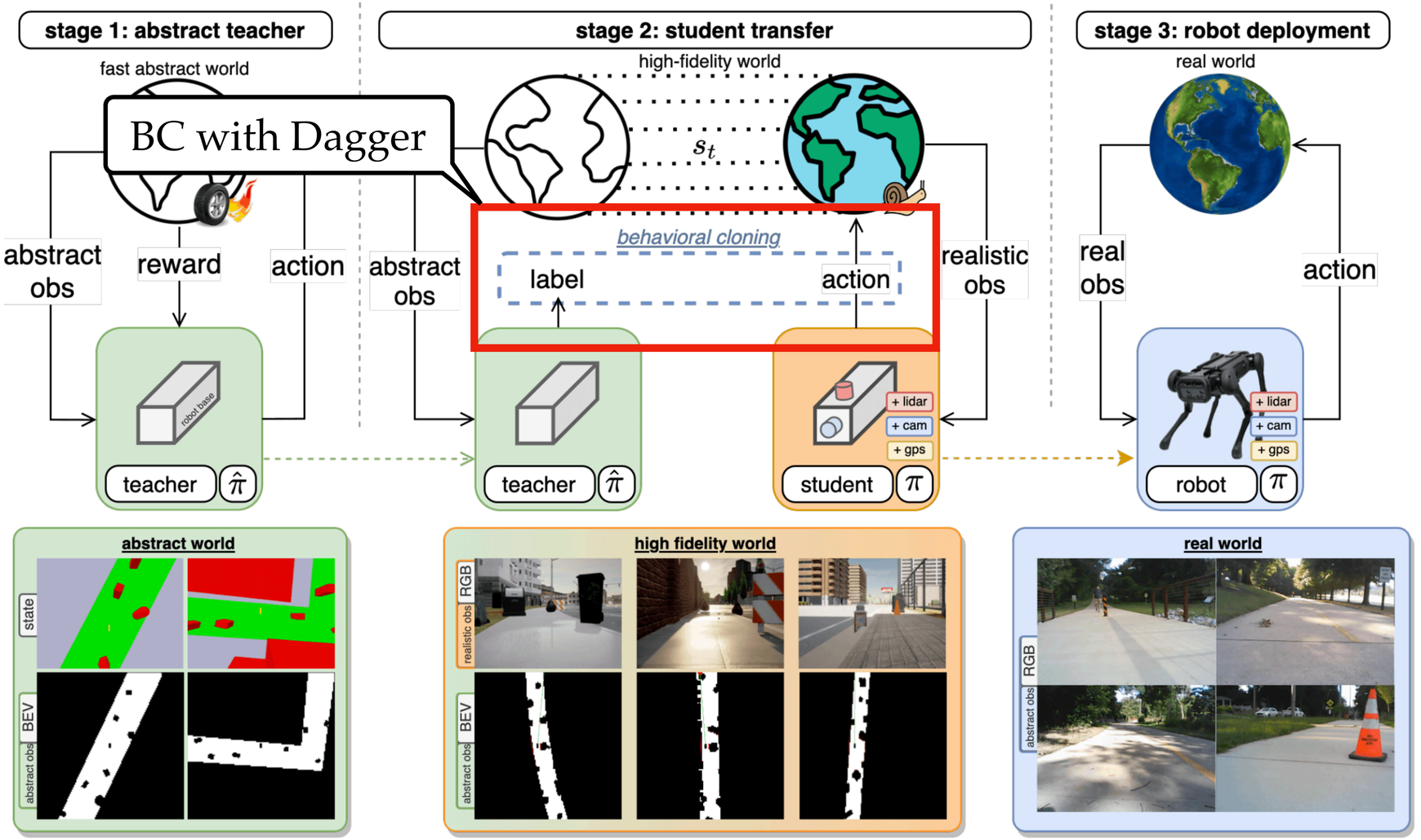
Lee et al. "DART: Noise Injection for Robust Imitation Learning", 2017

Dataset Aggregation

- What if we strategically collect data?
- While not converge:
  - Train π_θ from human demonstration $D = \{(s_1, a_1), (s_2, a_2), \cdots, (s_n, a_n)\}$
  - Run π_θ to get a list of states: $\{s_{n+1}, s_{n+2}, \cdots, s_{n+m}\}$
  - Ask an expert to label the actions: $\{o_{n+1}, o_{n+2}, \cdots, o_{n+m}\}$
  - Augment the data D

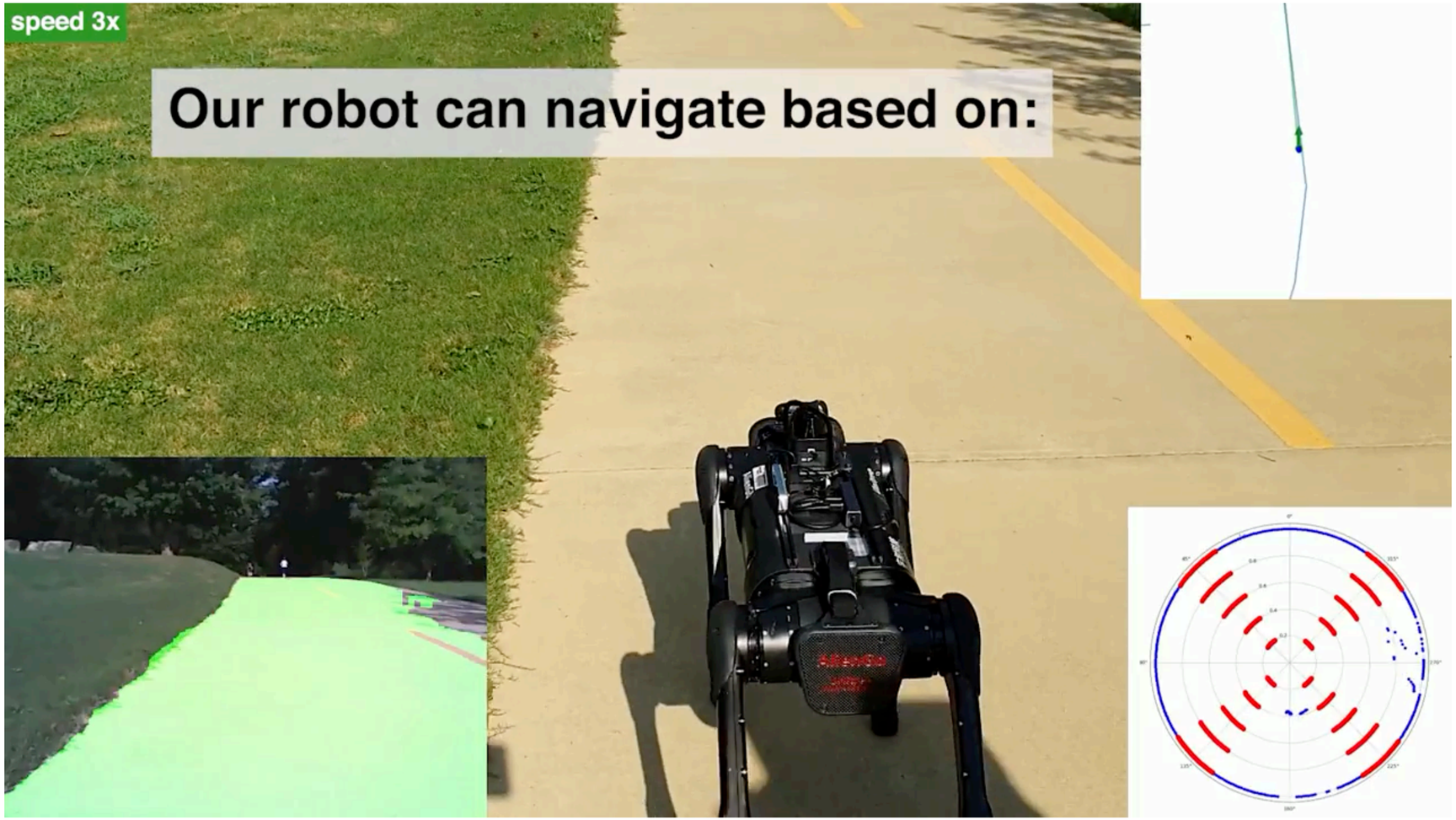
Ross et al. 2011

Dataset Aggregation



Sorokin et al. “Learning to Navigate Sidewalks in Outdoor Environments”, 2022

Dataset Aggregation



Sorokin et al. “Learning to Navigate Sidewalks in Outdoor Environments”, 2022

Connection to Inverse RL

- Inverse RL infers the reward function from the given expert behaviors.
- Combined with RL, it becomes Generative Adversarial Imitation Learning.

Algorithm 1 Generative adversarial imitation learning

Expert trajectories $\tau_E \sim \pi_E$, initial policy and discriminator parameters θ_0, w_0

$0, 1, 2, \ldots$ **do**

Sample trajectories $\tau_i \sim \pi_{\theta_i}$

Update the discriminator parameters from w_i to w_{i+1} with the gradient

$$\hat{\mathbb{E}}_{\tau_i}[\nabla_w \log(D_w(s,a))] + \hat{\mathbb{E}}_{\tau_E}[\nabla_w \log(1 - D_w(s,a))] \tag{17}$$

Take a policy step from θ_i to θ_{i+1}, using the TRPO rule with cost function $\log(D_{w_{i+1}}(s,a))$. Specifically, take a KL-constrained natural gradient step with

$$\hat{\mathbb{E}}_{\tau_i}[\nabla_\theta \log \pi_\theta(a|s) Q(s,a)] - \lambda \nabla_\theta H(\pi_\theta), \tag{18}$$
$$\text{where } Q(\bar{s},\bar{a}) = \hat{\mathbb{E}}_{\tau_i}[\log(D_{w_{i+1}}(s,a)) \mid s_0 = \bar{s}, a_0 = \bar{a}]$$

end for

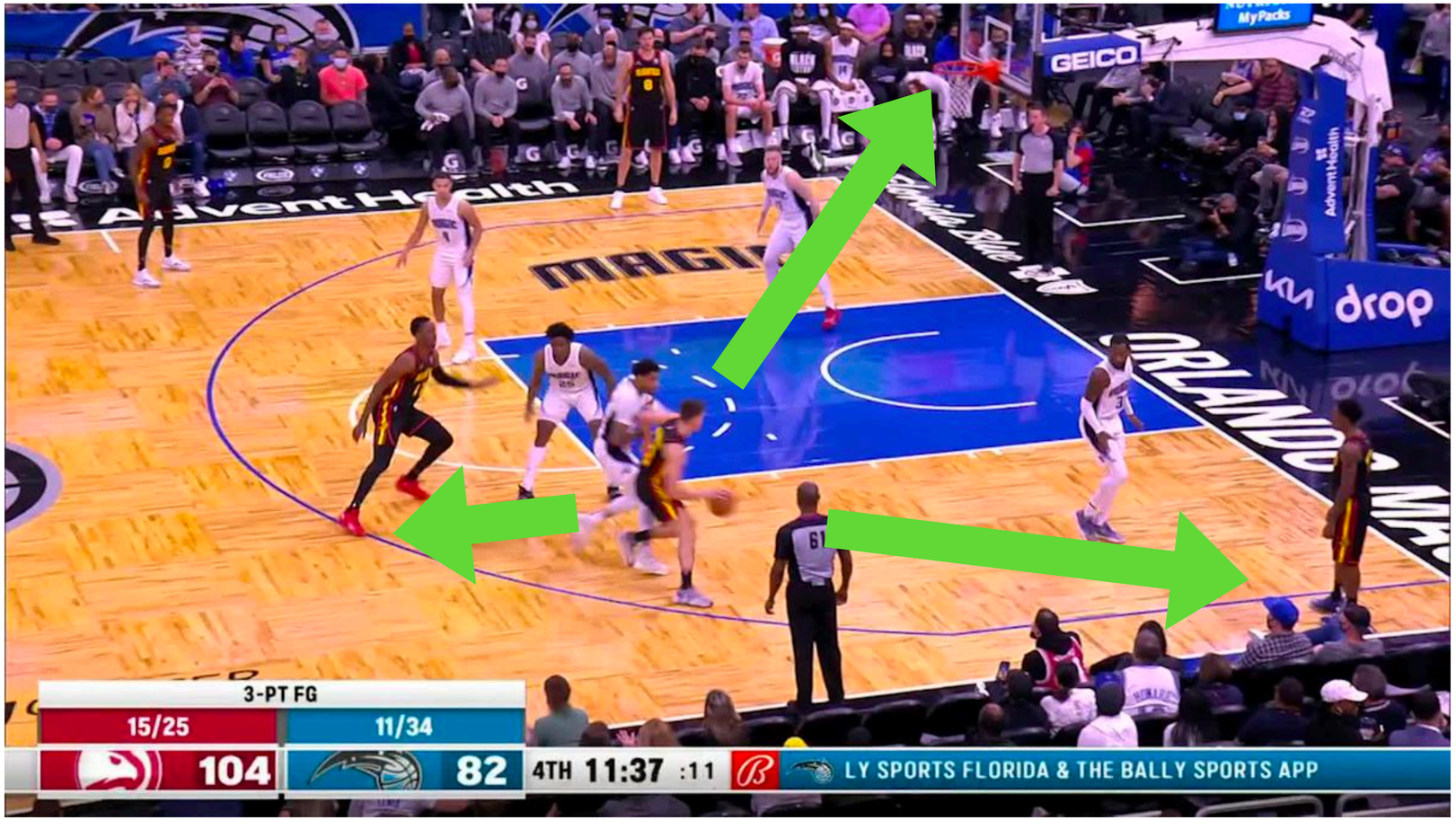
Learn a discriminator

Update a policy

Ho and Ermon. "Generative Adversarial Imitation Learning", 2016

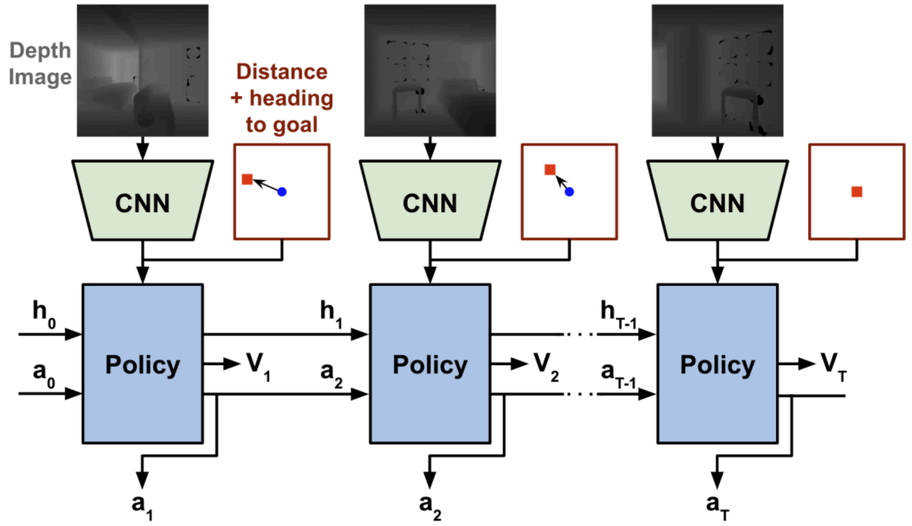
Non-Markovian Behavior

- What will the basketball player do? Pass? Shoot?



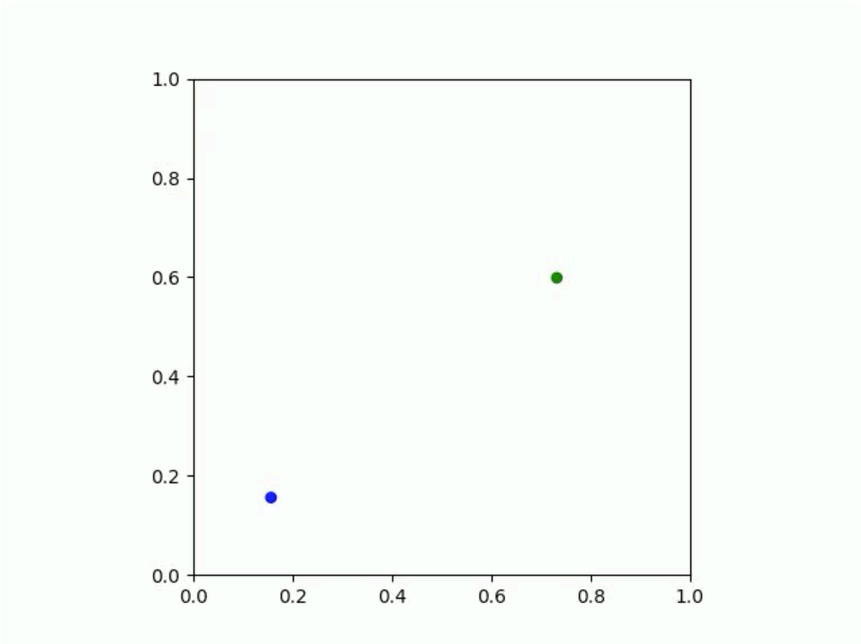
Non-Markovian Behavior

- The policy $\pi_\theta(a_t \mid s_t)$ only depends on the current state.
  - If we see the same thing twice, we do the same thing twice, regardless of what happened before.
- We can use the history: $\pi_\theta(a_t \mid s_{t-H}, \cdots, s_t)$
- Quiz: how can we use the entire history?
- Ans: Recurrent Neural Network (e.g., LSTM)



Expressiveness

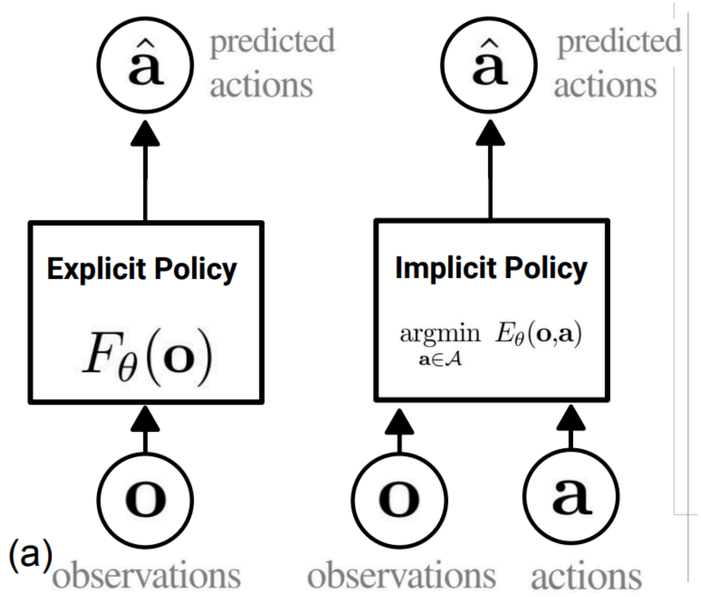
- Some behaviors are naturally multi-modal, discrete.
- They may be hard to be learned by behavior cloning.



Florence et al. “Implicit behavioral cloning”, 2022

Expressiveness

- Idea: learn a behavior as an implicit function. $\hat{a} = \text{argmin}\ E_\theta(o, a)$



Implicit

$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\arg\min}\ E_\theta(\mathbf{x}, \mathbf{y})$

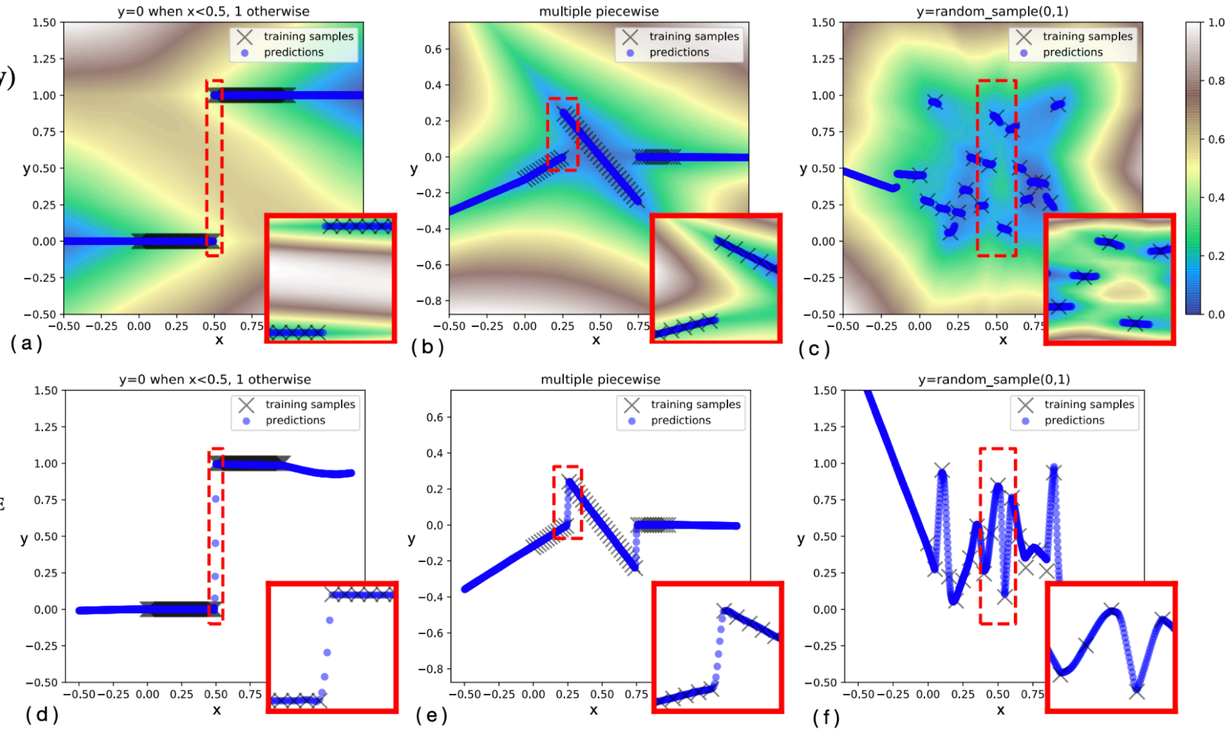
ReLU-MLP trained as EBM
2:512:512:1
5k steps

shown density is:
$\text{normalized}(E_\theta(\mathbf{x}, \mathbf{y}))$
$\mathbf{y}$

Explicit

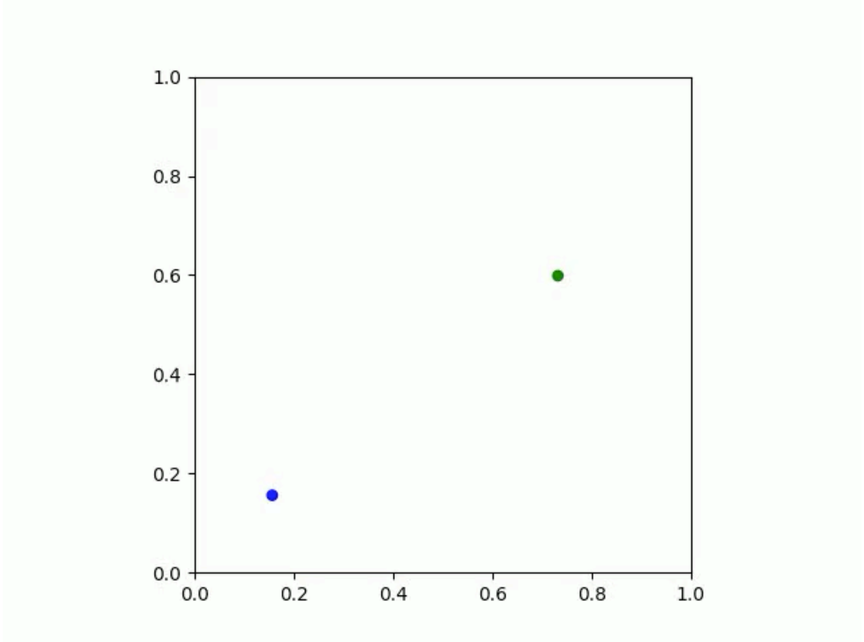
$\hat{\mathbf{y}} = f_\theta(\mathbf{x})$

ReLU-MLP trained with MSE
1:512:512:1
20k steps

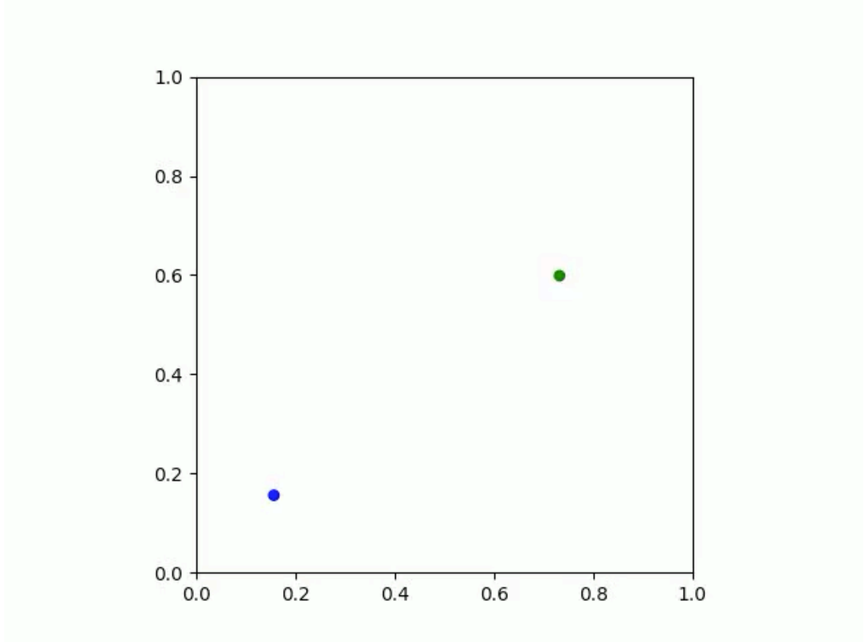


Expressiveness

- Better at learning complex behaviors.



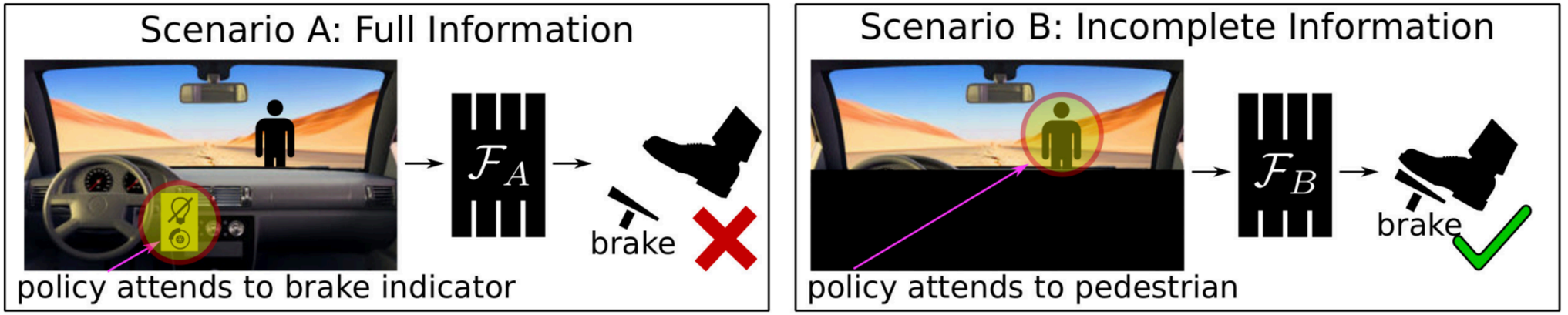
Implicit BC



Explicit BC

Causality

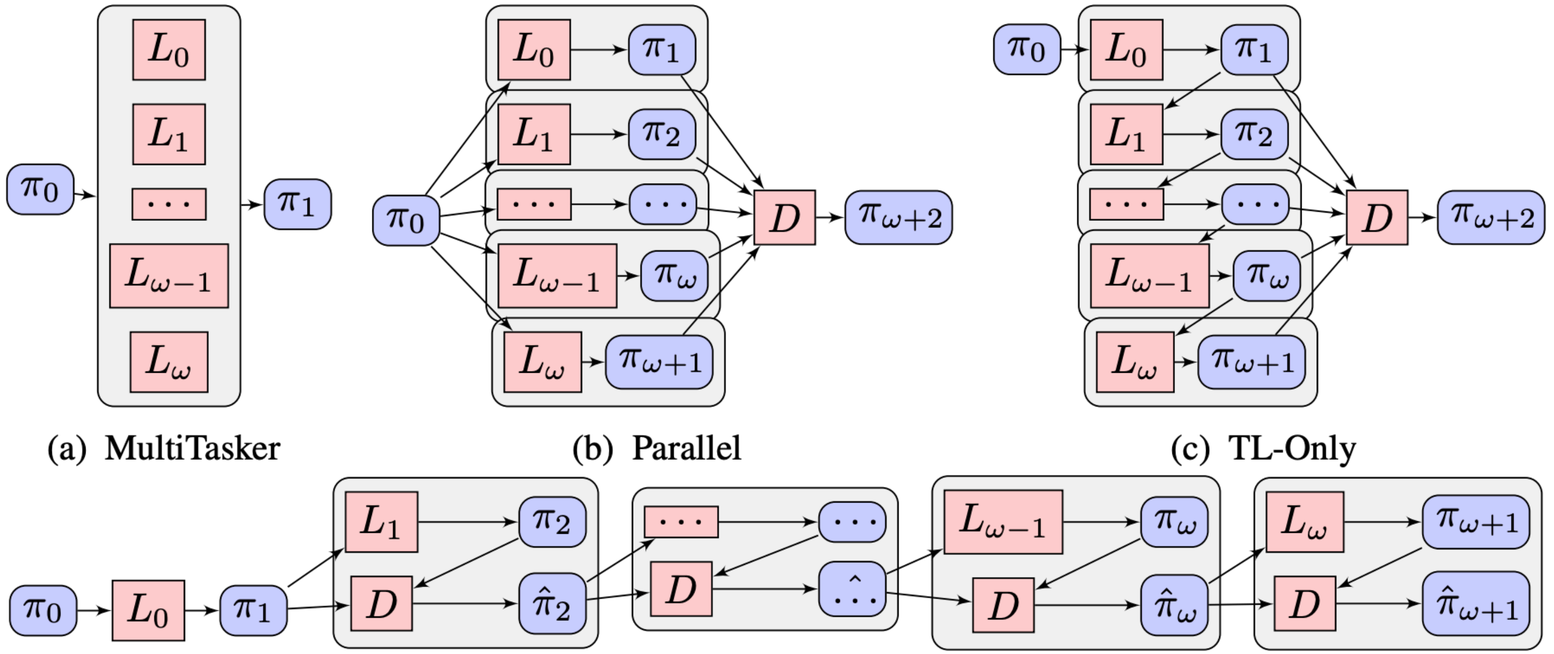
- Causal misidentification: more information yields worse imitation learning performance.
  - Model A relies on the braking indicator to decide whether to brake. Model B instead correctly attends to the pedestrian.



De Hann et al. “Causal Confusion in Imitation Learning”, 2019

Multi-task Learning

- Forgetting is a big issue for multi-task learning.
- We could combine “learning” and “distillation” (BC).



(a) MultiTasker (b) Parallel (c) TL-Only

Berseth et al. “PROGRESSIVE REINFORCEMENT LEARNING WITH DISTILLATION FOR MULTI-SKILLED MOTION CONTROL”, 2018

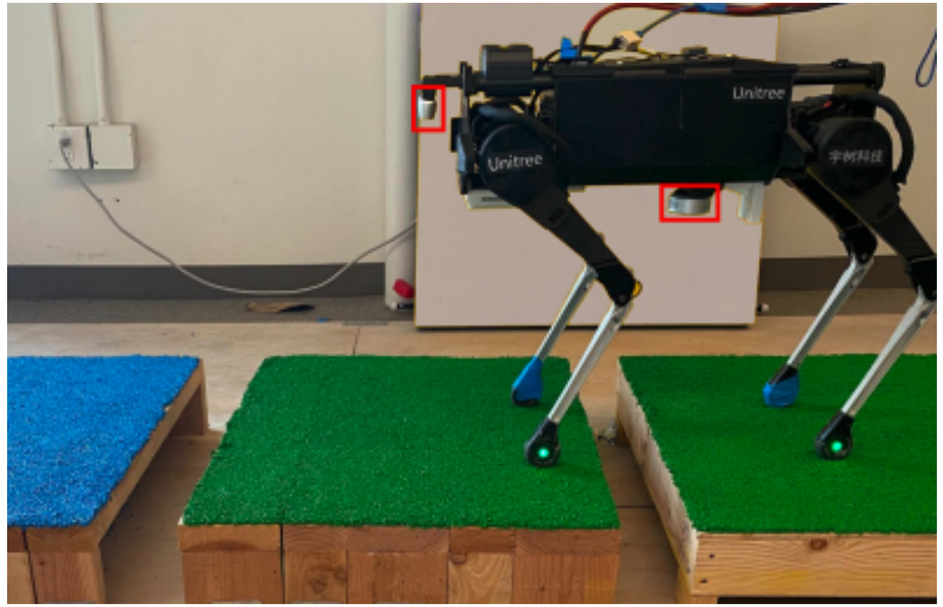
Brainstorming

- Can an RL agent do better than the teacher?
  - Ans: yes, if it combines with fine-tuning.

Ghosh et al. "Learning to reach goals via iterated supervised learning." *2019*

Brainstorming

- What about legged robots? Can we apply imitation learning for them?
  - If not, what are the challenges?
  - If yes, how?



Challenges

- Apply Vanilla BC to legged robots is not straightforward because it is hard to obtain expert demonstrations.
  - Data collection for manipulators: remote controller, VR, …
  - Data collection for autonomous driving: real driving experience…
  - Legged robots requires high dimensional, time-sensitive control.

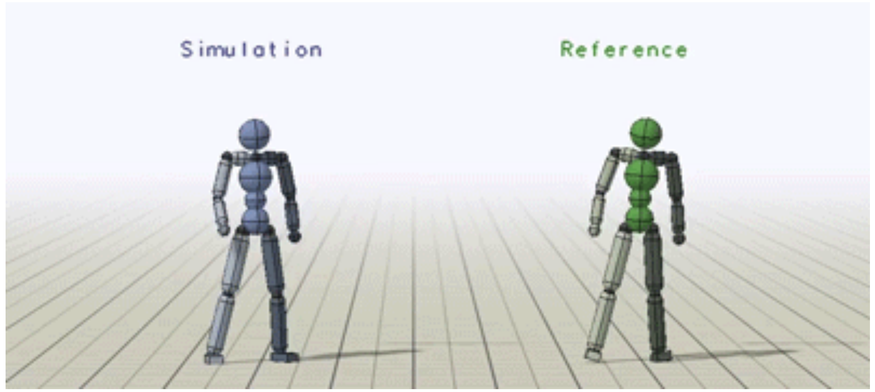
Motion Imitation

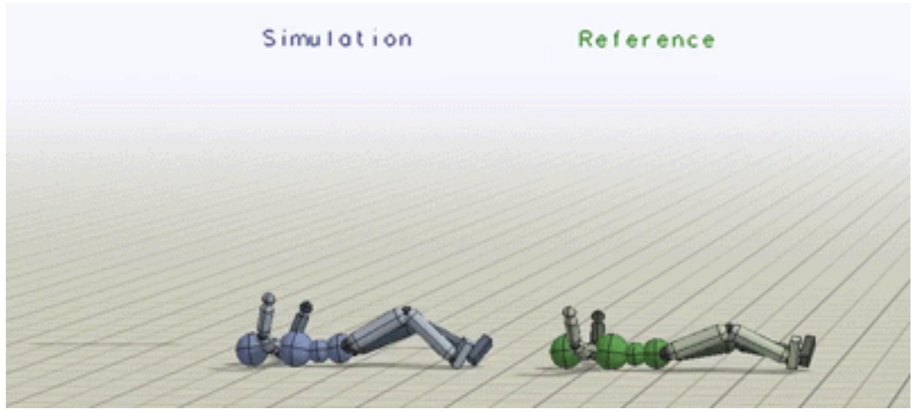
- Instead, we can try to imitate the motion trajectory of the human.
  - $r_t = \exp\left(-|q^t - \hat{q}^t|^2\right)$

Current pose

Reference pose

  - We can optimize the policy with PPO.



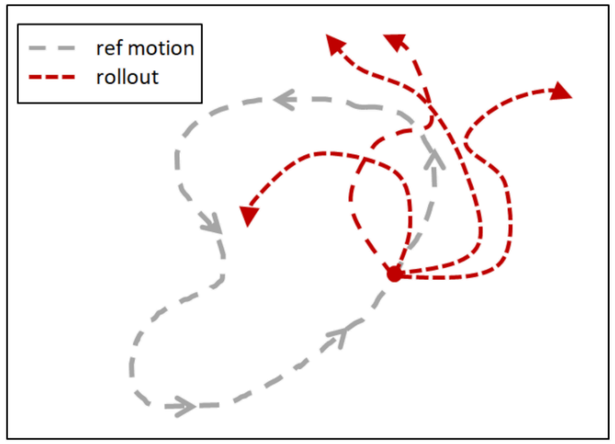


Peng et al. “Deepmimic: Example-guided deep reinforcement learning of physics-based character skills“

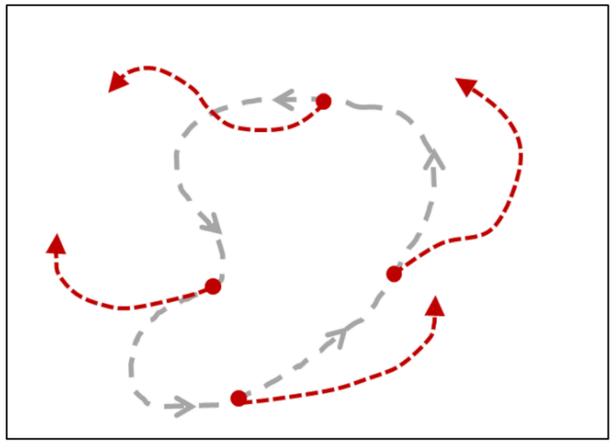
Motion Imitation

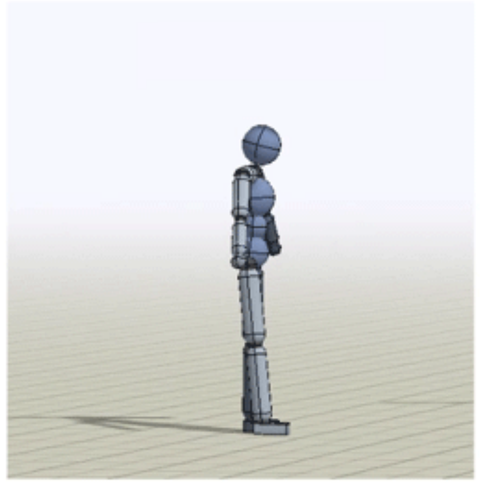
- Reference State Initialization
  - It is unlikely to accidentally execute a successful trajectory through random exploration.
  - Initialize by sampling the trajectory (reference state initialization).

Fixed Initial State

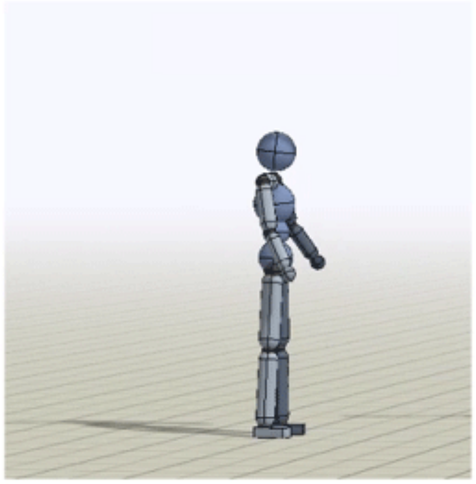


Reference State Initialization





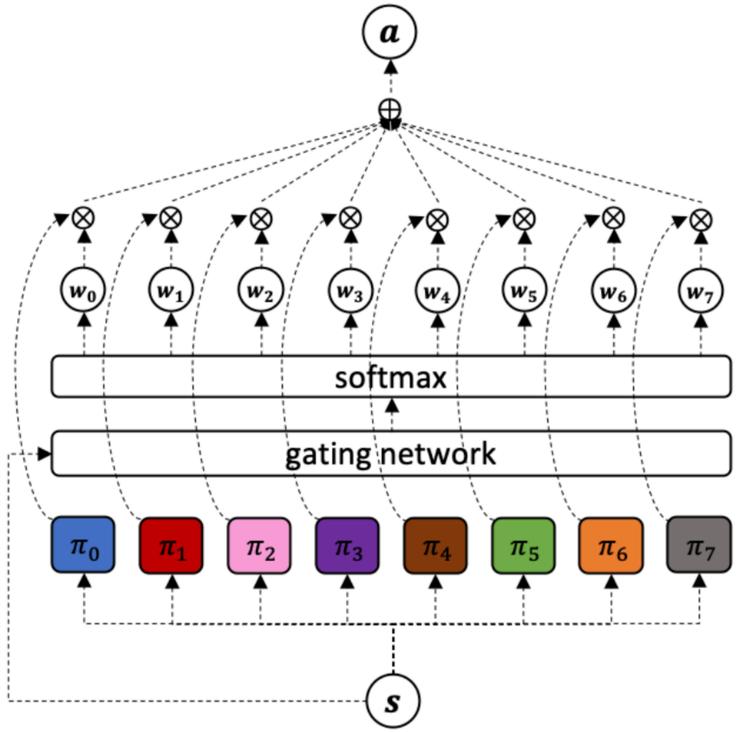
RSI



No RSI

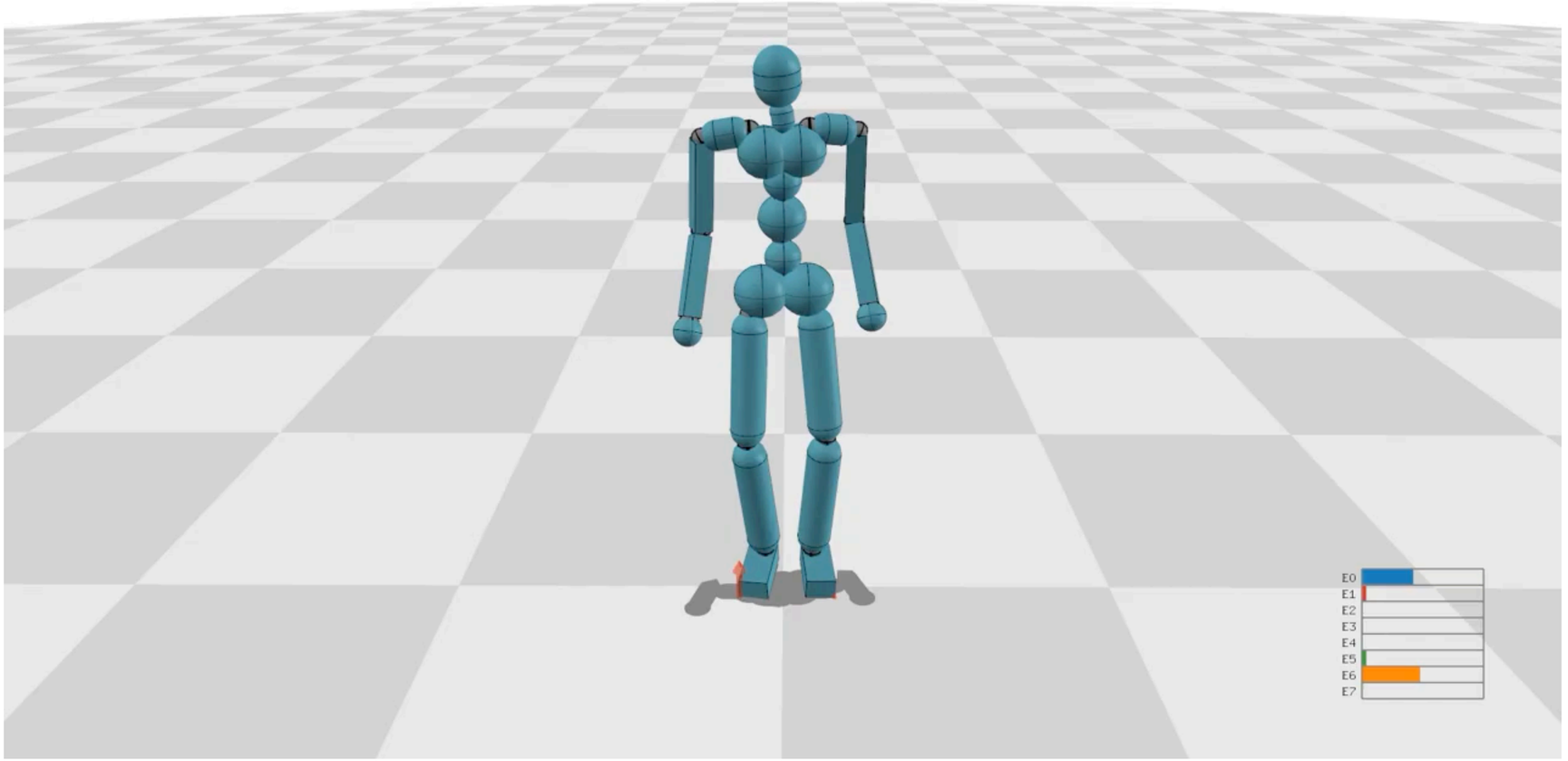
Scalable Motion Imitation

- What is we want to learn a larger motion data set?
- We can learn a mixture of expert for more efficient learning.



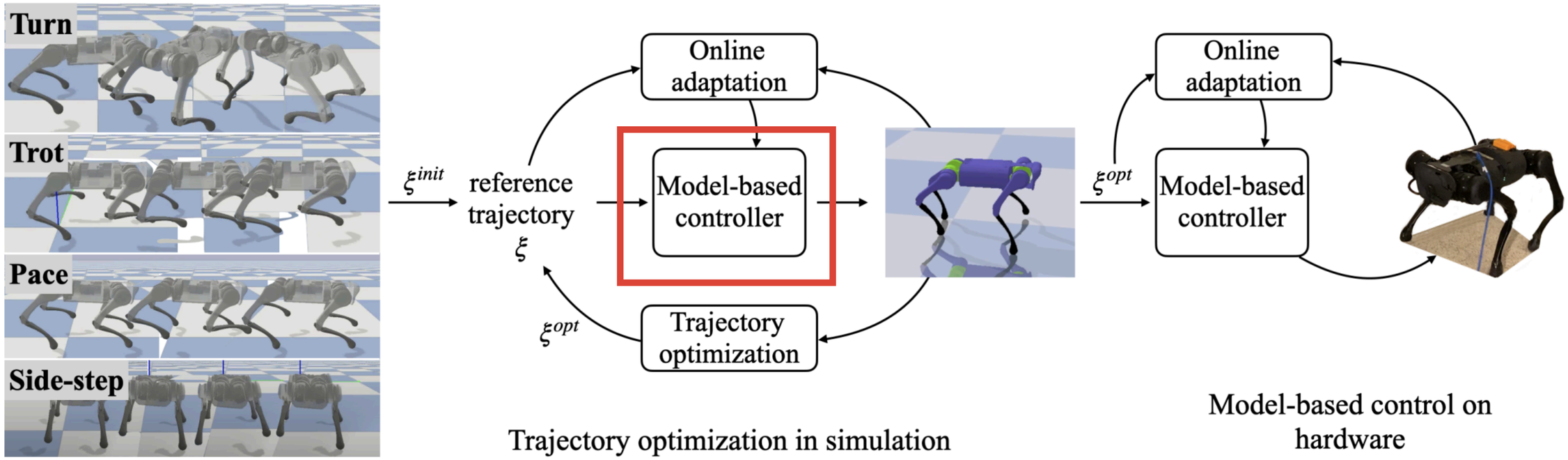
Won et al. "A Scalable Approach to Control Diverse Behaviors for Physically Simulated Characters", 2020

Scalable Motion Imitation



Motion Imitation as Control

Can we directly develop a model-based controller that can imitate the given motion?



Li et al. “Model-based Motion Imitation for Agile, Diverse and Generalizable Quadupedal Locomotion“, 2022

Motion Imitation as Control

Rule 1. Compute contact forces to track the given COM motion.

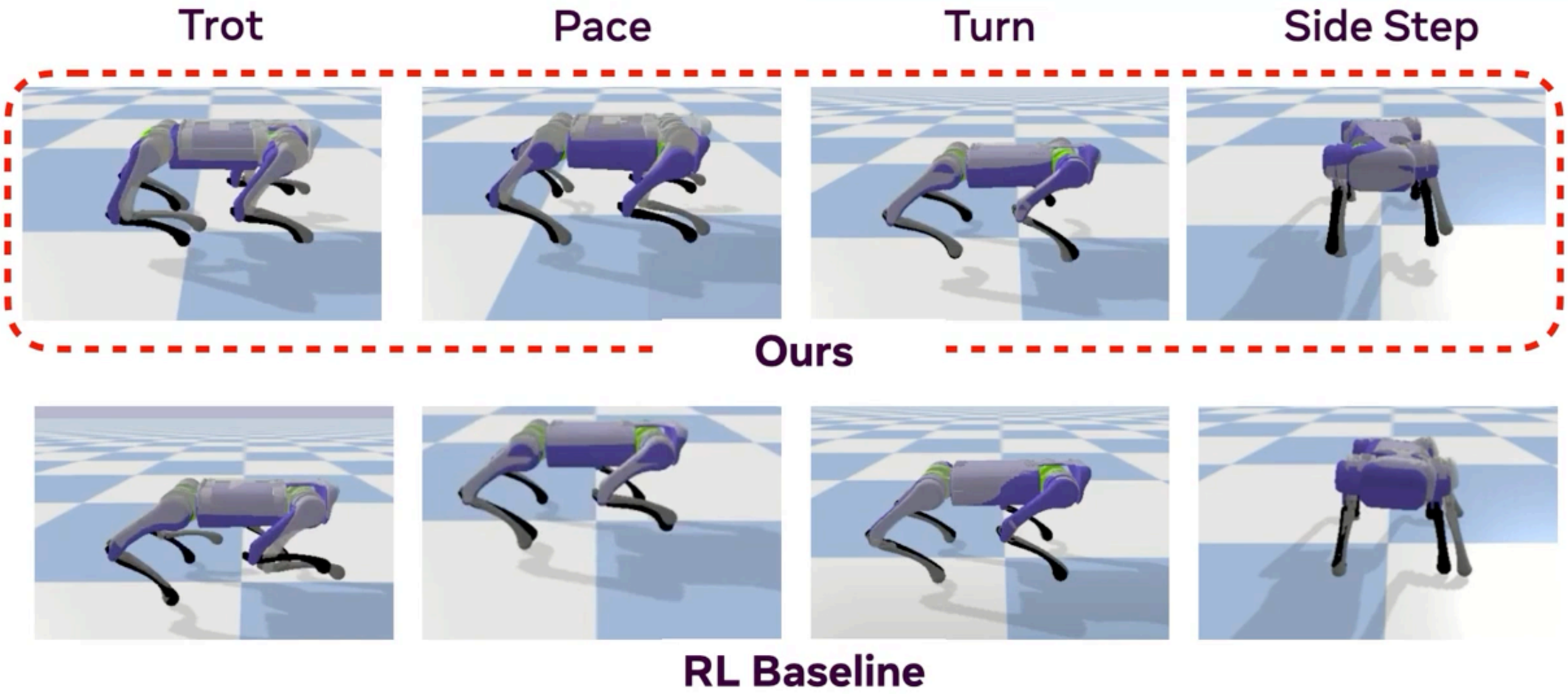
$$\begin{aligned} \bar{\mathbf{f}} = \arg\min_{\mathbf{f}} \ & \|\bar{\mathbf{M}}\mathbf{f} - \widetilde{g} - \ddot{\bar{\mathbf{x}}}\|_Q + \|\mathbf{f}\|_R \\ \text{s.t. } & f_{z,i} \geqslant f_{z,min}, \ \text{if Stance}, \quad f_{z,i} = 0, \text{if Swing} \\ & \mu f_{x,i} \geqslant f_{z,i} \geqslant -\mu f_{x,i}, \quad \mu f_{y,i} \geqslant f_{z,i} \geqslant -\mu f_{y,i}, \end{aligned} \tag{1}$$

Rule 2. Adjust the feet positions if moves too fast or too slow.

$$\hat{\mathbf{x}}^f_{t+1,i} = \mathbf{x}^f_{t,i} - \dot{\bar{\mathbf{x}}}^{\text{robot}}_t \cdot dt. \tag{2}$$

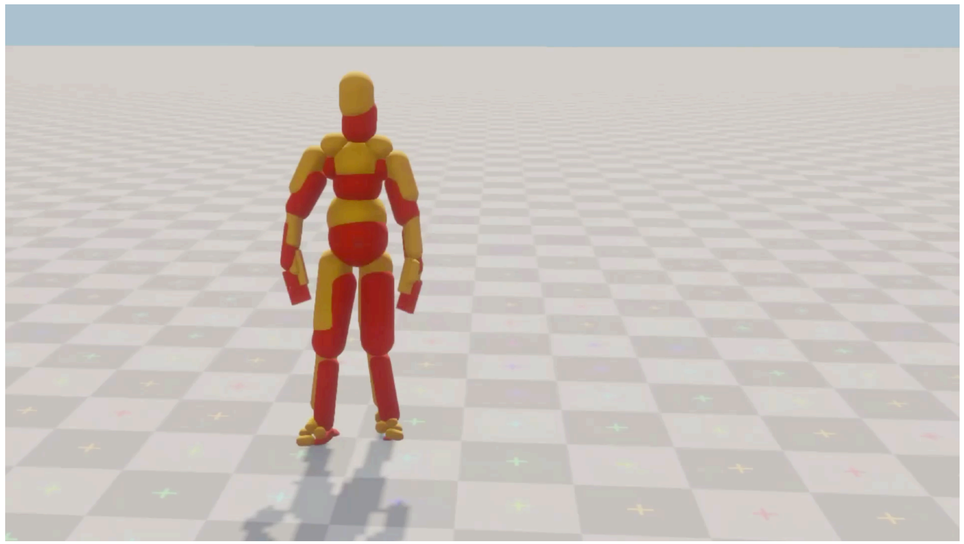
$$\mathbf{x}^{f,ref}_i = \bar{\mathbf{x}}^f_i - K(\dot{\bar{\mathbf{x}}}^{\text{robot}} - \dot{\mathbf{x}}^{\text{robot}}) \tag{3}$$

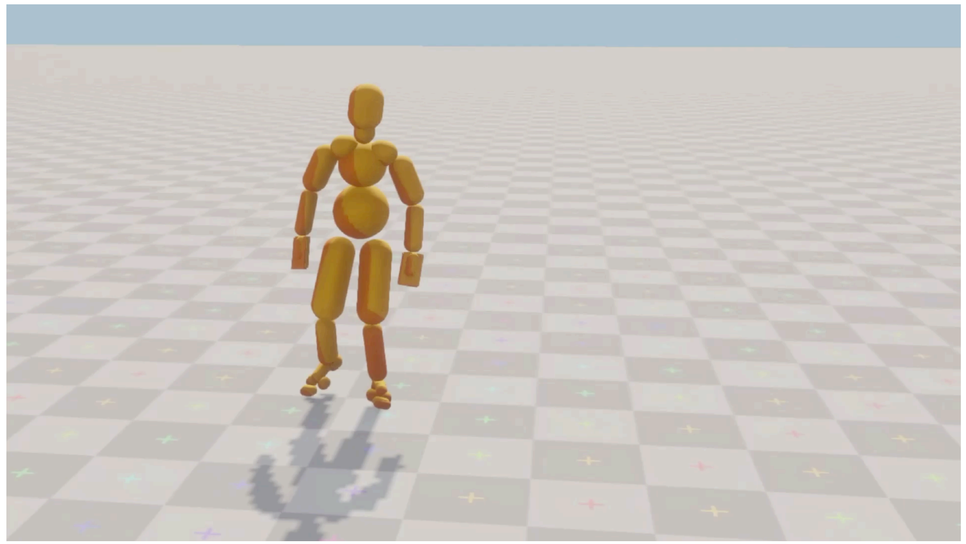
Motion Imitation as Control



Motion Imitation as Supervised Learning

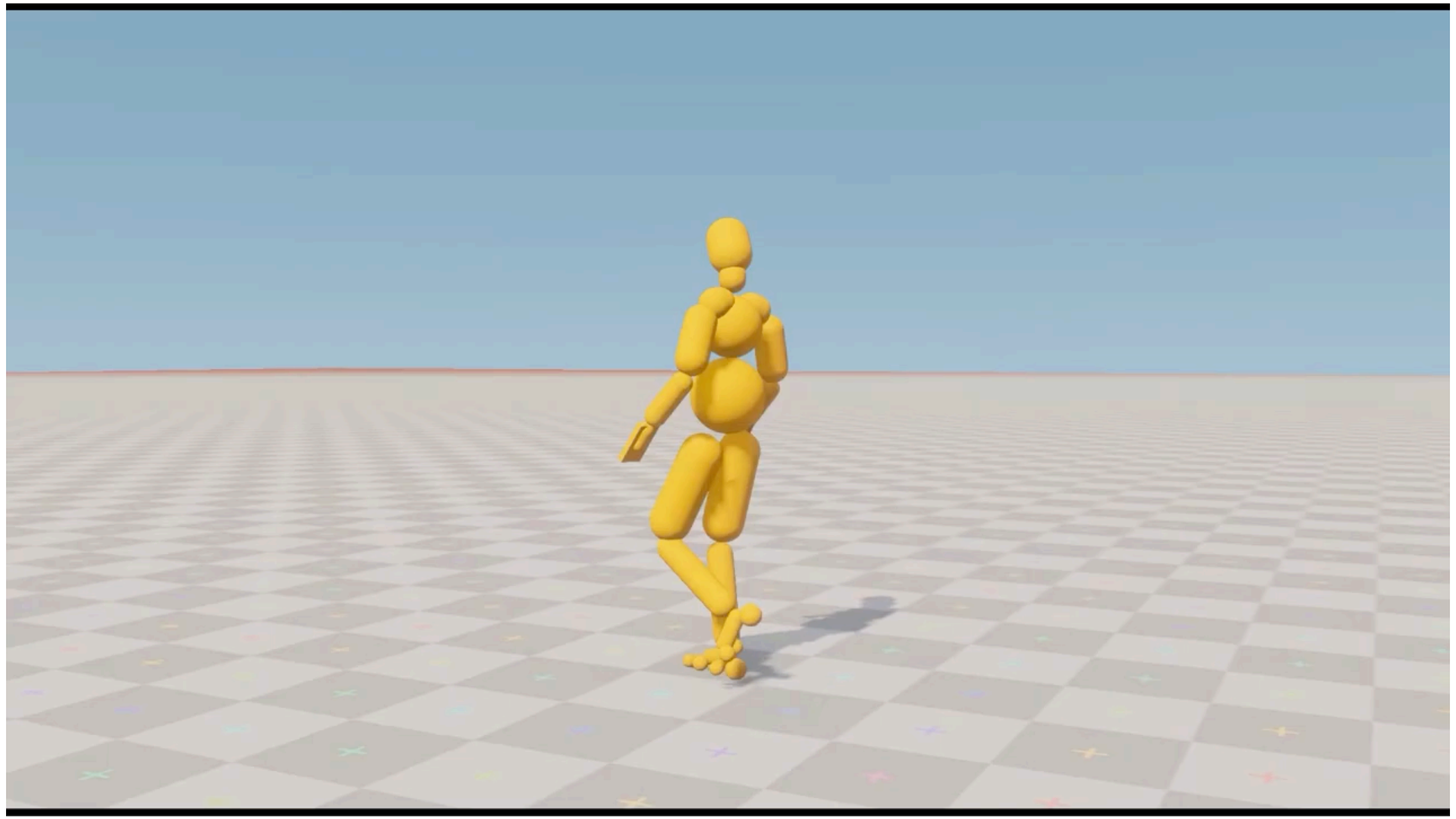
- You can learn a policy to match the pose after a short time horizon.
  - If your world model is differentiable, the problem can be SL.
  - You can iteratively learn the world model and train a policy.





Fussel et al. “SuperTrack: motion tracking for physically simulated characters using supervised learning“, 2021

Motion Imitation as Supervised Learning



Reading List

DAgger

- Paper #1: Ross, S., Gordon, G. and Bagnell, D., 2011, June. 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.

Trending RL + BC Algorithm

- Paper #2: Chen, D., Zhou, B., Koltun, V. and Krähenbühl, P., 2020, May. Learning by cheating. In Conference on Robot Learning (pp. 66-75). PMLR.

Adversarial Imitation

- Paper #3: Merel, J., Tassa, Y., TB, D., Srinivasan, S., Lemmon, J., Wang, Z., Wayne, G. and Heess, N., 2017. Learning human behaviors from motion capture by adversarial imitation. arXiv preprint arXiv:1707.02201.

Classic Motion Imitation

- Paper #4: Peng, X.B., Abbeel, P., Levine, S. and van de Panne, M., 2018. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. ACM Transactions on Graphics (TOG), 37(4), pp.1-14.

Reading List

Inverse Reinforcement Learning

- Runner-up: Abbeel, P. and Ng, A.Y., 2004, July. Apprenticeship learning via inverse reinforcement learning. In Proceedings of the twenty-first international conference on Machine learning (p. 1).

Multi-task Learning with BC

- Runner-up: Berseth, G., Xie, C., Cernek, P. and Van de Panne, M., 2018. Progressive reinforcement learning with distillation for multi-skilled motion control. arXiv preprint arXiv:1802.04765.

Summary

- Imitation learning helps us to learn a robust and natural policy from the expert demonstrations.
- It requires careful coordination about the details, like data management, policy representation, learning process, and so on.