ETH zürich

# Reinforcement Learning from Demonstrations

## Methods and Applications in Surgical Digital Twin Simulations

**Doctoral Examination, 24.06.2024**


Candidate: Ivan Ovinnikov

Supervisor: Prof. Dr. Joachim M. Buhmann
Co-examiner: Prof. Dr. Andreas Krause
Co-examiner: Dr. Raimundo Sierra

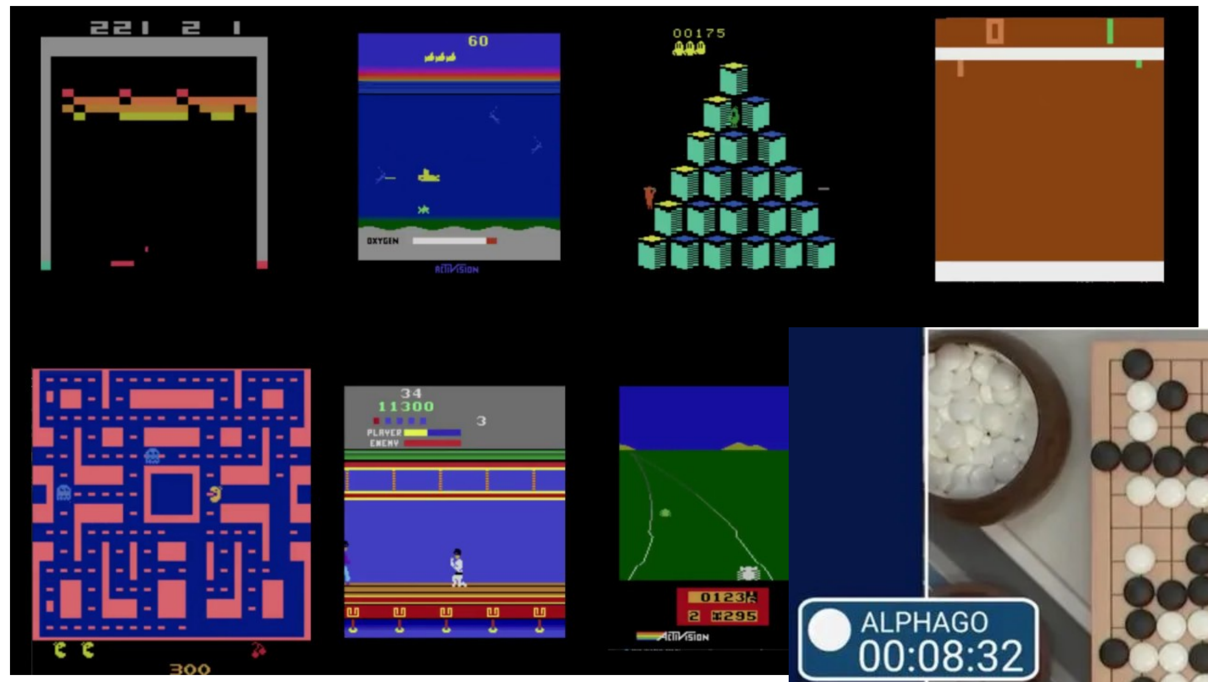
Chair: Prof. Dr. Markus Gross

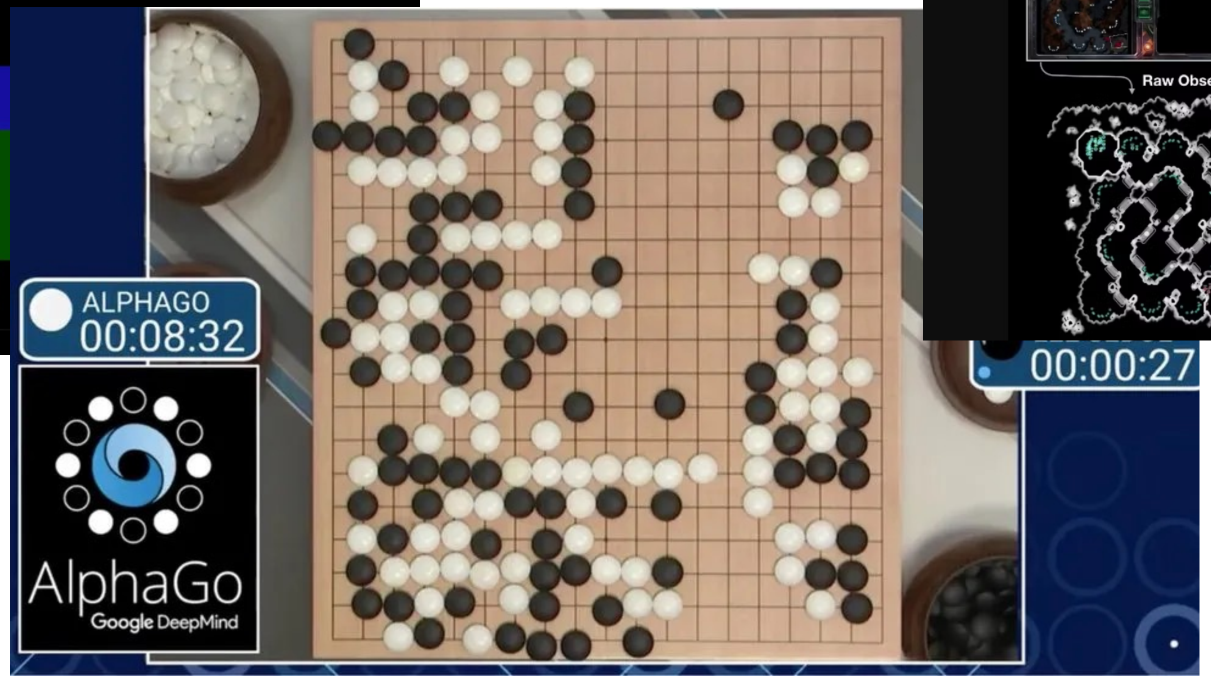
# Reinforcement Learning as Behaviour Formalism

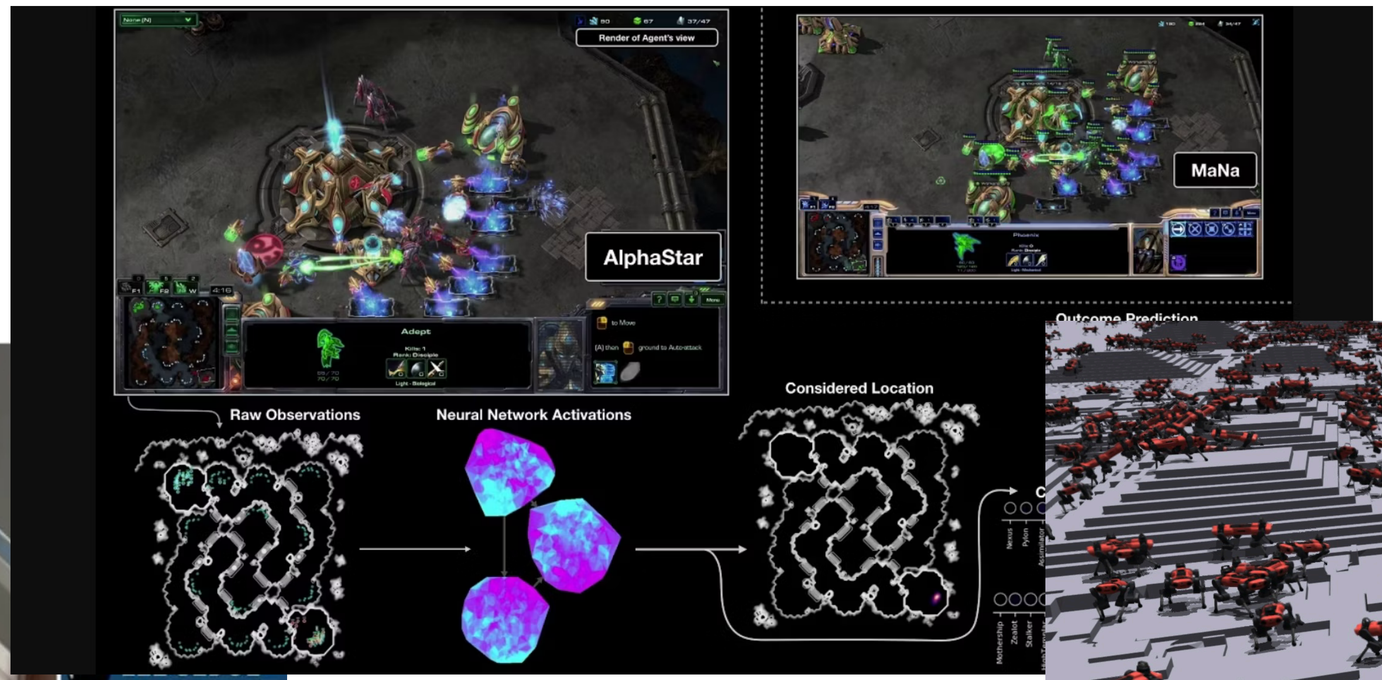
Action

Observation, Reward

**Main Objective: maximize gathered reward**





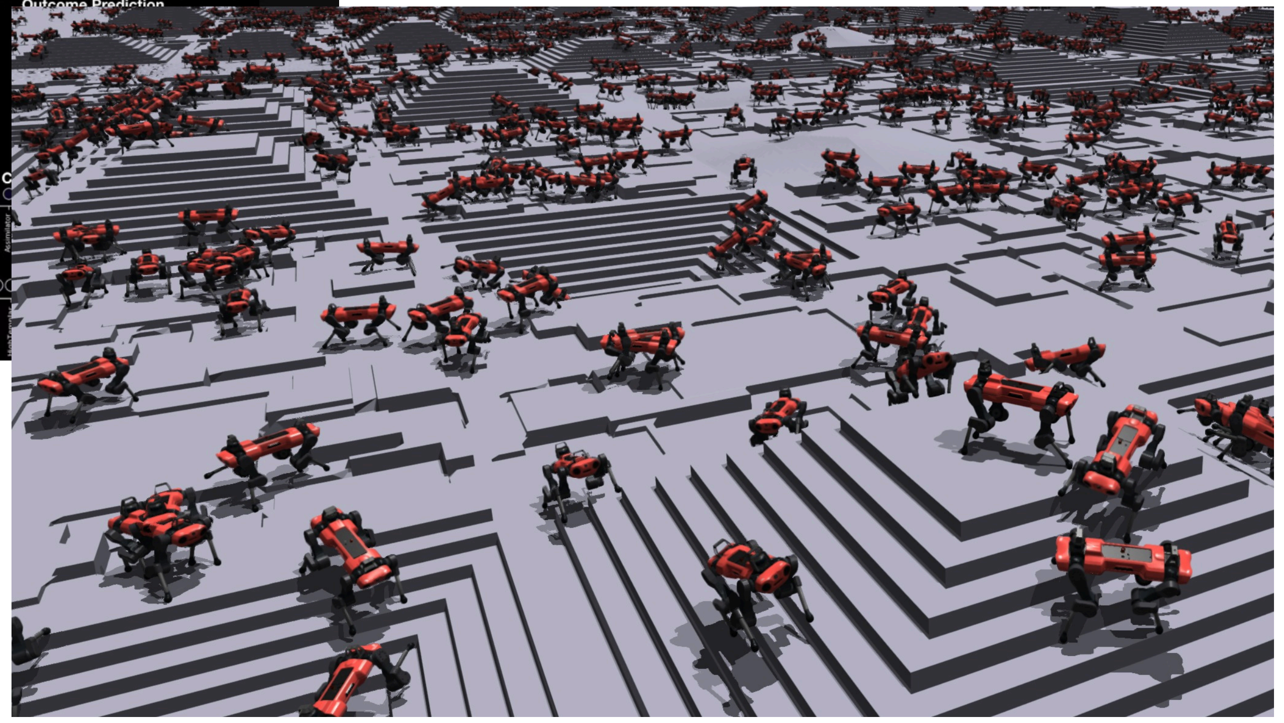



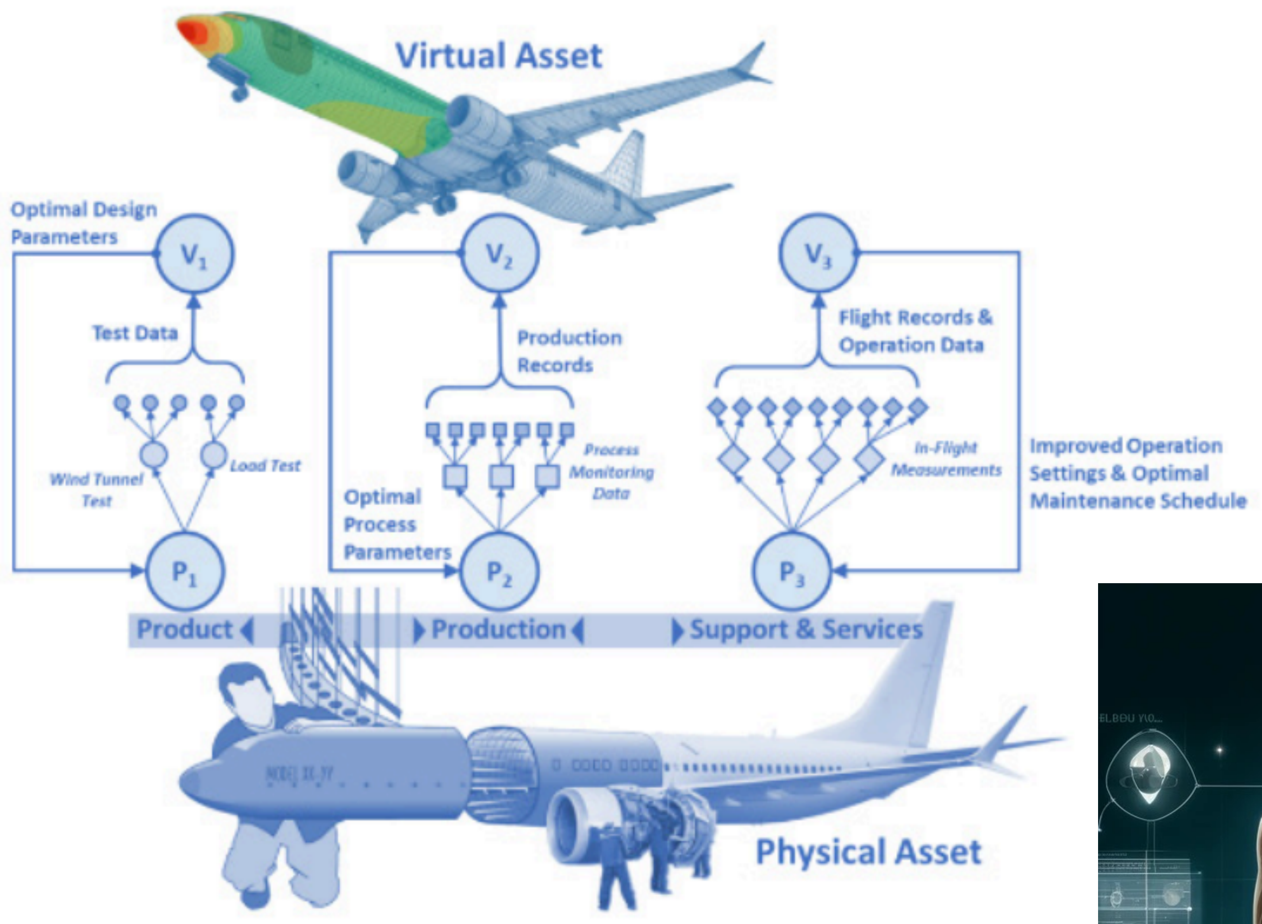
*Image sources: ChatGPT, Google Deepmind, NVIDIA IsaacGym*

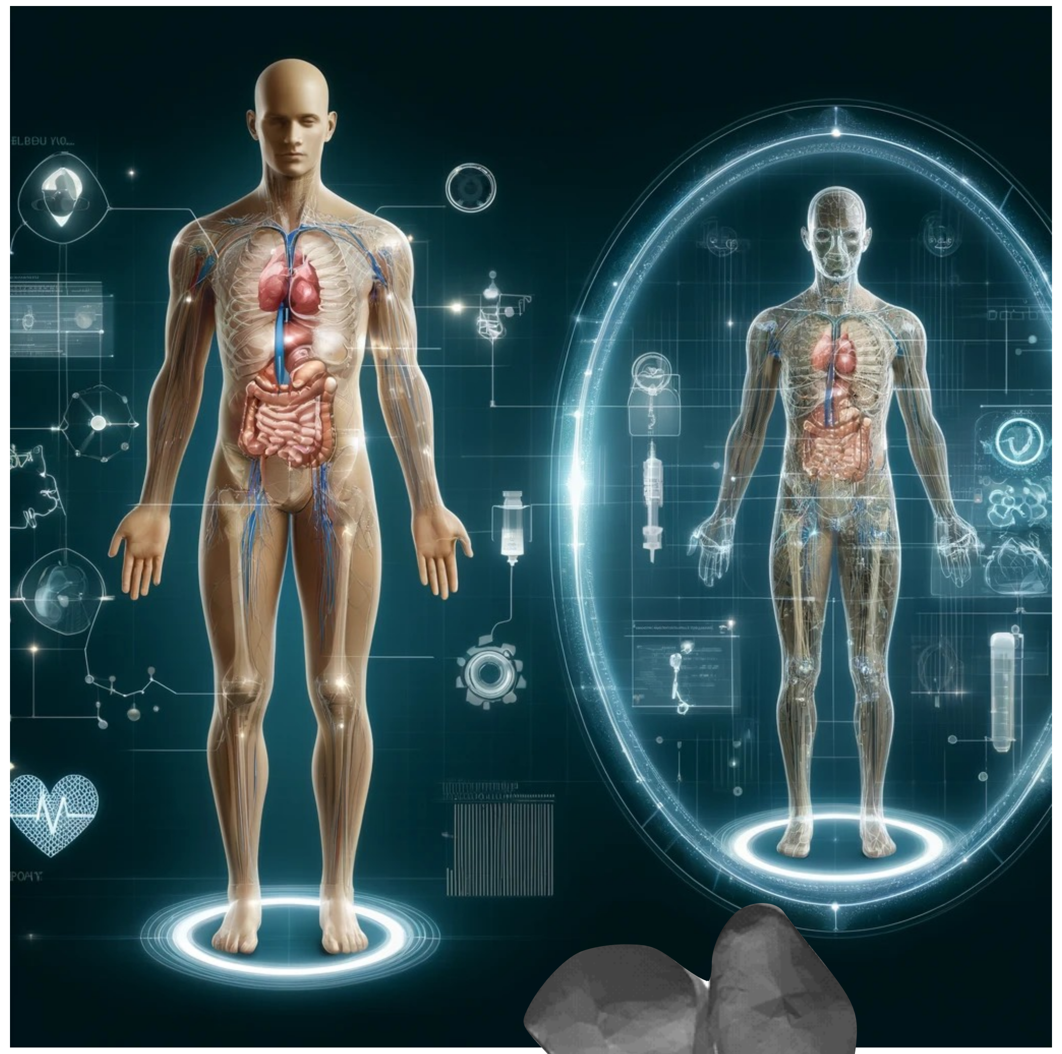
# Digital twins

## Definition & Examples

A set of virtual information constructs that mimics the structure, context and behaviour of an individual / unique physical asset, or a group of physical assets, is dynamically updated with data from its physical twin throughout its life cycle and informs decisions that realize value.

*AIAA Position Paper, 2020*



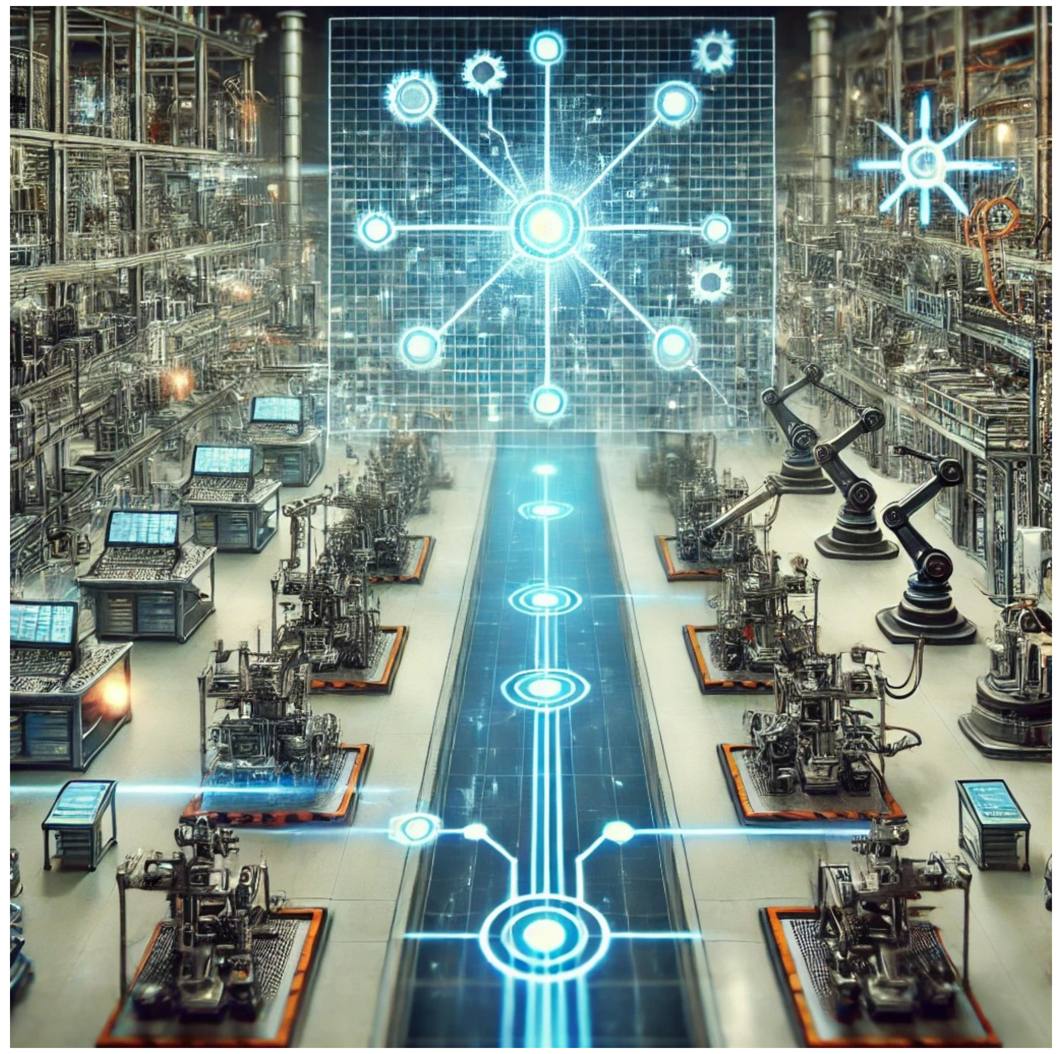
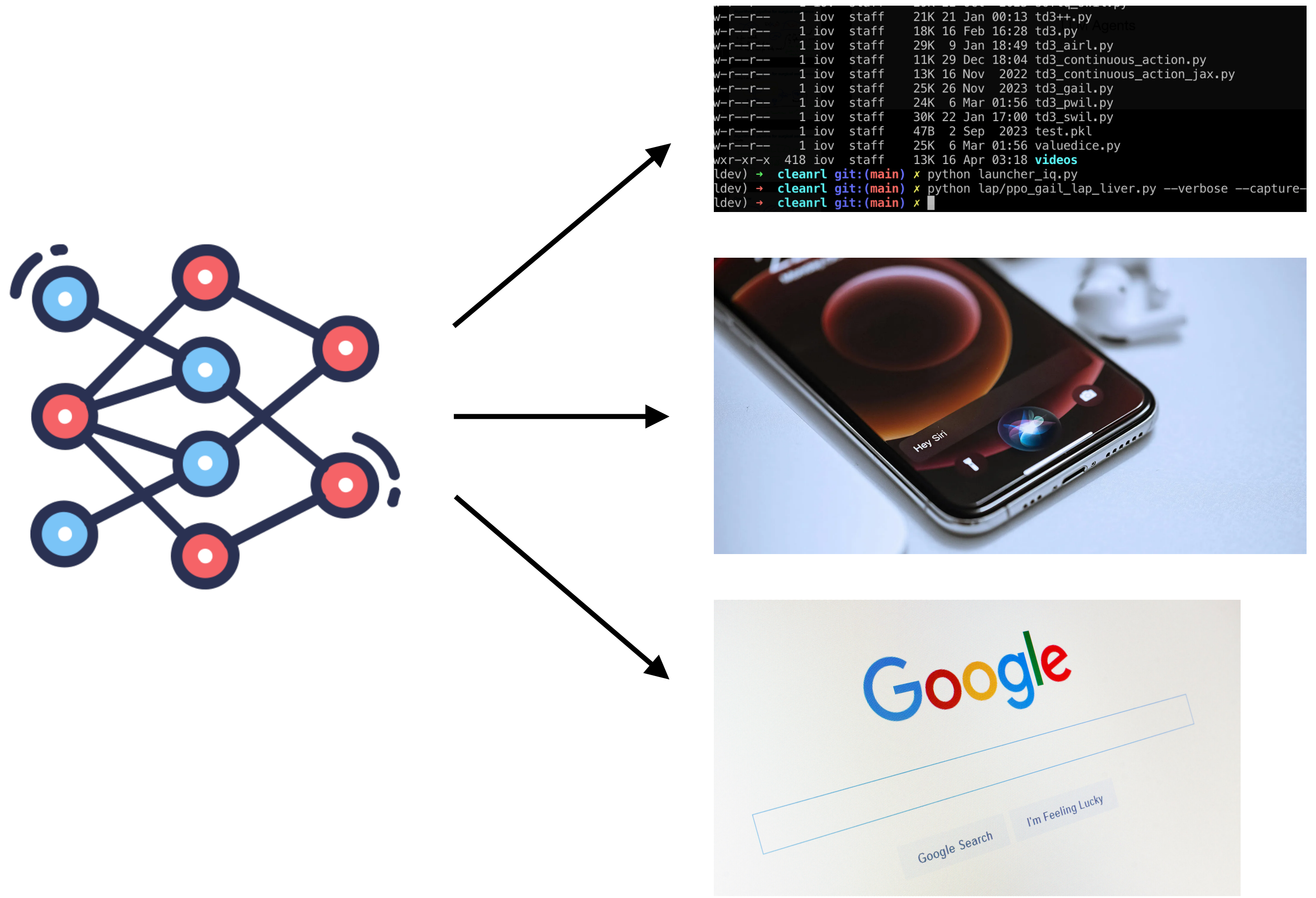


*Image sources: AIAA, ChatGPT, F. Laumer*

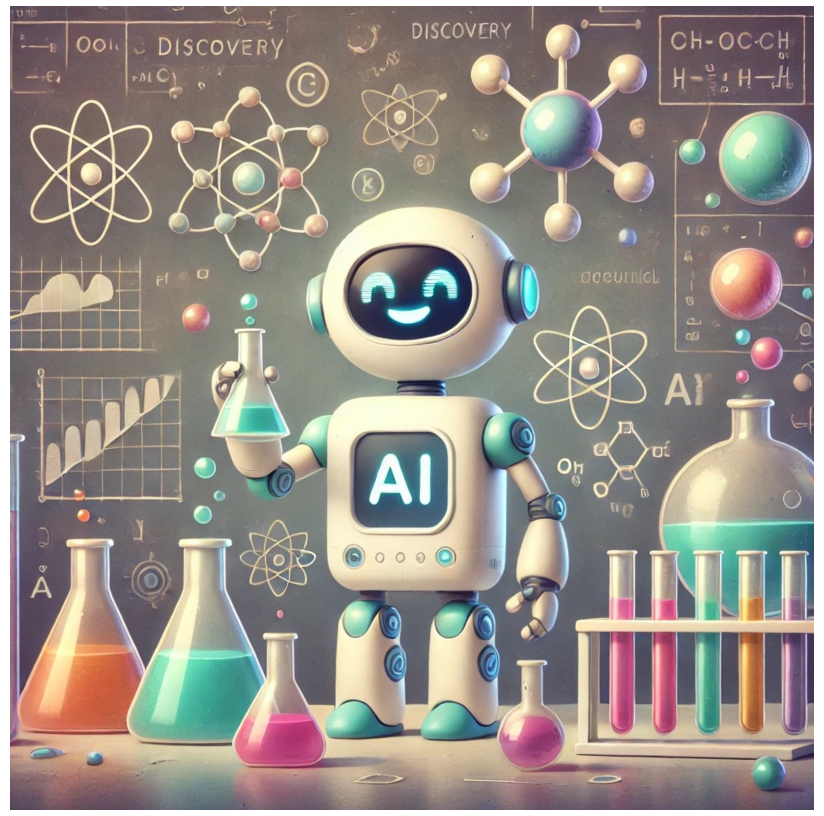
# RL in digital twins

## Environment models for agents

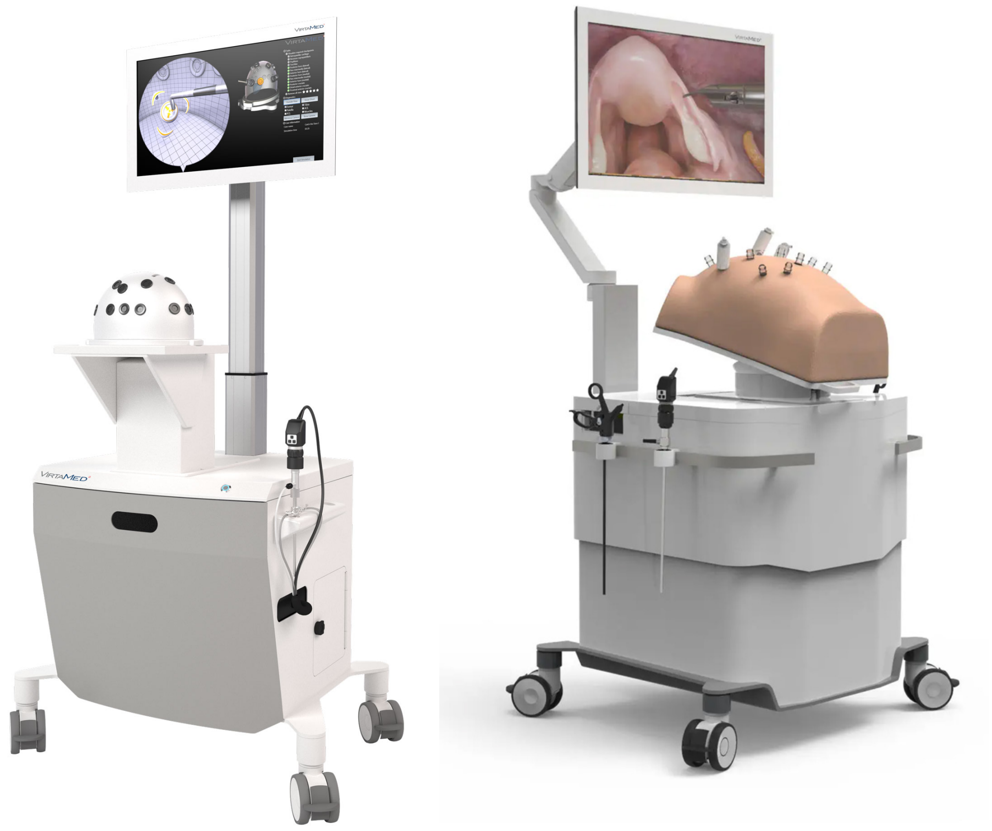
**Digital twins provide an interactive data source to learning models**




LLM Agents




Scientific Discovery

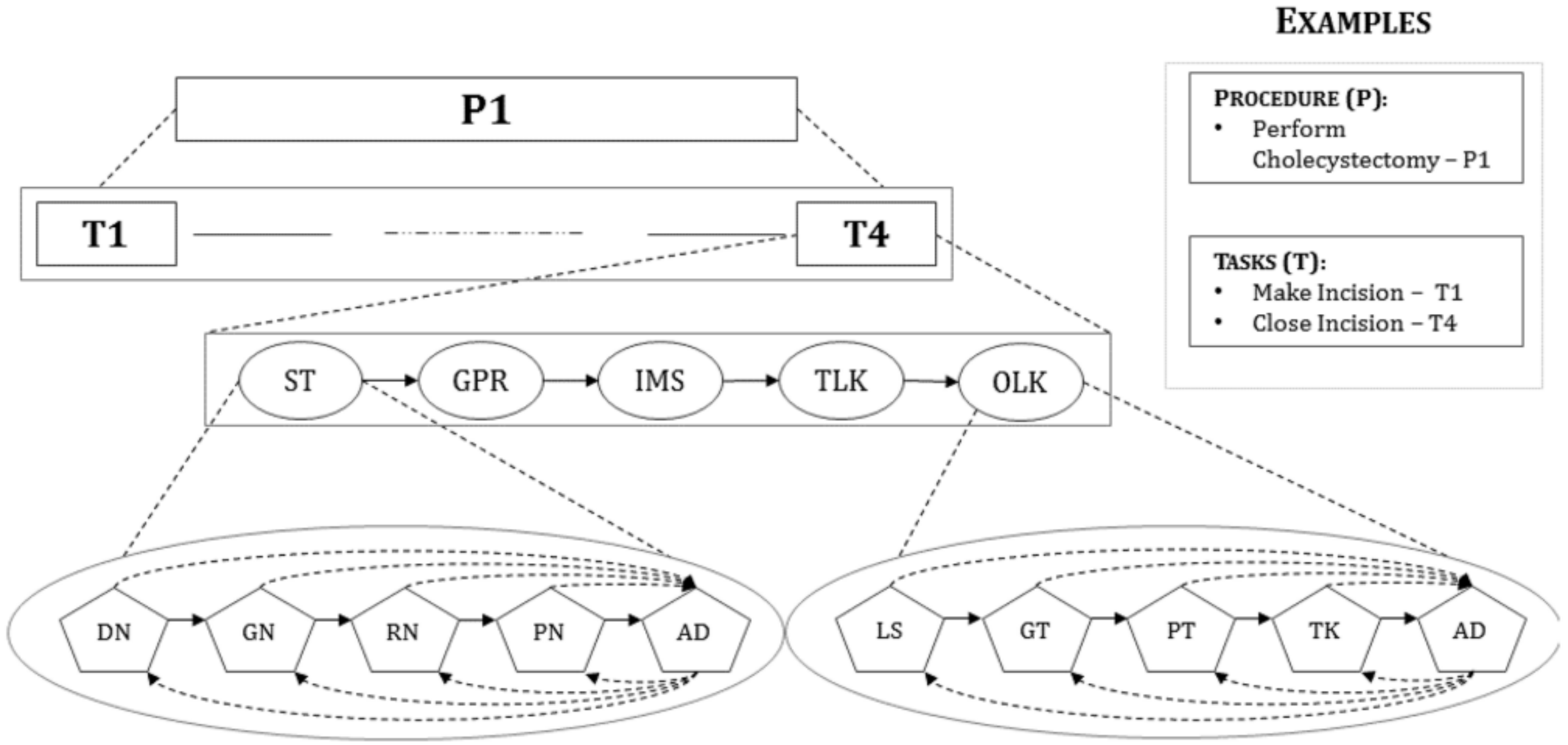


VIRTAMED
WE SIMULATE REALITY

Surgical Simulation

# RL in digital twins

## Eliciting behaviours in digital twins




| MANEUVERS: | |
|---|---|
| ST | Suture Throw |
| GPR | Pull/Run Suture out of Tissue |
| TLK[e] | Two Loop Knot |
| OLK[e] | One Loop Knot |
| IMS | Inter Maneuver Segment |

| GESTURES | | | |
|---|---|---|---|
| DN[e] | Drive Needle enter Skin | LS[e] | Loop_Suture_for_knot_tying |
| GN[e] | Grab Needle | GT[e] | Grasp_Suture_Tail_in_knot_tying |
| RN[e] | Rotate Needle through Skin | PT[e] | Pull_Suture_Tail_through_knot |
| PN[e] | Pull Needle Suture | TK[b] | Pull_Ends_of_Suture_taut |
| AD | Grab Suture; Grab Skin; Transfer Needle/Suture; Align Needle for Driving through Skin; Align Needle on Needle Driver; Camera Zoom In/Out; Camera Rotation | | |

**Fig 1. Hierarchical semantic decomposition of surgical activity.** [e] denotes that the segment can be performed using either of the robotic arms, [b] denotes that the segment is performed using both the robotic arms.

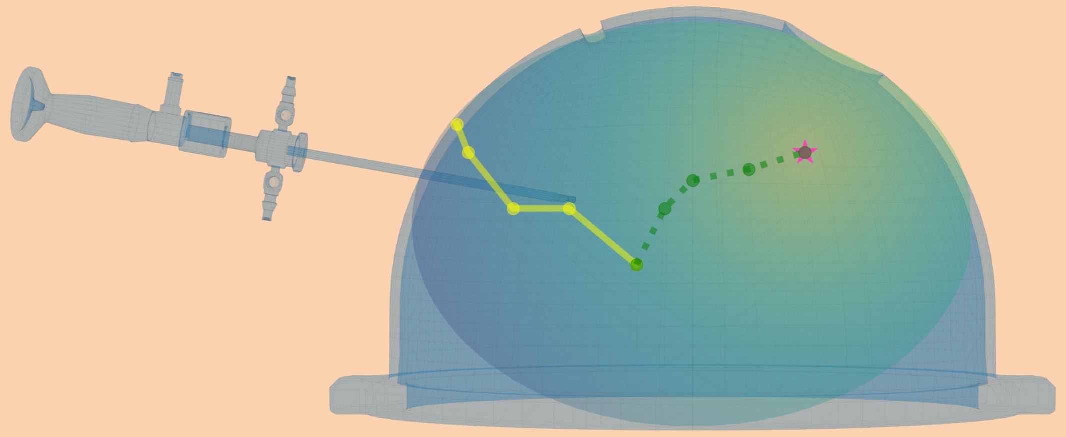
doi:10.1371/journal.pone.0149174.g001

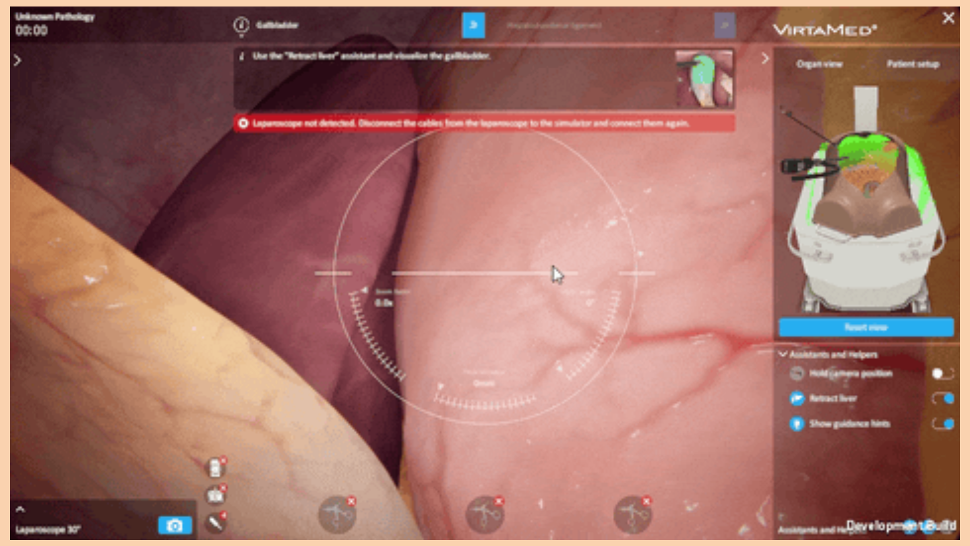
Complex environments require increasingly complex behaviours

Leverage demonstration data

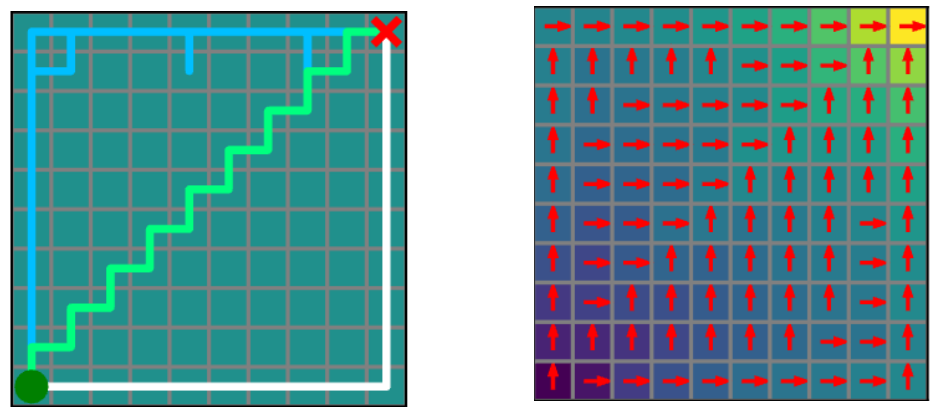
# Overview

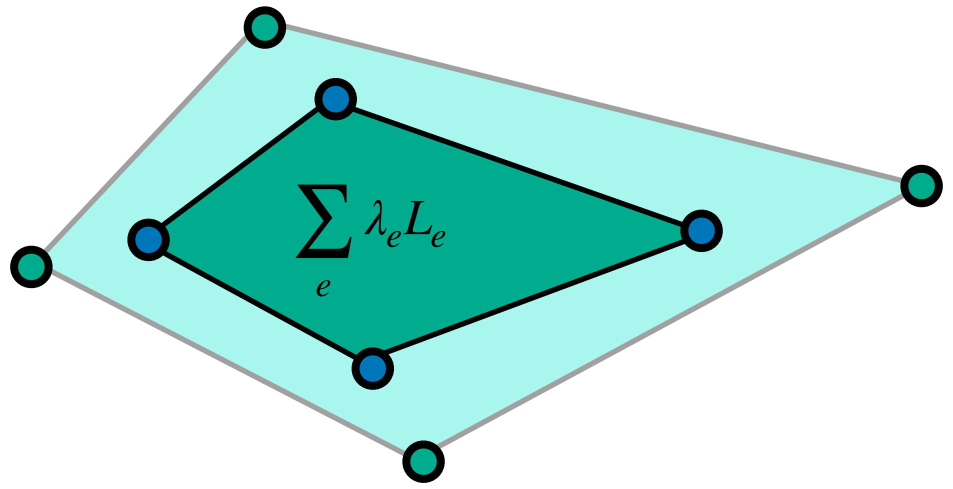
**Application:
Algorithmic RL pipeline in
Surgical Digital Twins**



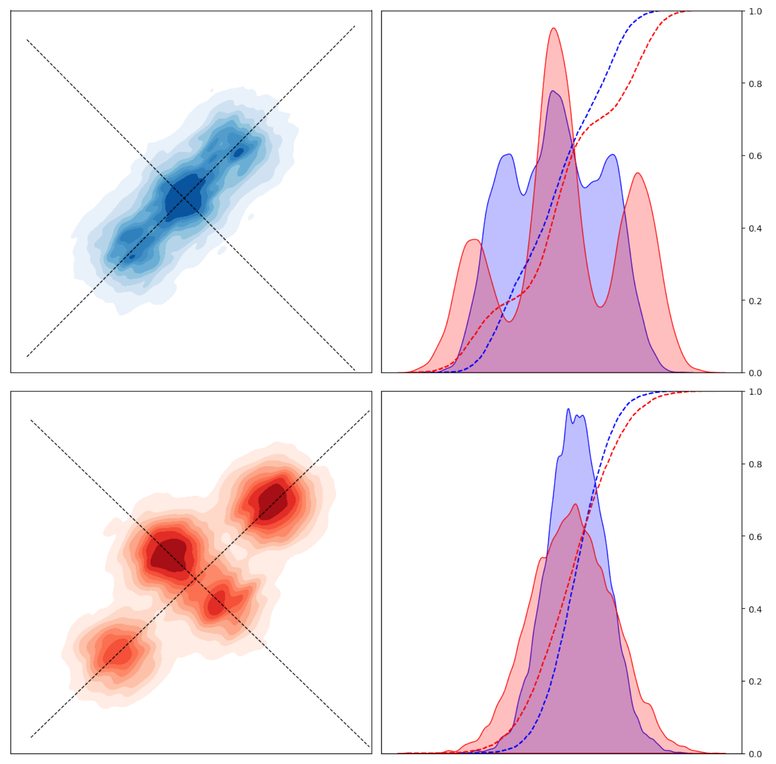


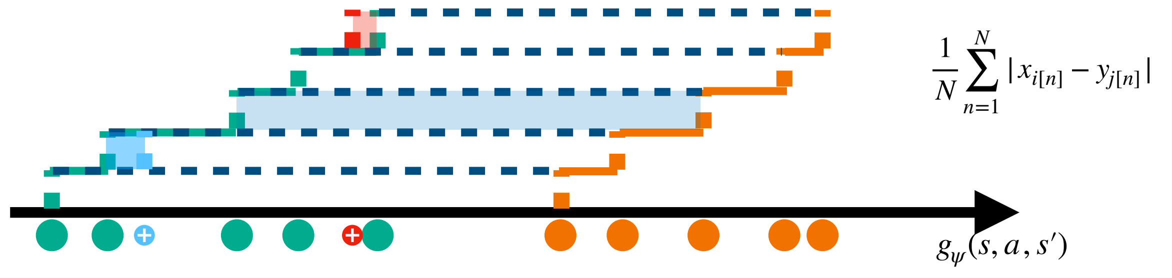
**Method I:
Addressing Reward
Generalization using Causal
Invariance**





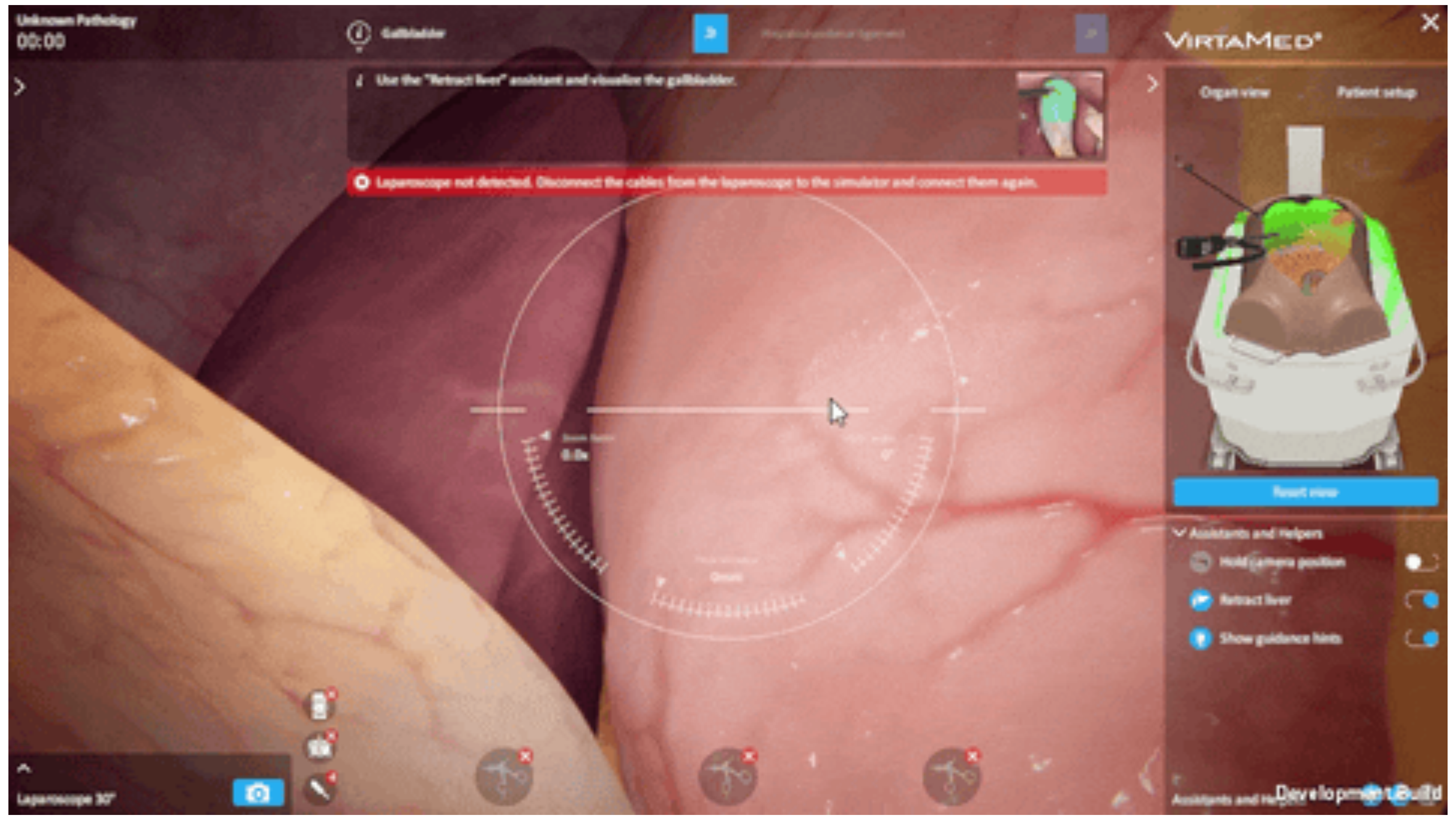

**Method II:
Addressing Data Efficiency in
Imitation Learning using Sliced
Optimal Transport**

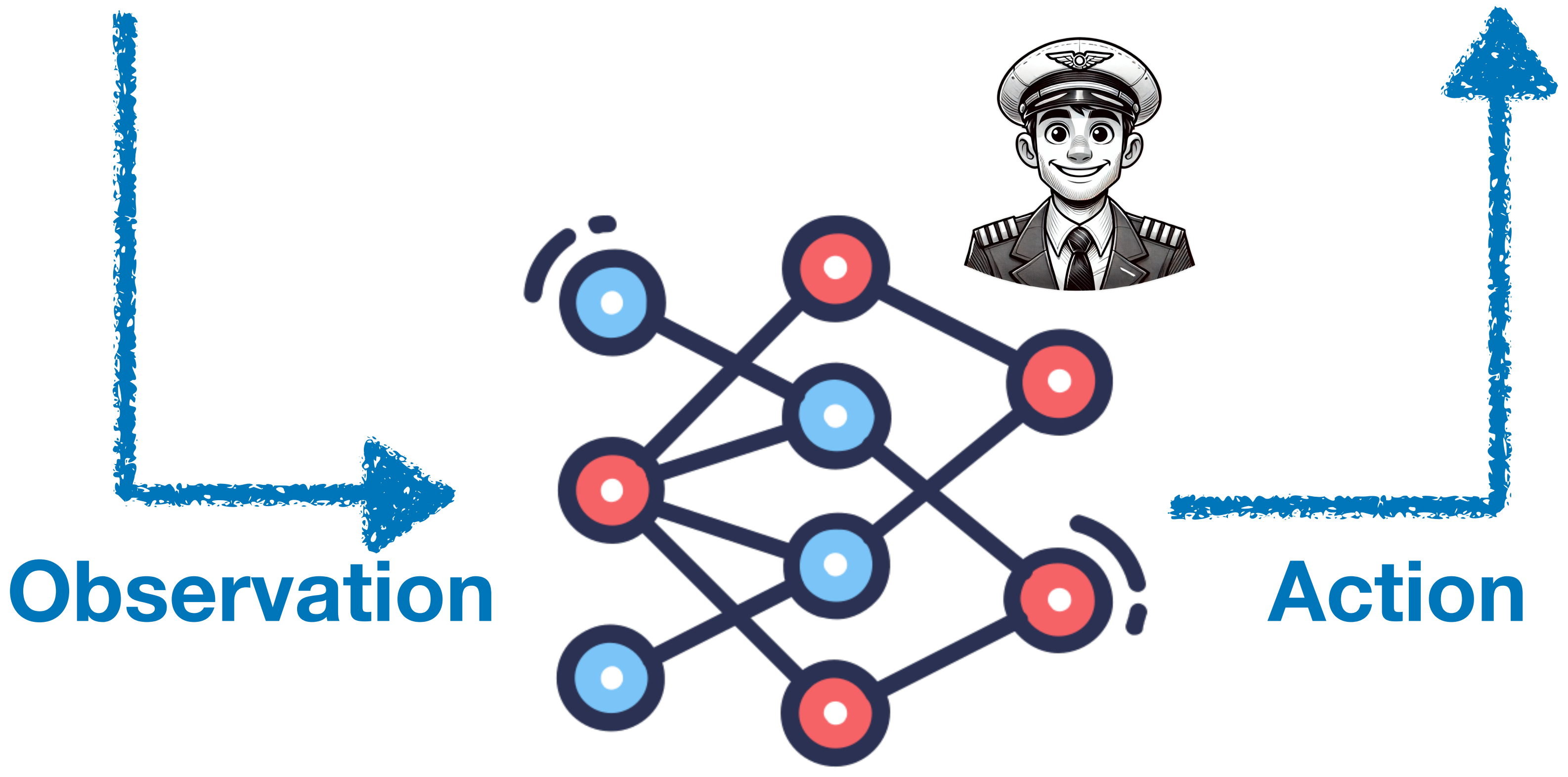




# RL in Surgical Digital Twins

## Research questions




Behavioural Policy (and Reward)

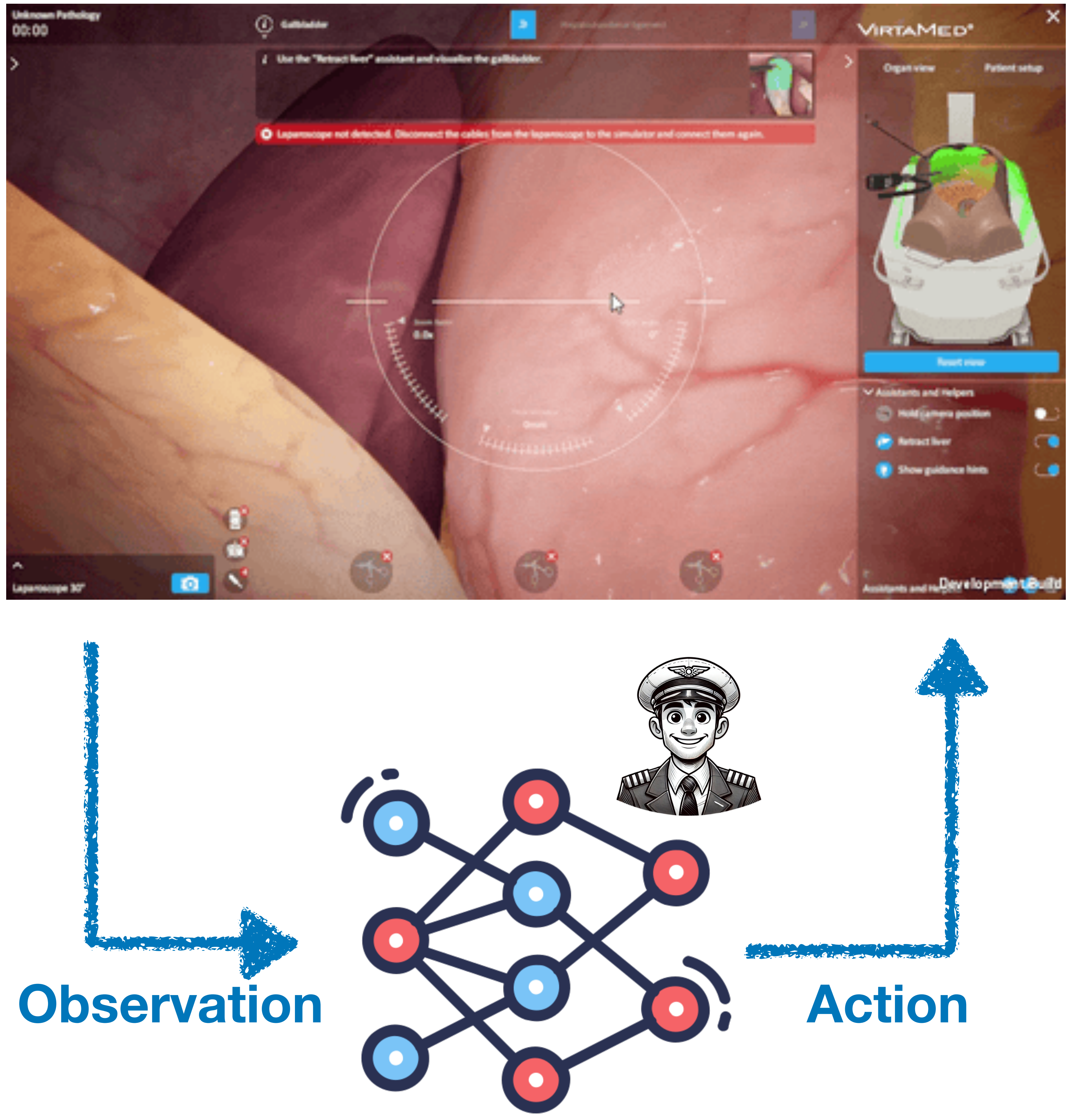
**How to leverage ML technology?**

- How to evaluate trainees in a data-driven manner?
- How to provide meaningful feedback and assistance?

Devise algorithmic pipeline based on RL agents serving as surgical co-pilots

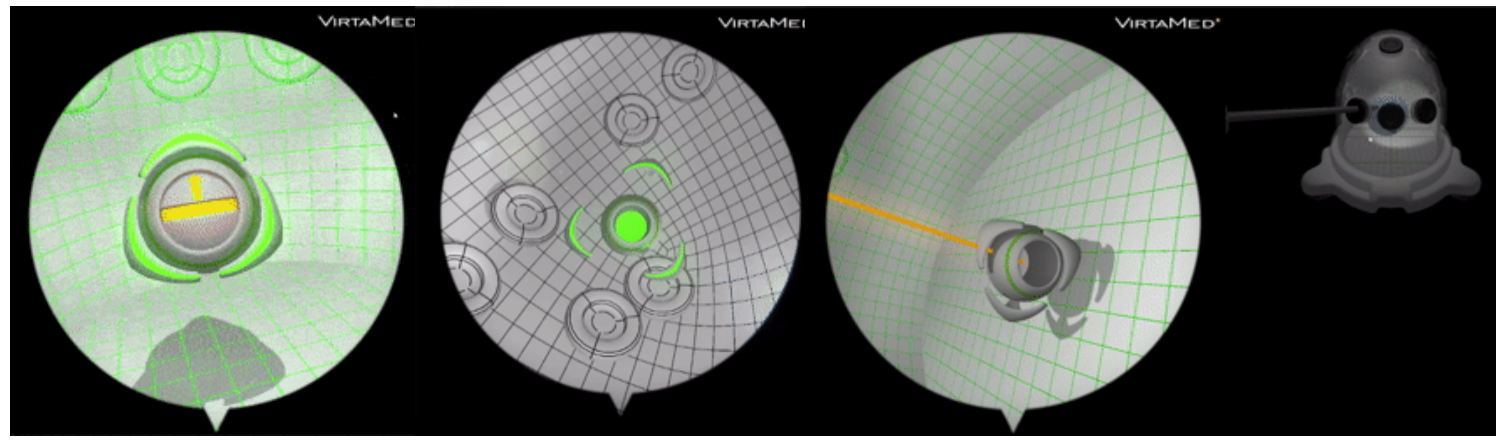
# Training surgical agents

## Leveraging control loop approach

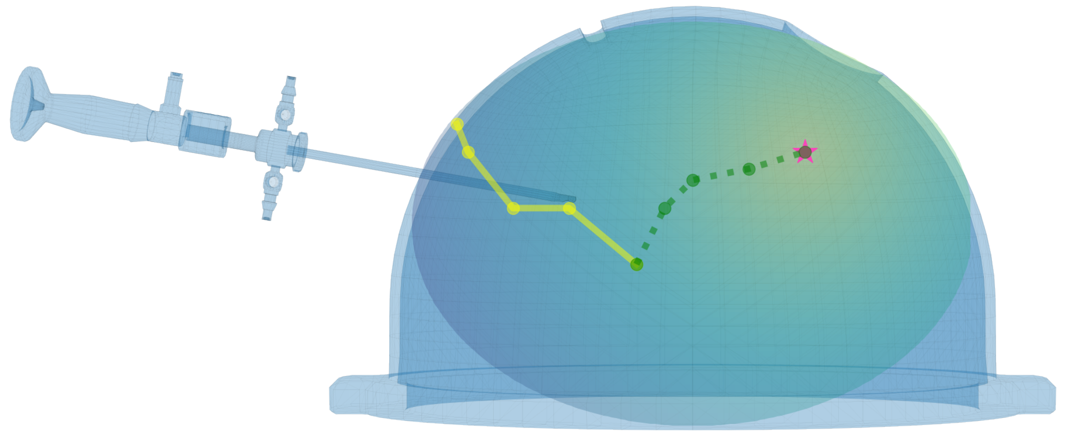



Behavioural Policy (and Reward)

**Benchmark suite of tasks for surgical RL agents**

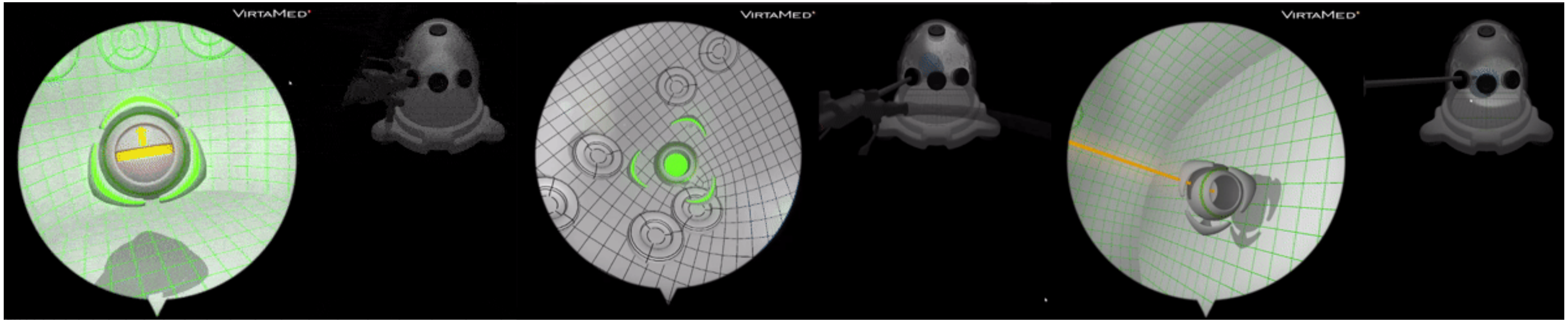


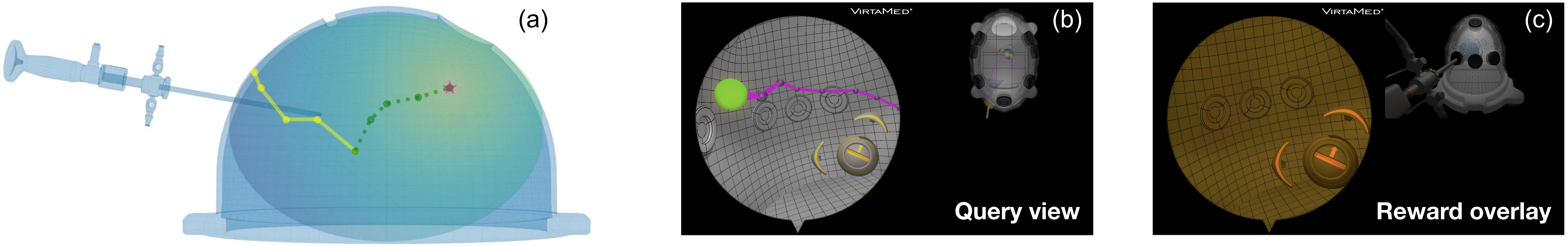
**Learning from demonstrations for purposes of evaluation**



# FASTRL Benchmark suite

## Fundamentals of Arthroscopic Surgery Training



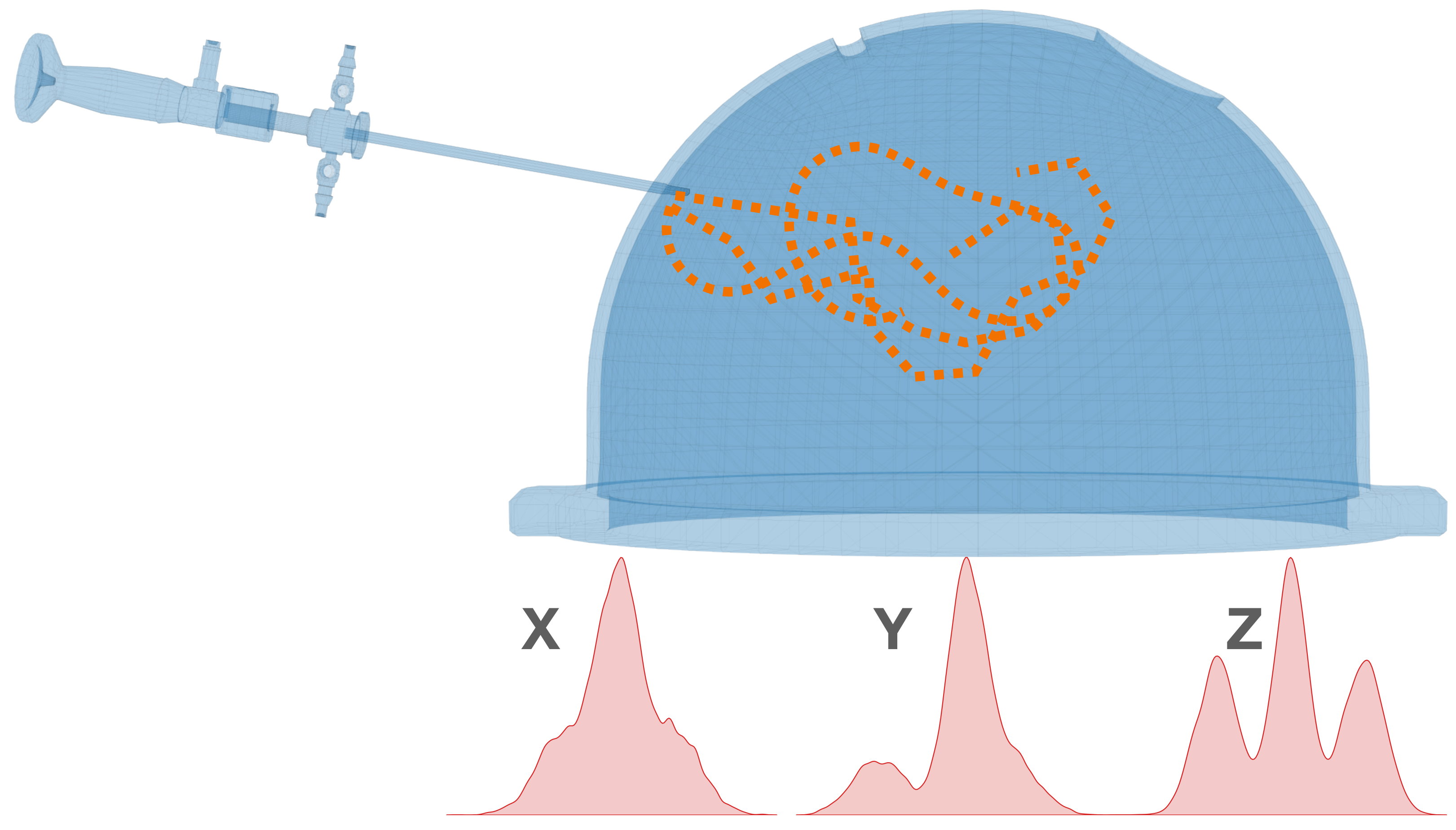



# Algorithmic pipeline for surgical assistance

## Learning from demonstrations

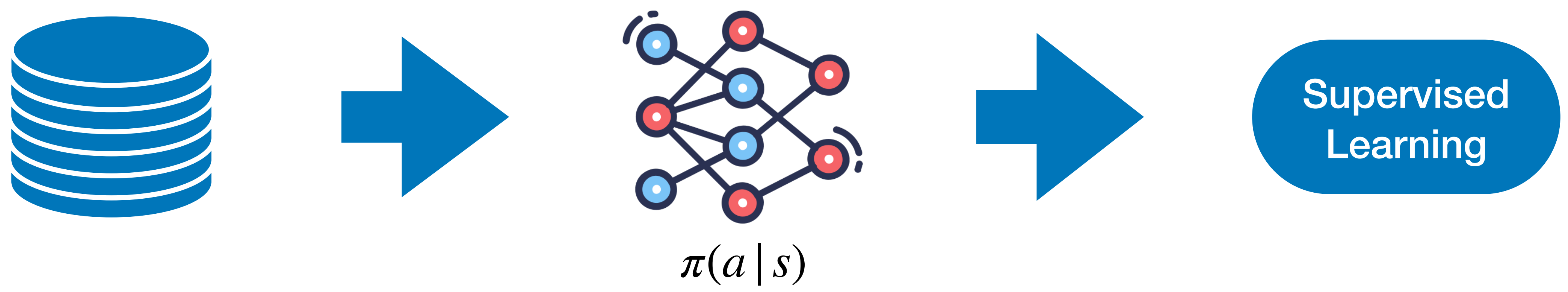
### Given a dataset of trajectories

$$\mathscr{D}_E = \{\xi_i\}_{i \le N} = \{(s_0, a_0, \ldots, a_{T_i-1}, s_{T_i})\}_{i \le N}$$

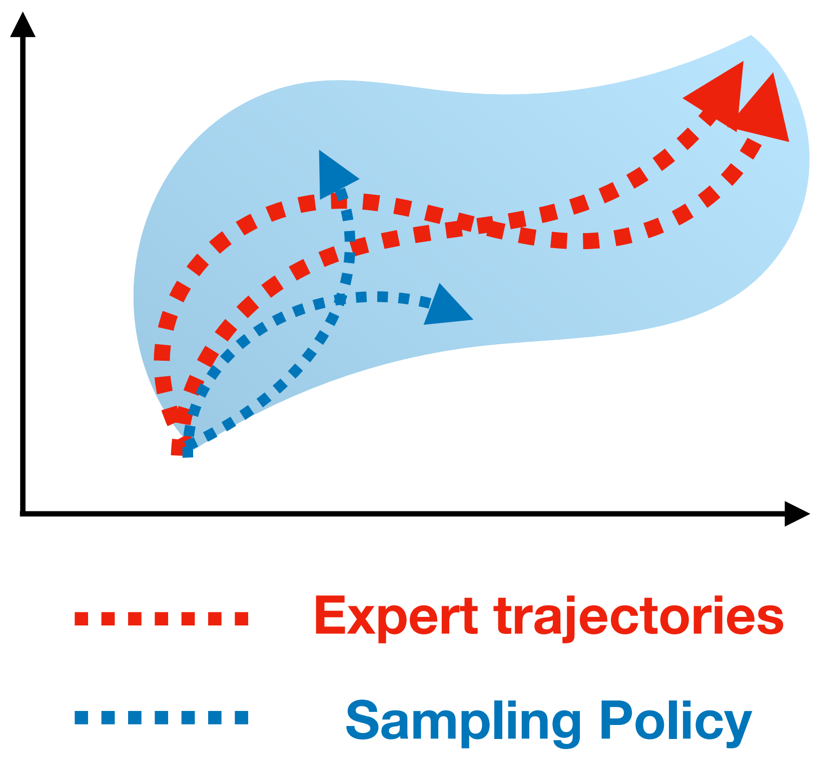



### How to learn behaviours from demonstrations?

**1. Behavioural cloning**

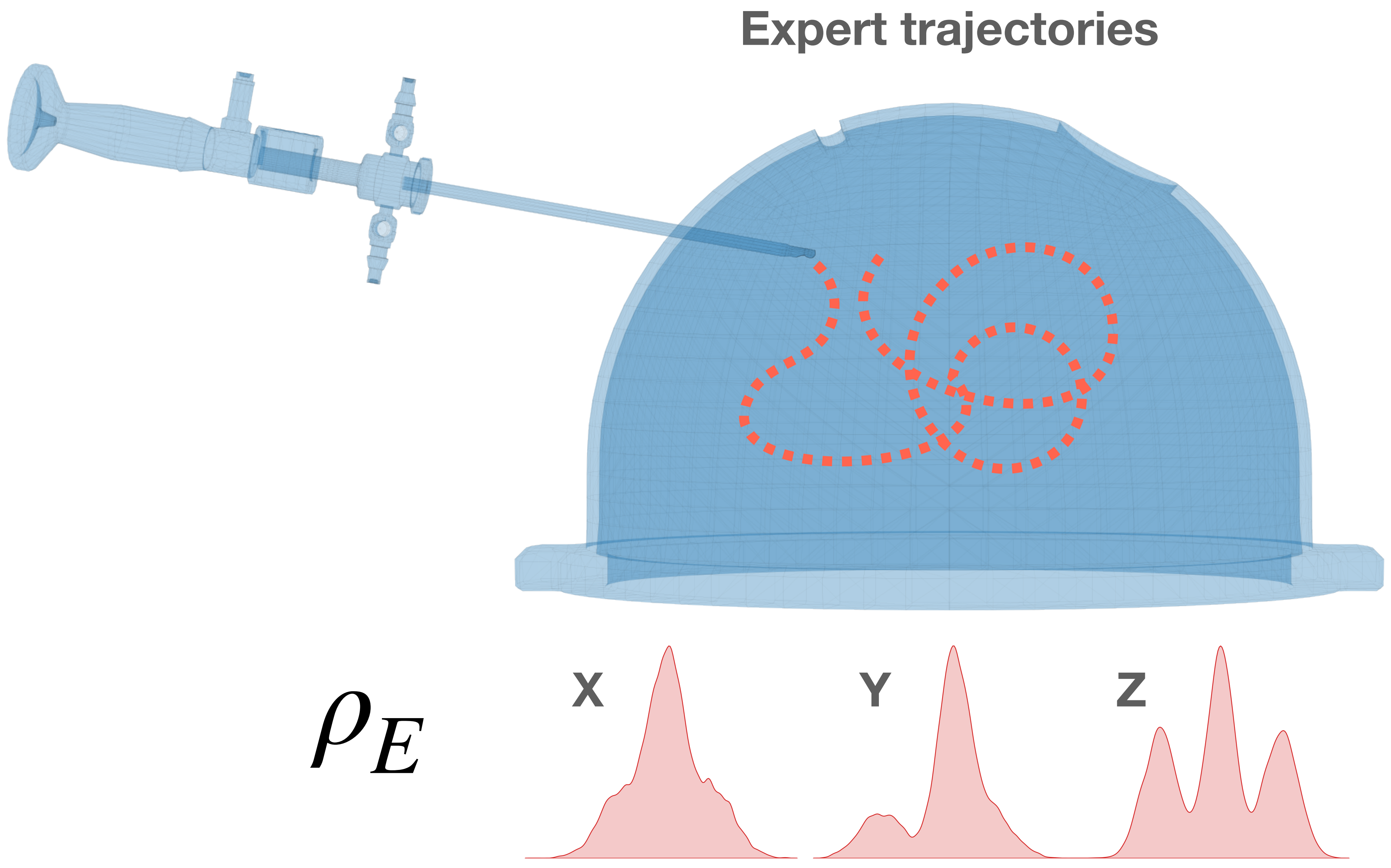


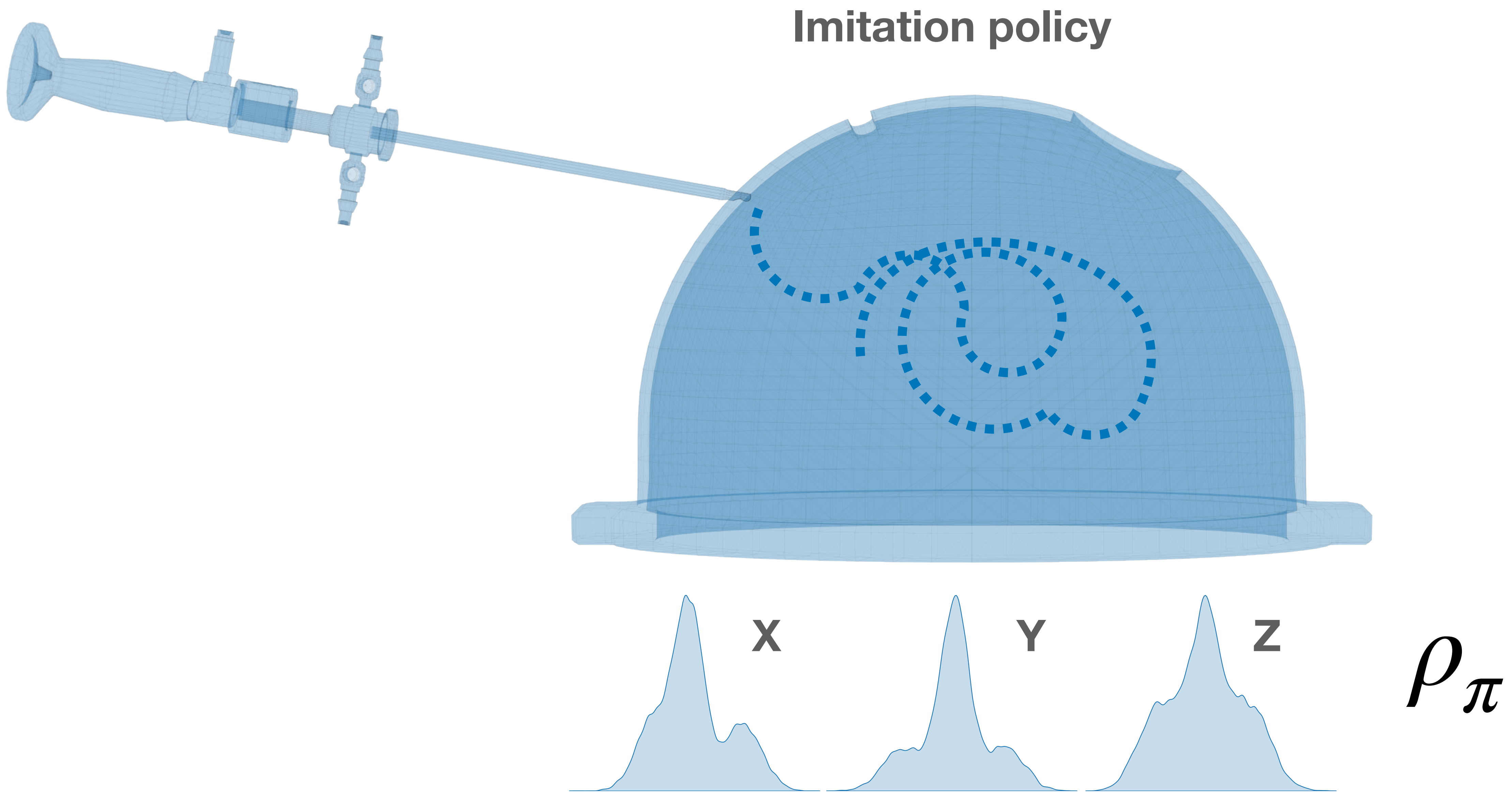

**2. Distribution matching**

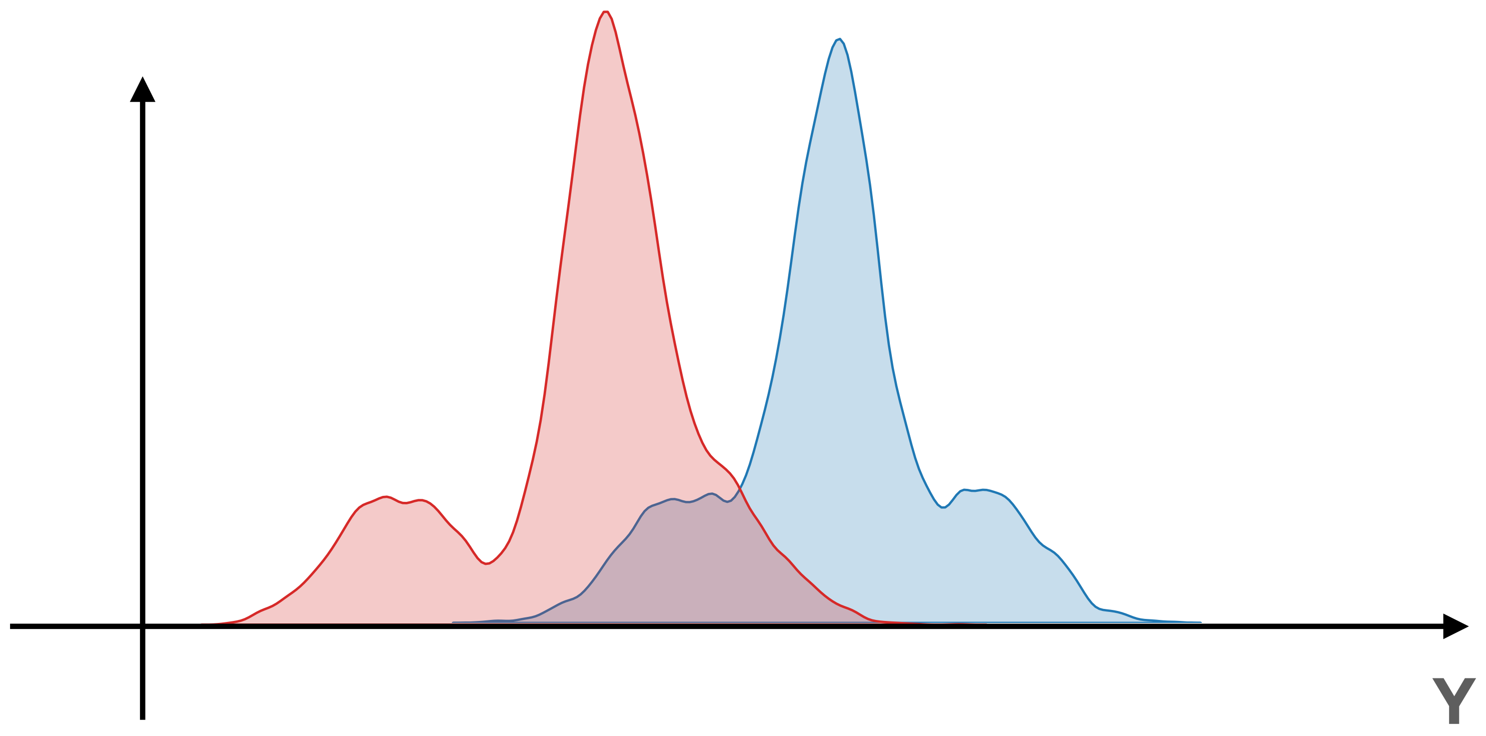



# Algorithmic pipeline for surgical assistance

## Distribution matching as basis for imitation learning



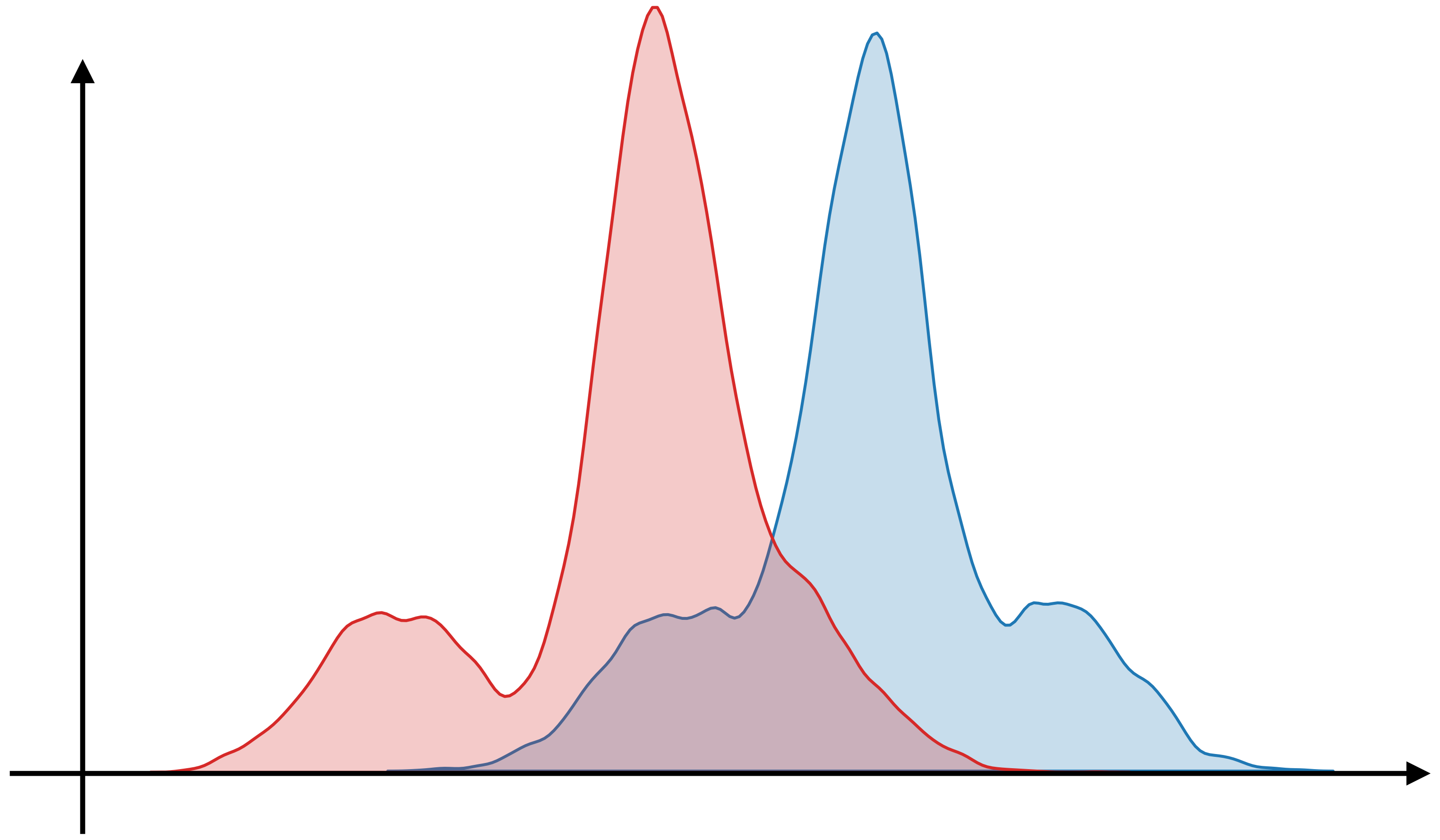

$$\min_{\pi} D(\rho_E, \rho_\pi)$$

Imitation Learning Objective

# Algorithmic pipeline for surgical assistance

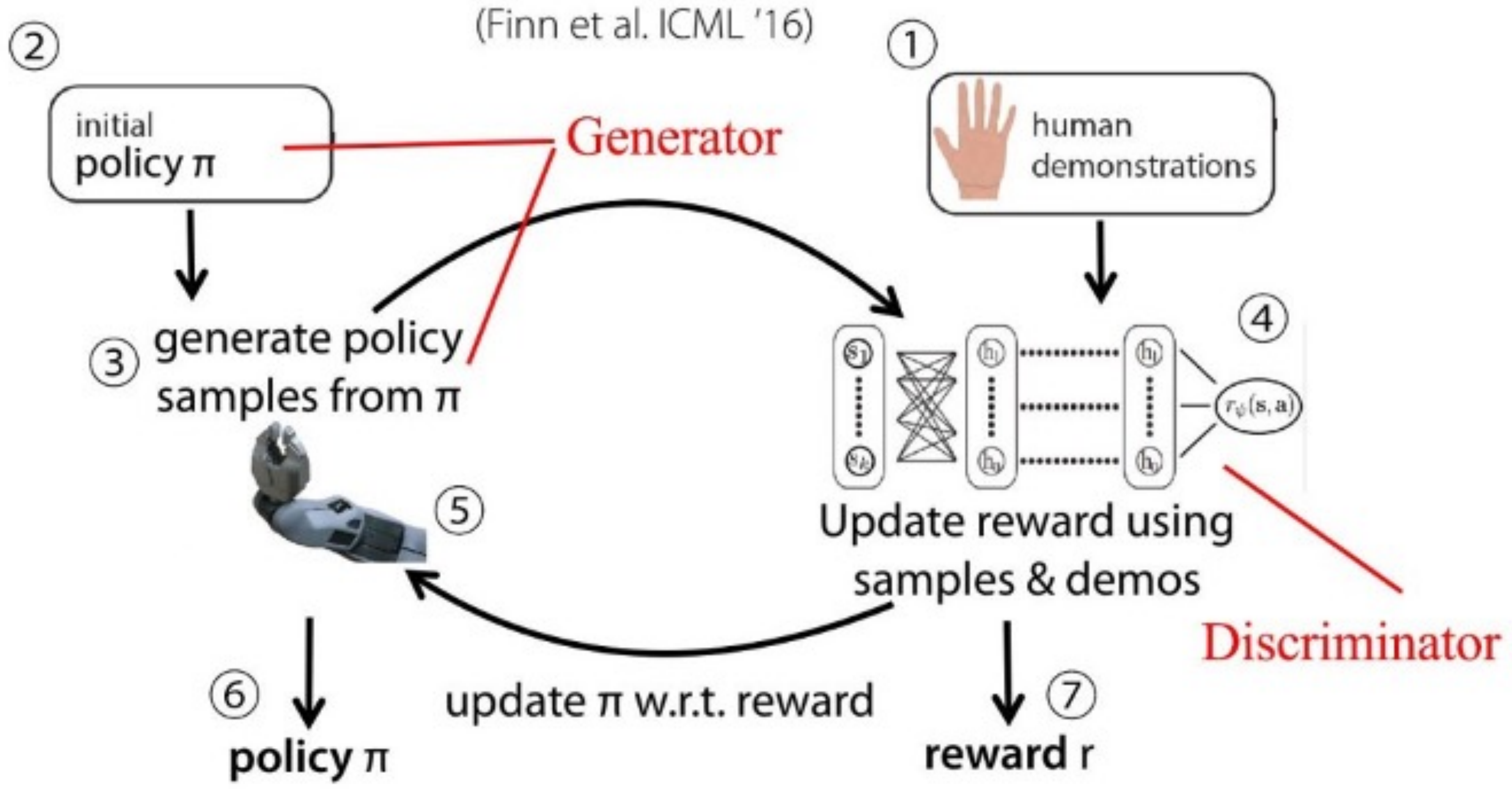
## Distribution matching: algorithmic choices

$$\min_{\pi} D(\rho_E, \rho_\pi)$$



How can we recover a policy and reward?

**How to optimize** $\min_{\pi} D(\rho_E, \rho_\pi)$ **?** **Min-max problem**

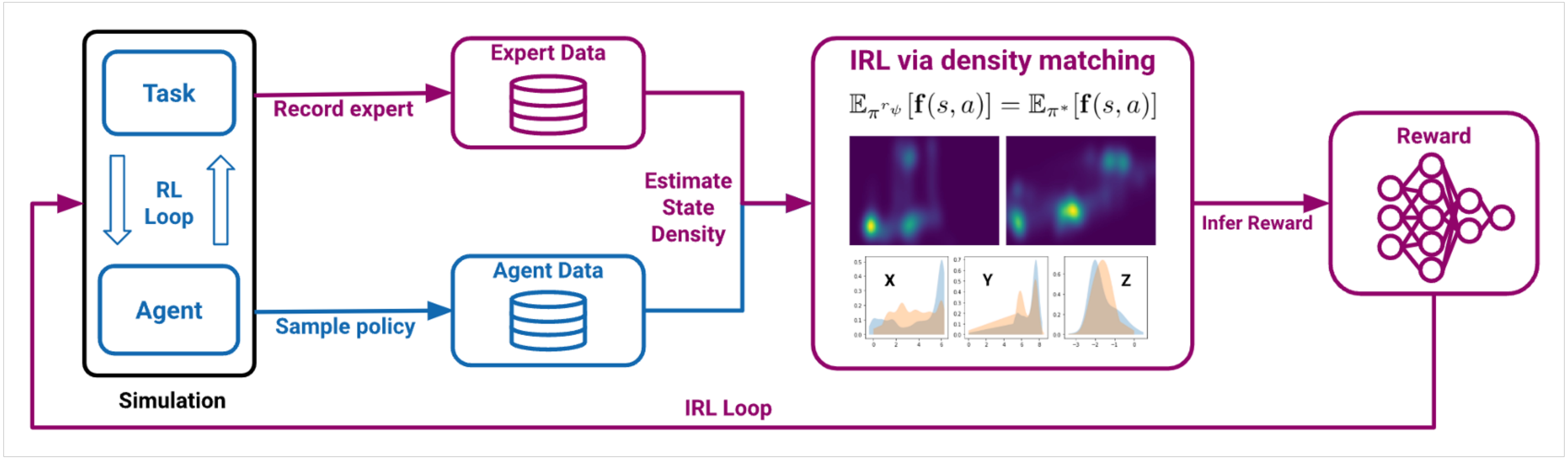



**Variational dual of f-divergences**

$$D_f(\rho_E, \rho_\pi) = \sup_{g} \mathbb{E}_{\rho_E}[g(x)] - \mathbb{E}_{\rho_\pi}[f^*(g(x))]$$

# Algorithmic pipeline for surgical assistance

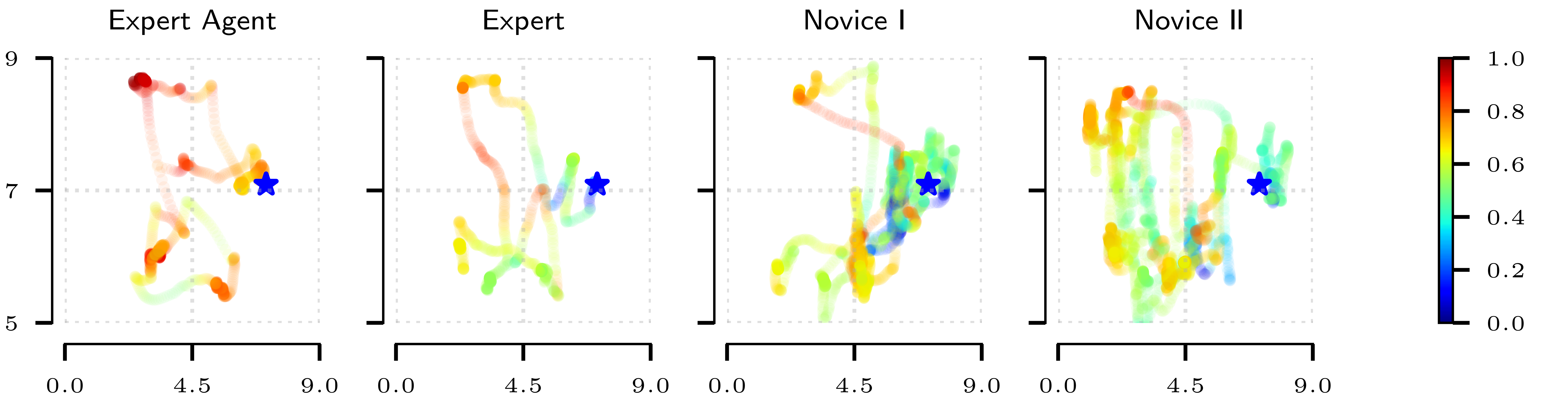
## Putting it all together

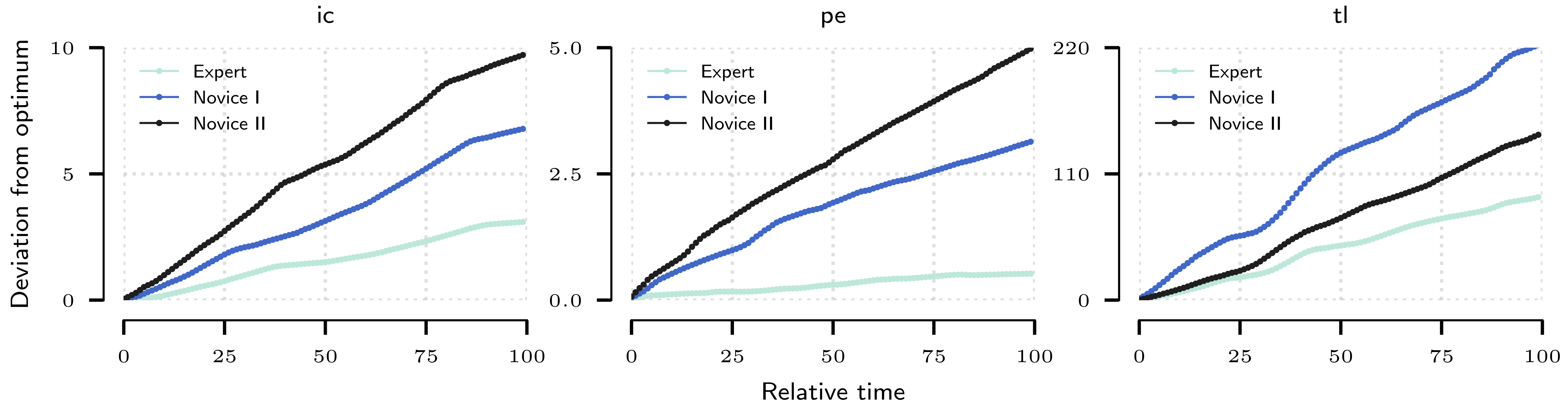



**Can use policy and reward for assistance and evaluation**

# Algorithmic pipeline for surgical assistance

## Results: human evaluation example

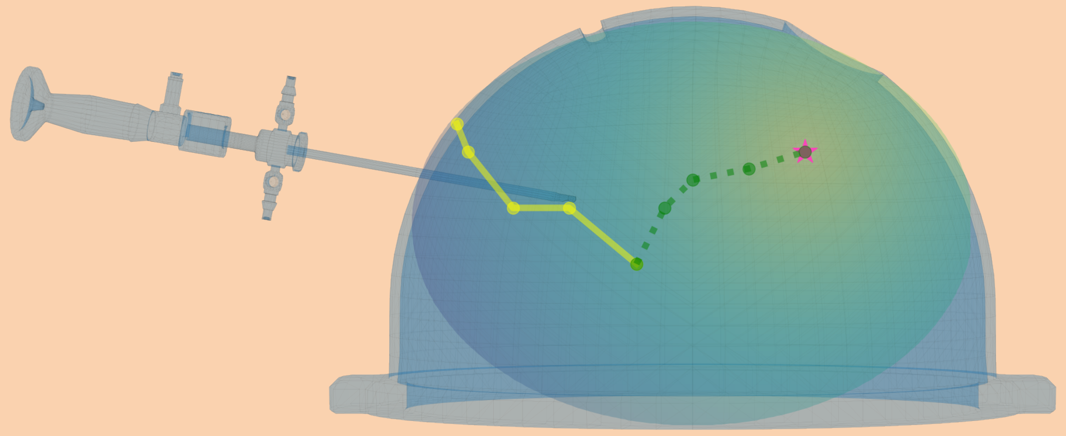


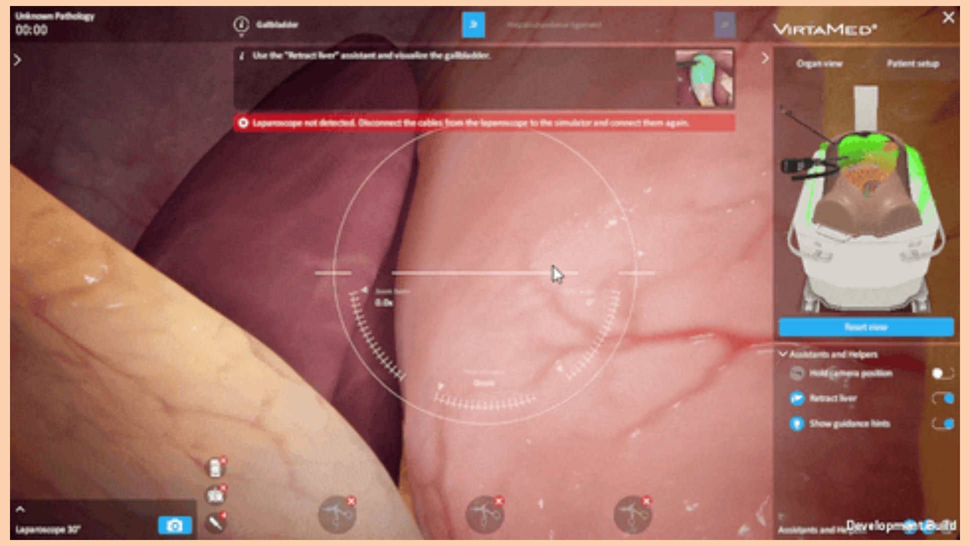




| Agent ID | $r_\psi$ | $V_\phi$ | $r_{heur}$ | Trajectory length |
|---|---|---|---|---|
| fw | $0.993_{\pm 0.004}$ | $0.876_{\pm 0.076}$ | $0.999_{\pm 0.0002}$ | $1600.6_{\pm 971.0}$ |
| fm | $0.732_{\pm 0.019}$ | $0.728_{\pm 0.016}$ | $0.742_{\pm 0.016}$ | $1821.5_{\pm 104.8}$ |
| an | $0.720_{\pm 0.042}$ | $0.716_{\pm 0.040}$ | $0.729_{\pm 0.039}$ | $1929.2_{\pm 277.4}$ |
| mk | $0.617_{\pm 0.117}$ | $0.627_{\pm 0.121}$ | $0.634_{\pm 0.106}$ | $2578.5_{\pm 745.8}$ |
| mv | $0.518_{\pm 0.094}$ | $0.521_{\pm 0.094}$ | $0.529_{\pm 0.095}$ | $3313.5_{\pm 665.5}$ |
| io | $0.374_{\pm 0.271}$ | $0.339_{\pm 0.249}$ | $0.384_{\pm 0.278}$ | $4290.8_{\pm 1981}$ |

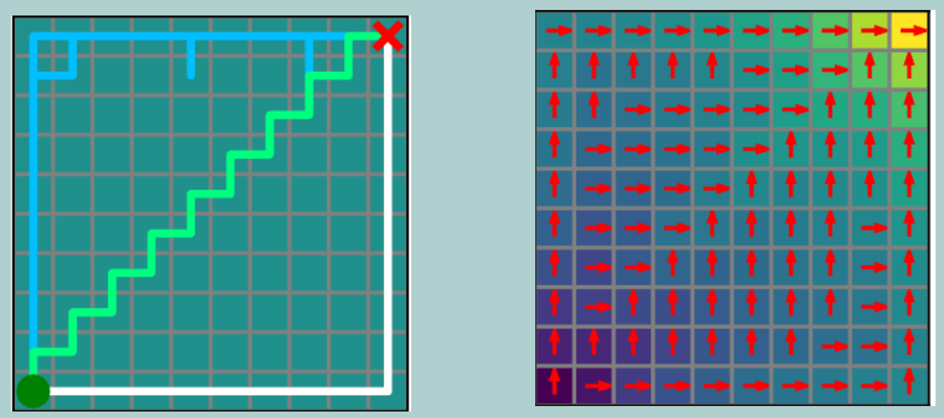
# Overview

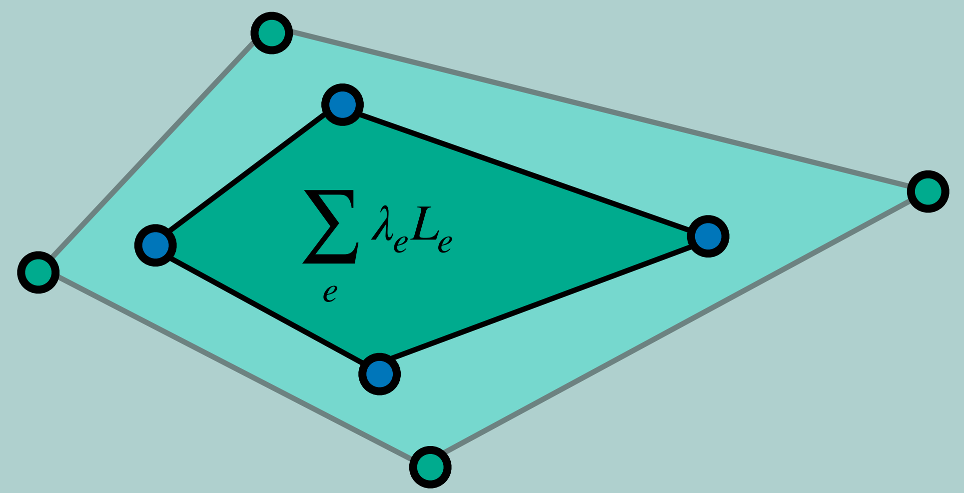
**Application:**
**Algorithmic RL pipeline in Surgical Digital Twins**



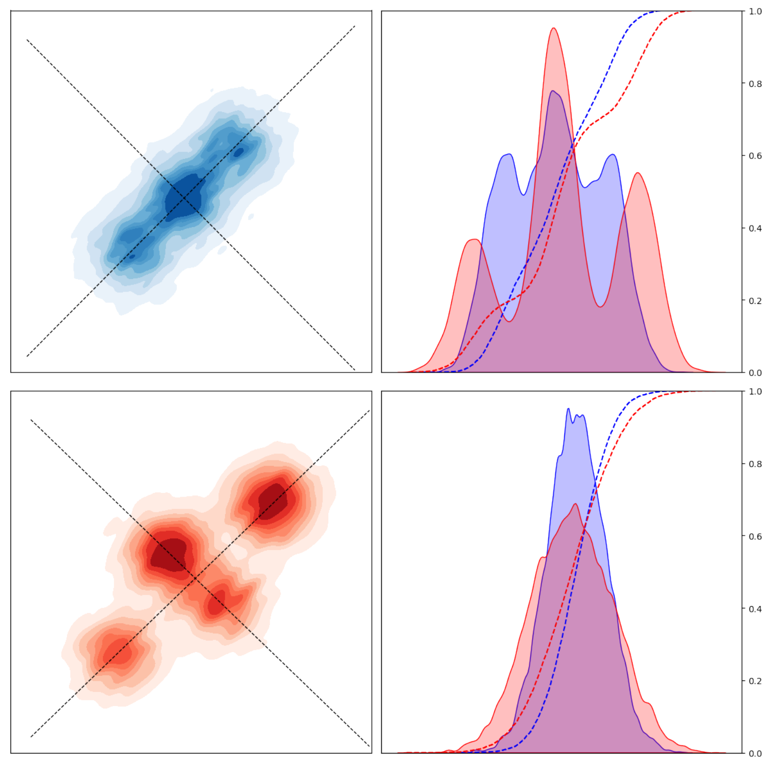


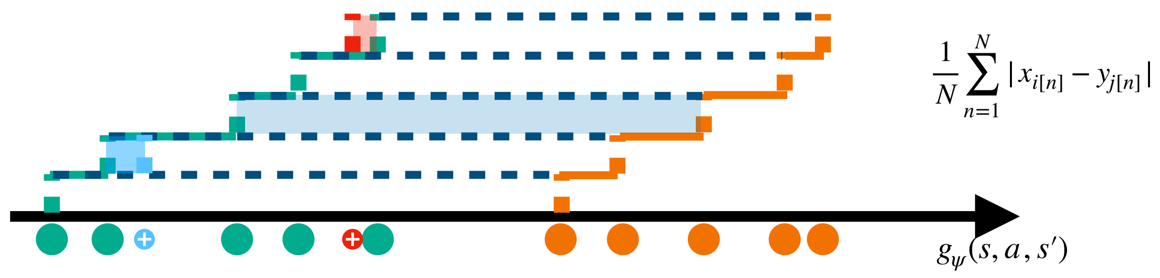
**Method I:**
**Addressing Reward Generalization using Causal Invariance**






**Method II:**
**Addressing Data Efficiency in Imitation Learning using Sliced Optimal Transport**

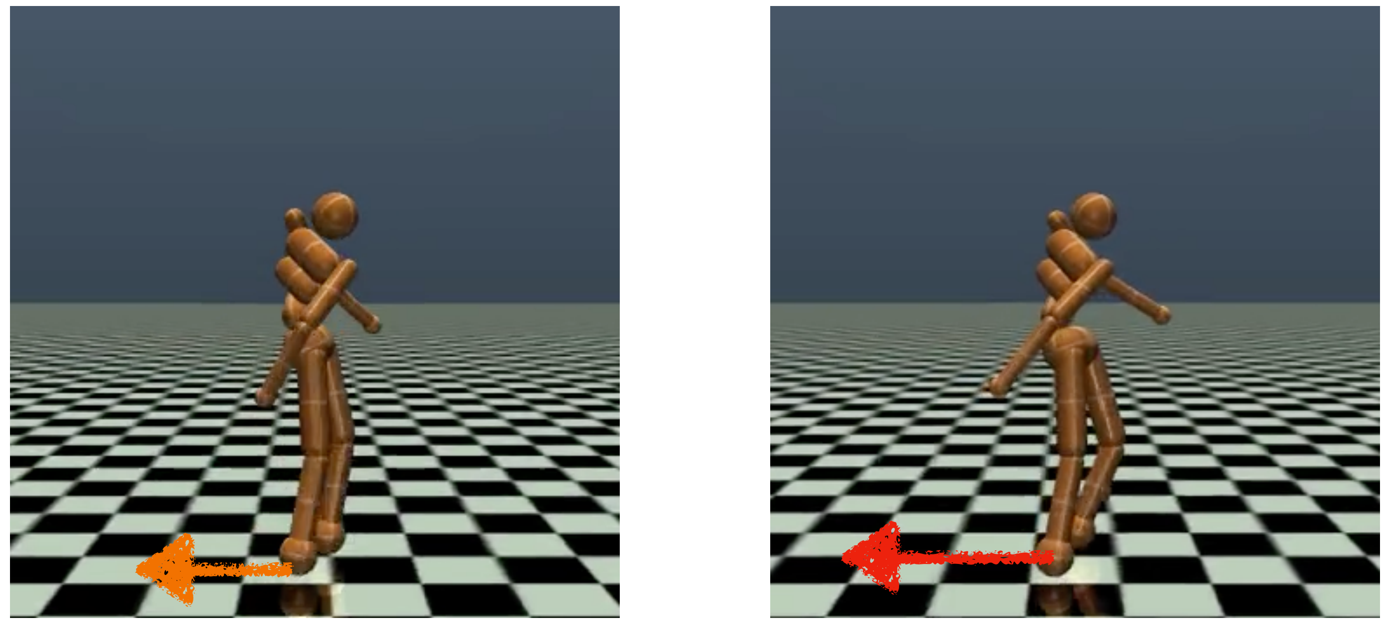




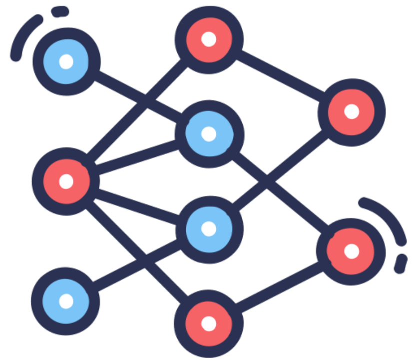
# Method: addressing generalization

## Defining generalization

### Generalization in reinforcement learning



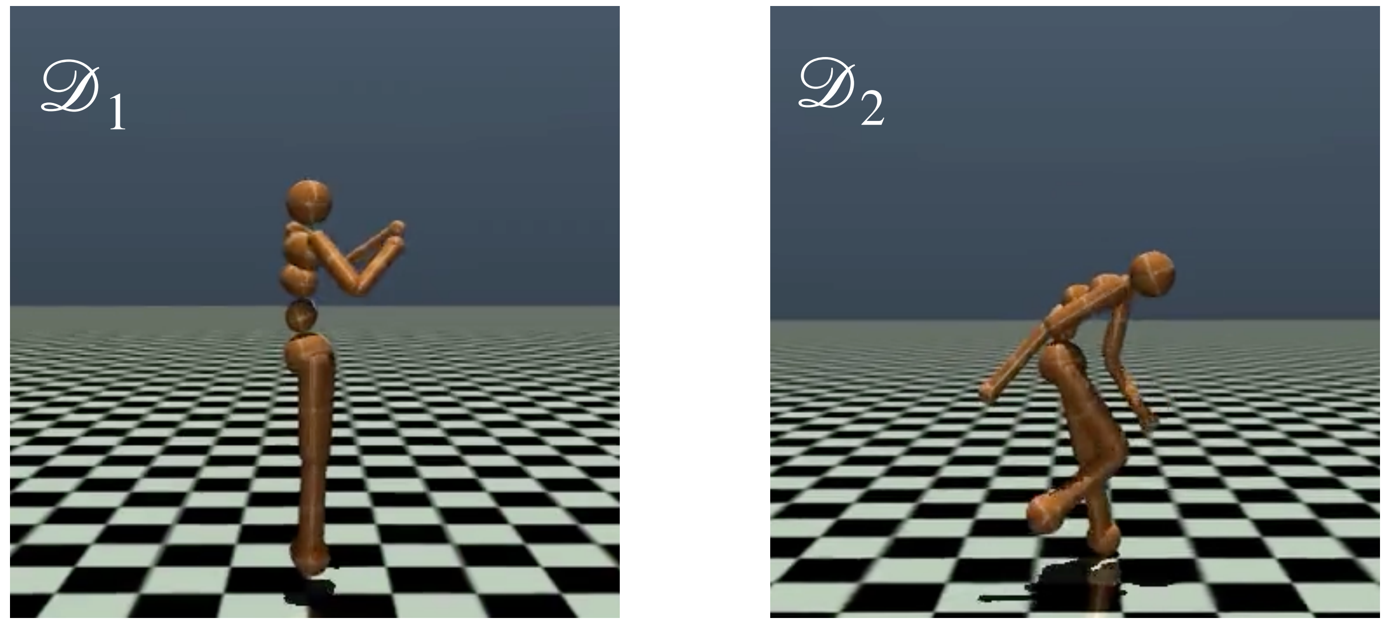
Changing friction coefficient



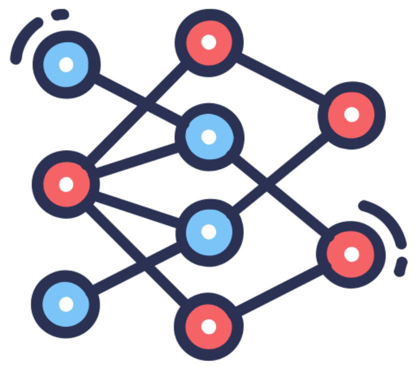
Policy

Can policy perform well on unseen dynamics?

### Generalization in inverse reinforcement learning




Datasets with varying gait



Reward

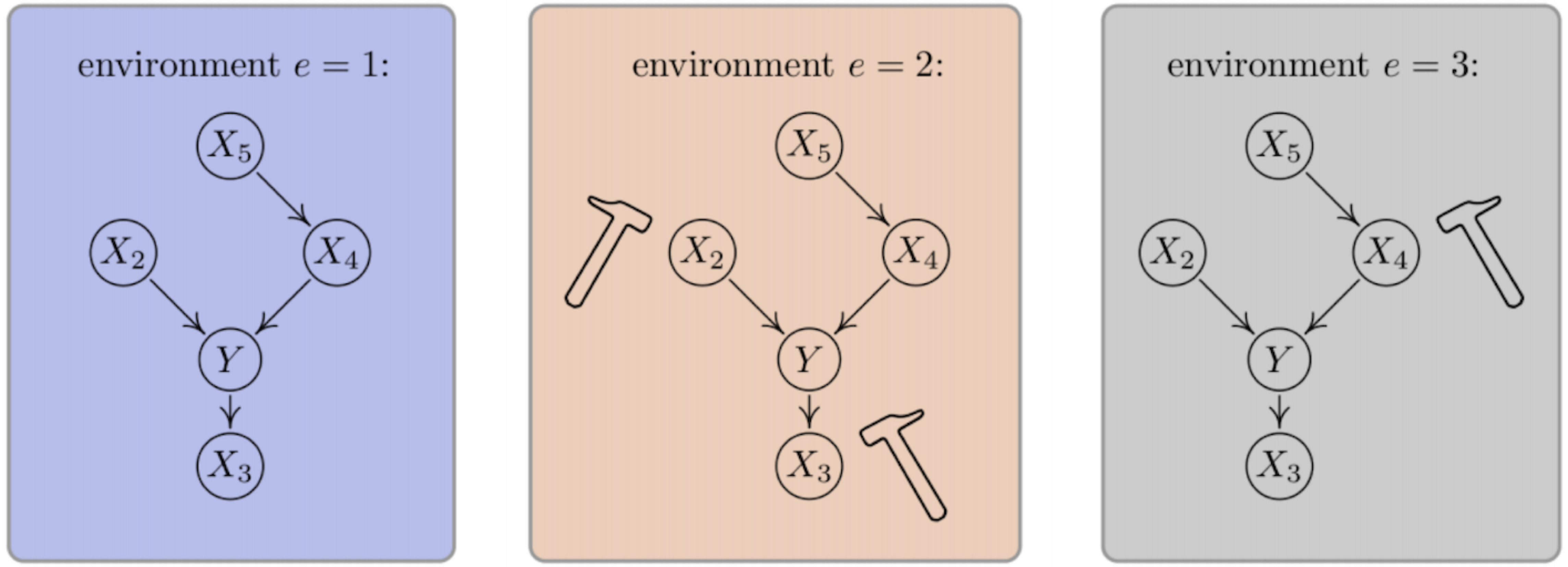
Can we recover reward which provides useful training signal on unseen dynamics?

*Image: Ministry of Silly Walks, BBC*

*Zhang et al. Invariant Causal Prediction for Block MDPs*

# A little excursion

## Invariant causal prediction




Peters et al. 2015

$$\mathcal{D}_e = (\mathbf{X}^e, Y^e) \quad e = \{1 \ldots K\}$$

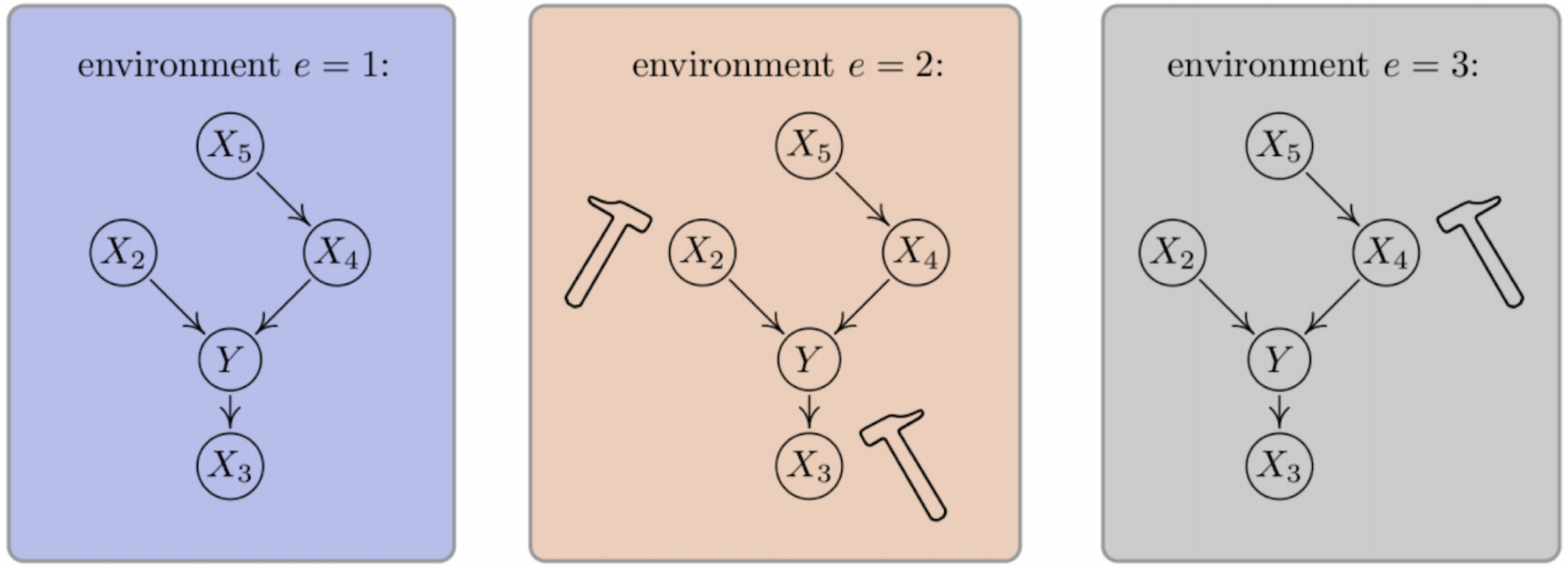
**Prediction Task** $g : \mathcal{X} \rightarrow \mathcal{Y}$

$$P(Y \mid X_1, \ldots, X_n)$$

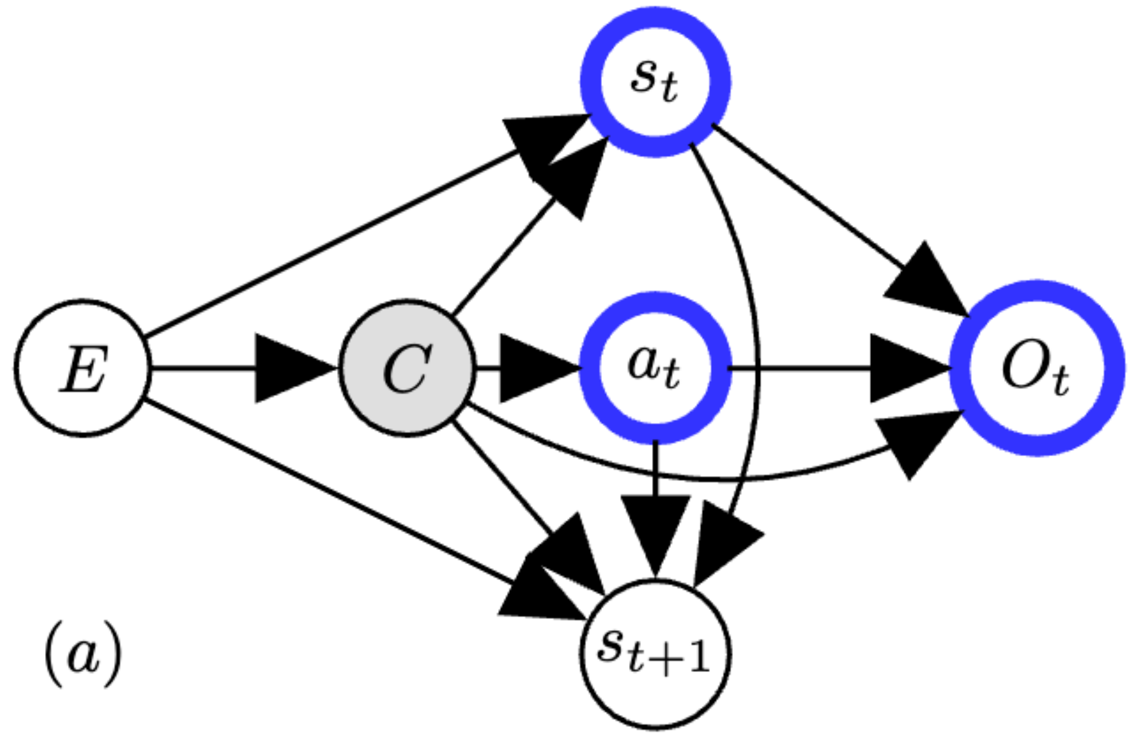
**Should be stable across settings!**

# Reward generalization

## Mapping ICP intuition to IRL




Peters et al. 2015




**Inverse reinforcement learning corresponds to learning the optimality conditional** $P(O_t \mid s_t, a_t)$

**Main intuition: optimality label distribution should be stable across expert demonstrations**

# Reward generalization

## How to adapt principle to IRL setting?

**Primal case: Gibbs MLE Problem over trajectories**

$$\max_{\varphi,\psi} \mathbb{E}_{\xi \sim \mathscr{D}_E}\left[\log \frac{1}{Z_{\psi,\varphi}} p(\xi \,|\, \psi, \varphi)\right]$$

**Dual case: total variation distance over transitions**

$$\max_{g \in \mathscr{G}} \min_{q \in \mathscr{P}} \mathbb{E}_{(s,a,s') \sim \mathscr{D}_E}[g(s,a,s')] - \mathbb{E}_{(s,a,s') \sim q}[g(s,a,s')]$$

**Invariant Risk Minimization**

**Decompose reward / critic into representation $\varphi$ and predictor $\psi$**

$$\longrightarrow \boxed{\varphi} \longrightarrow \boxed{\psi} \longrightarrow$$

**IRM bi-level optimization problem**

$$\max_{\varphi,\psi} \sum_{e \in \mathscr{E}_{tr}} \sum_{\xi \in \mathscr{D}_e} L_e(\xi, \psi, \varphi)$$

$$\textbf{s.t. } \psi \in \operatorname*{argmax}_{\bar{\psi}} \sum_{\xi \in \mathscr{D}_e} L_e(\xi, \bar{\psi}, \varphi)$$

**IRMv1 penalty**

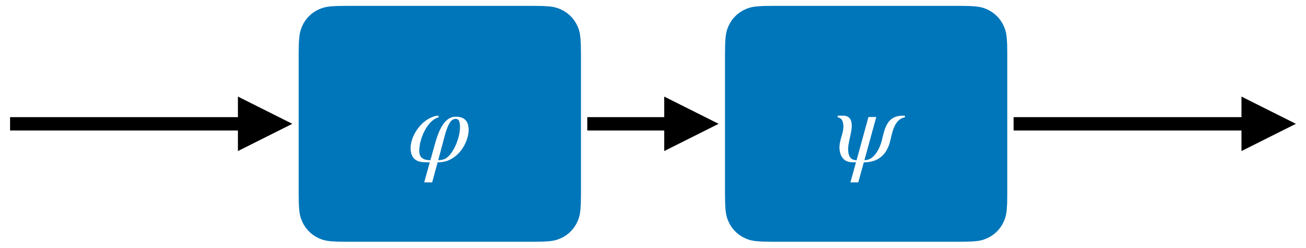
$$\mathbb{D}(\psi, \varphi; e) = ||\nabla_\psi \mathscr{L}^e(\psi, \varphi)|_{\psi=1.0}||^2$$

# Reward generalization

## How to adapt principle to IRL setting?

**Invariant Risk Minimization**

**Decompose reward / critic into representation $\varphi$ and predictor $\psi$**




**IRM bi-level optimization problem**

$$\max_{\varphi,\psi} \sum_{e \in \mathscr{E}_{tr}} \sum_{\xi \in \mathscr{D}_e} L_e(\xi, \psi, \varphi)$$

$$\textbf{s.t. } \psi \in \operatorname*{argmax}_{\bar{\psi}} \sum_{\xi \in \mathscr{D}_e} L_e(\xi, \bar{\psi}, \varphi)$$

**IRMv1 penalty**

$$\mathbb{D}(\psi, \varphi; e) = \left|\left| \nabla_\psi L^e(\psi, \varphi) |_{\psi=1.0} \right|\right|^2$$

**IRMv1 penalty applied to IRL as regularizer**

**Primal:** $$\max_{\varphi,\psi} \sum_{e \in \mathscr{E}_{tr}} \left( \mathbb{E}_{\xi \in \mathscr{D}_e} \left[ \log \left( \frac{1}{Z_{\psi,\varphi}} \exp(\psi^T \varphi(\xi)) \right) + \lambda \mathbb{D}(\psi, \varphi, e) \right] \right)$$

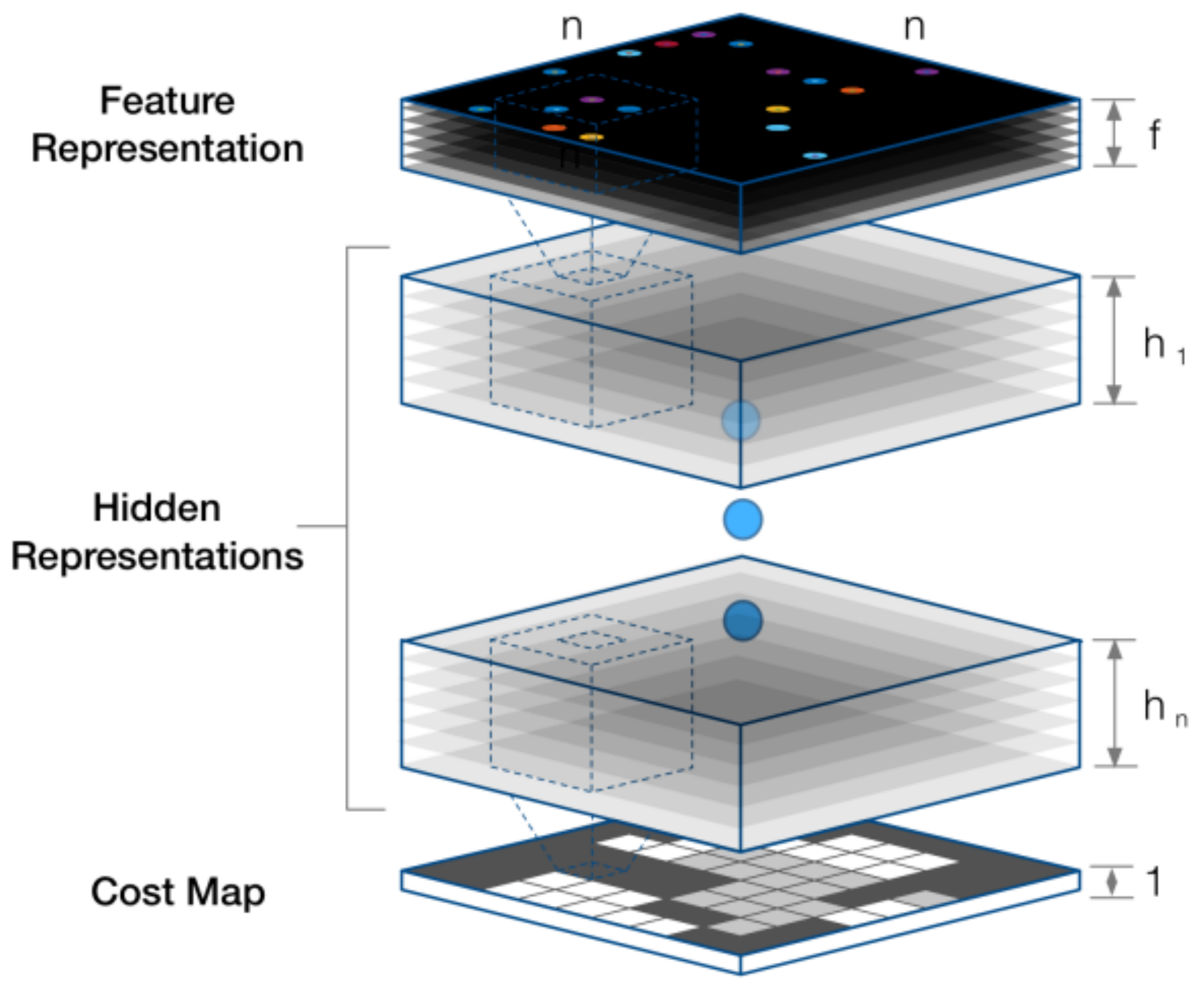
**Dual:** $$\max_{\psi,\varphi} \sum_{e \in \mathscr{E}_{tr}} \min_q \left( \mathbb{E}_{(s,a,s') \sim \mathscr{D}_E^e}[g_{\psi,\varphi}(s,a,s')] - \mathbb{E}_{(s,a,s') \sim q}[g_{\psi,\varphi}(s,a,s')] + \lambda \mathbb{D}(\psi, \varphi, e) \right)$$

# Reward generalization

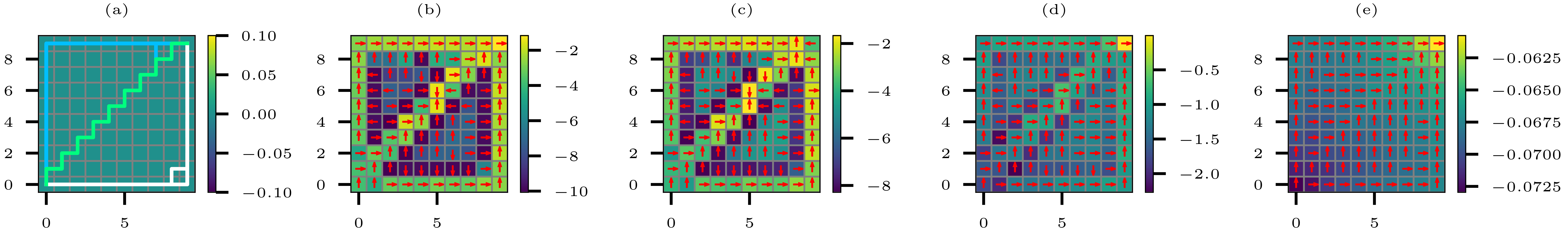
## Evaluation: gridworld navigation

**Solve primal problem using diverse demonstration data**

$$\max_{\varphi,\psi} \sum_{e\in\mathcal{E}_{tr}} \left( \mathbb{E}_{\xi\in\mathcal{D}_e} \left[ \log \left( \frac{1}{Z_{\psi,\varphi}} \exp(\psi^T \varphi(\xi)) \right) + \lambda \mathbb{D}(\psi, \varphi, e) \right] \right)$$



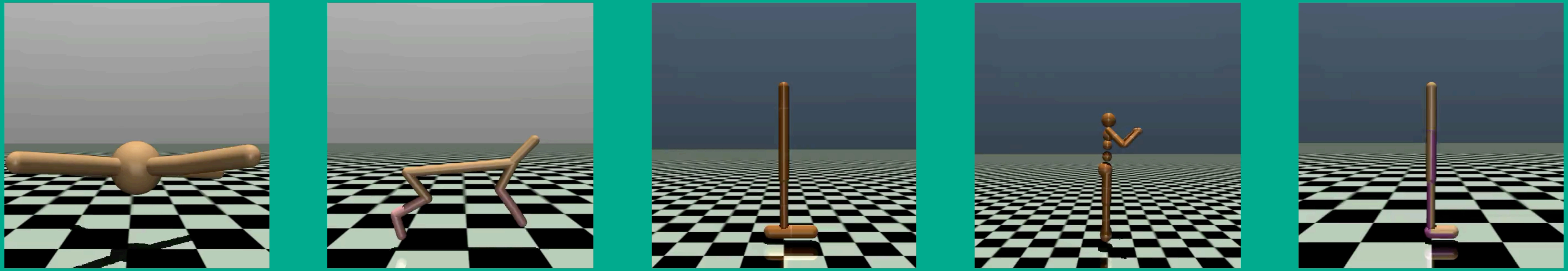

*Wulfmeier et al., 2016*



# Reward generalization

## Evaluation: robotic locomotion transfer setting

**Mujoco Benchmark**

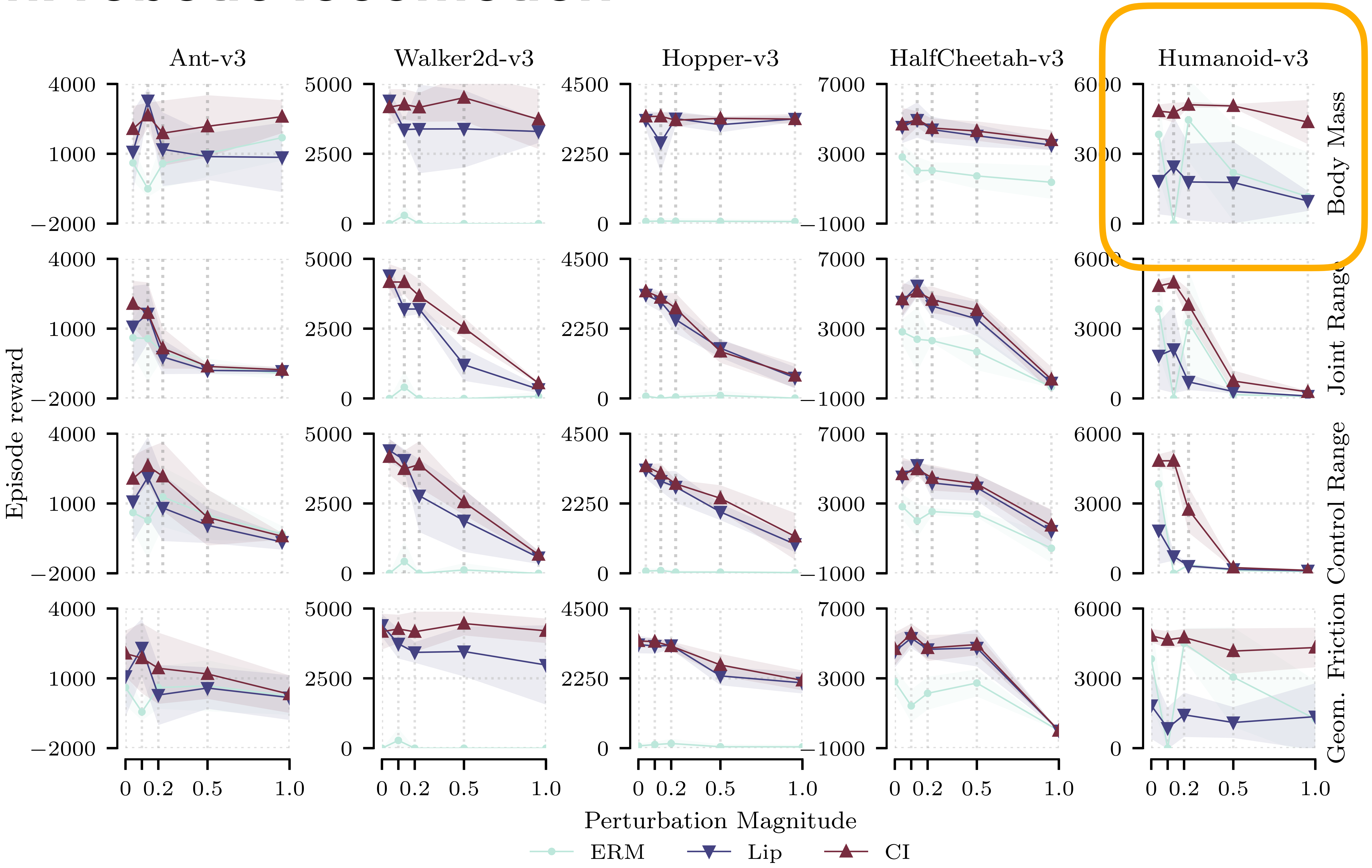


**Evaluation protocol:**

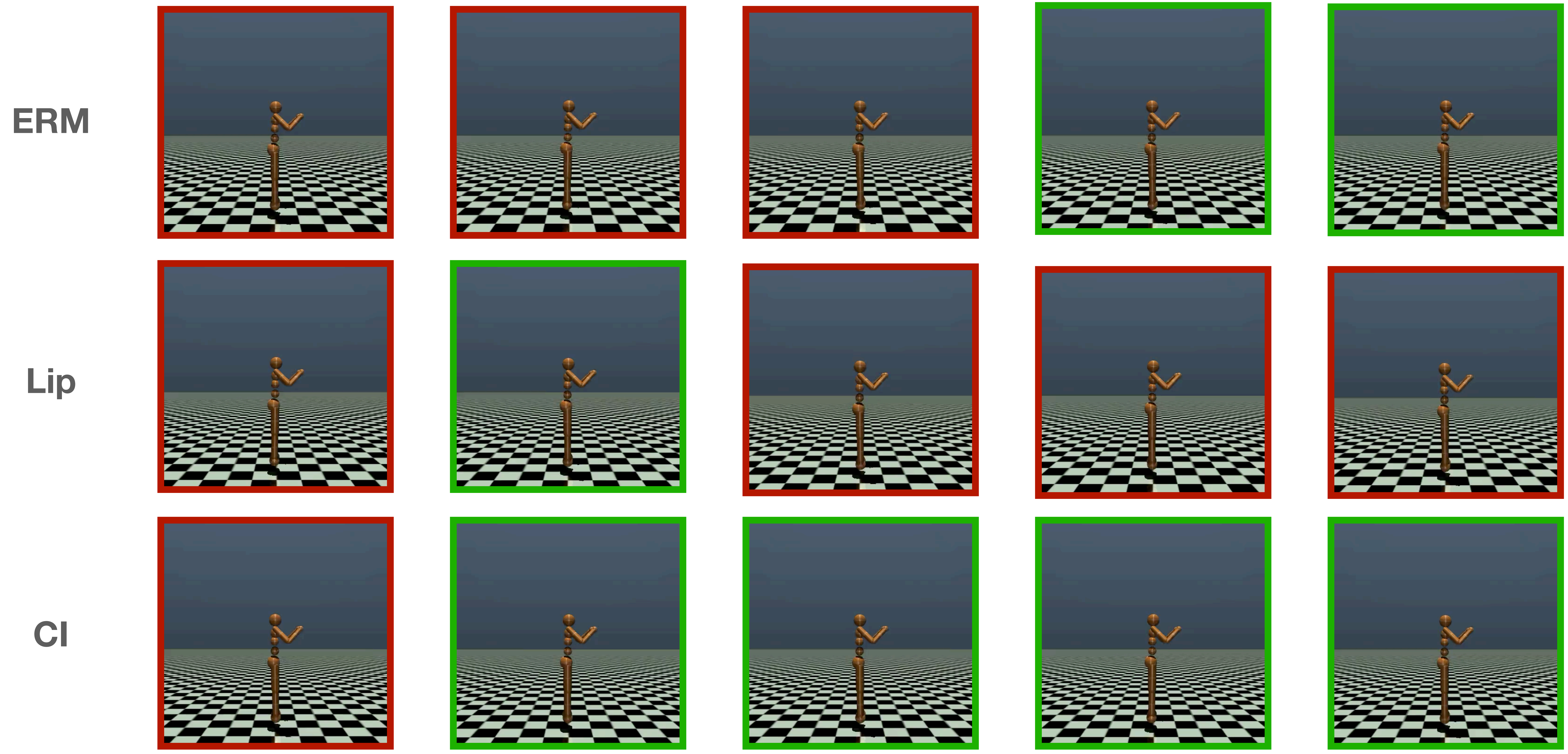
1. Generate diverse trajectories by **applying structured noise**
2. **Solve dual problem by training adversarial IRL algorithms to recover rewards**
3. Use recovered reward to **train policies on perturbed dynamics**

# Reward generalization

## Evaluation: robotic locomotion

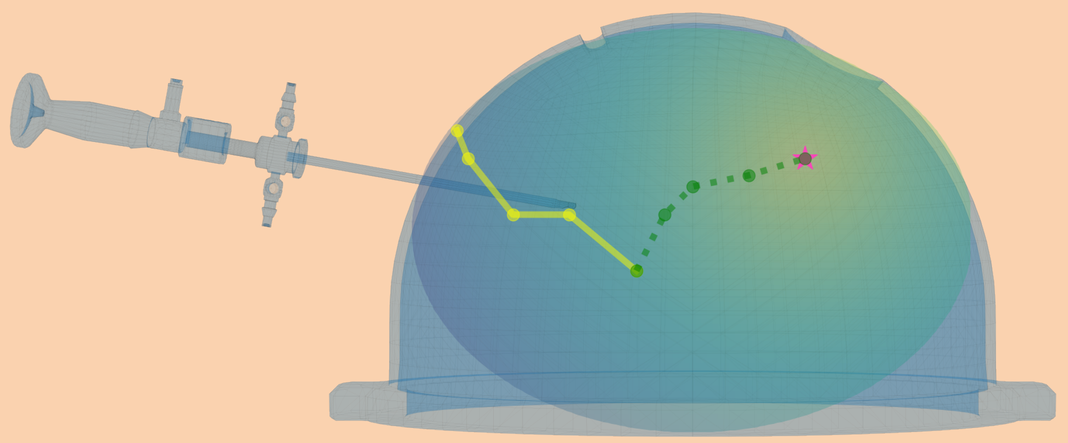


# Ministry of silly walks

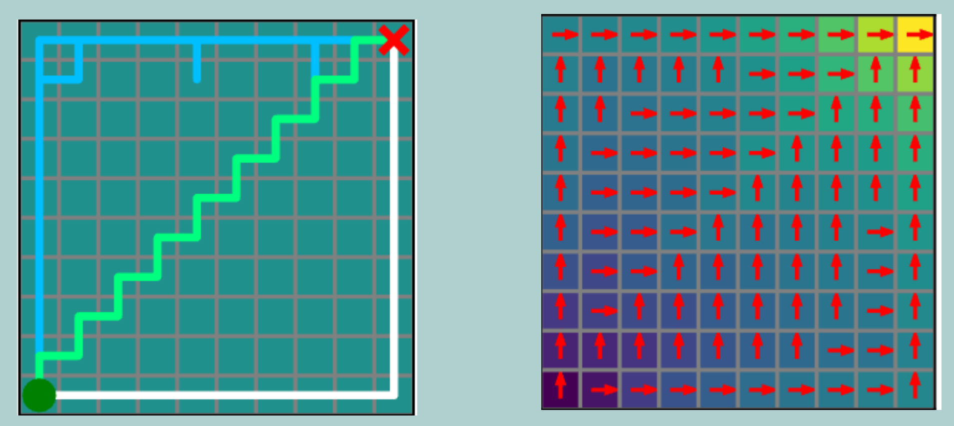


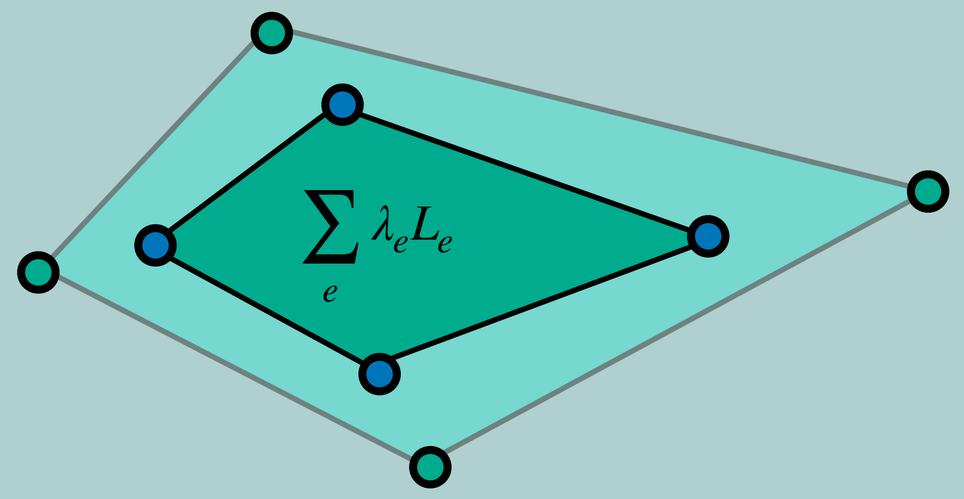
# Overview

**Application:**
**Algorithmic RL pipeline in Surgical Digital Twins**

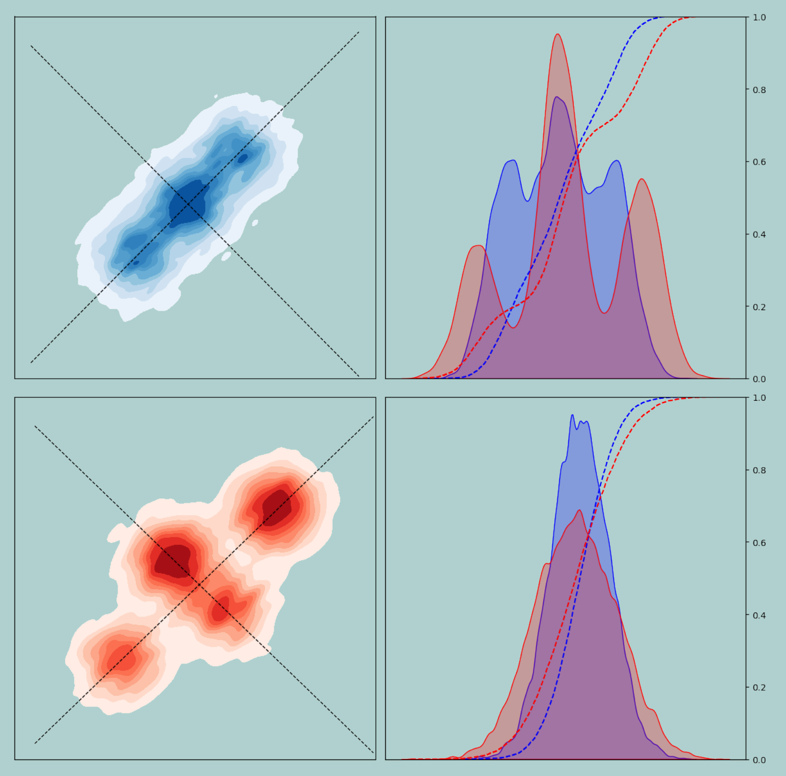


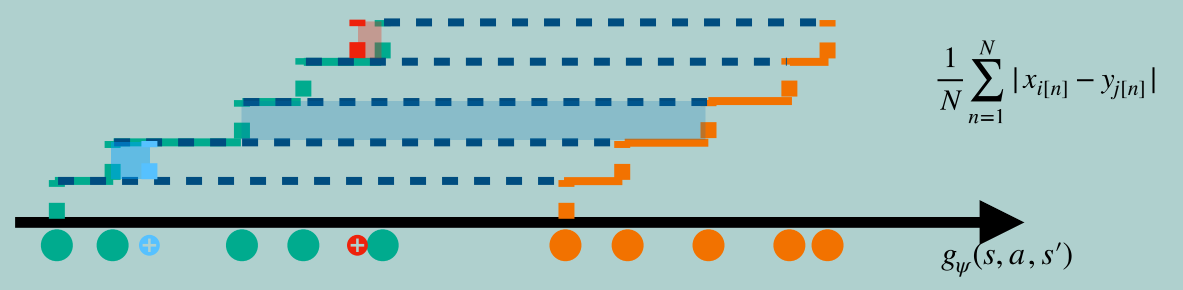
**Method I:**
**Addressing Reward Generalization using Causal Invariance**






**Method II:**
**Addressing Data Efficiency in Imitation Learning using Sliced Optimal Transport**





# Addressing Data Efficiency

## Back to distribution matching

**So far, focus on distribution matching methods based on f-divergences**

$$D_f(P,Q) := \begin{cases} \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right] & P \ll Q \\ +\infty & \text{otherwise} \end{cases}$$

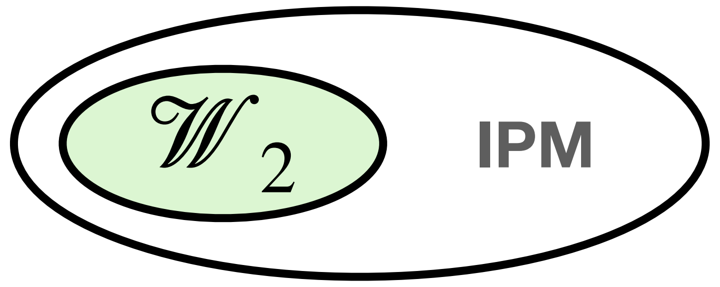
**Two issues**

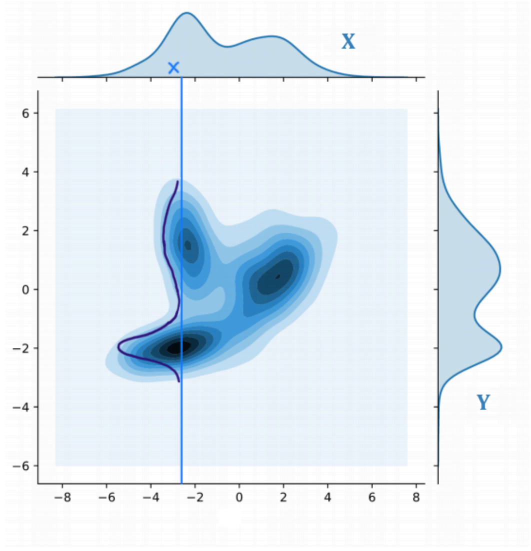
- **Ill-defined for disjoint support**
- **Requires brittle optimization procedures**

**Alternative: Integral Probability Metrics**

$$v_{\mathcal{F}}(P,Q) := \sup_{f \in \mathcal{F}} |\mathbb{E}_P[f] - \mathbb{E}_Q[f]|$$

**Depending on $\mathcal{F}$, different metrics: MMD, Kantorovich, Dudley**

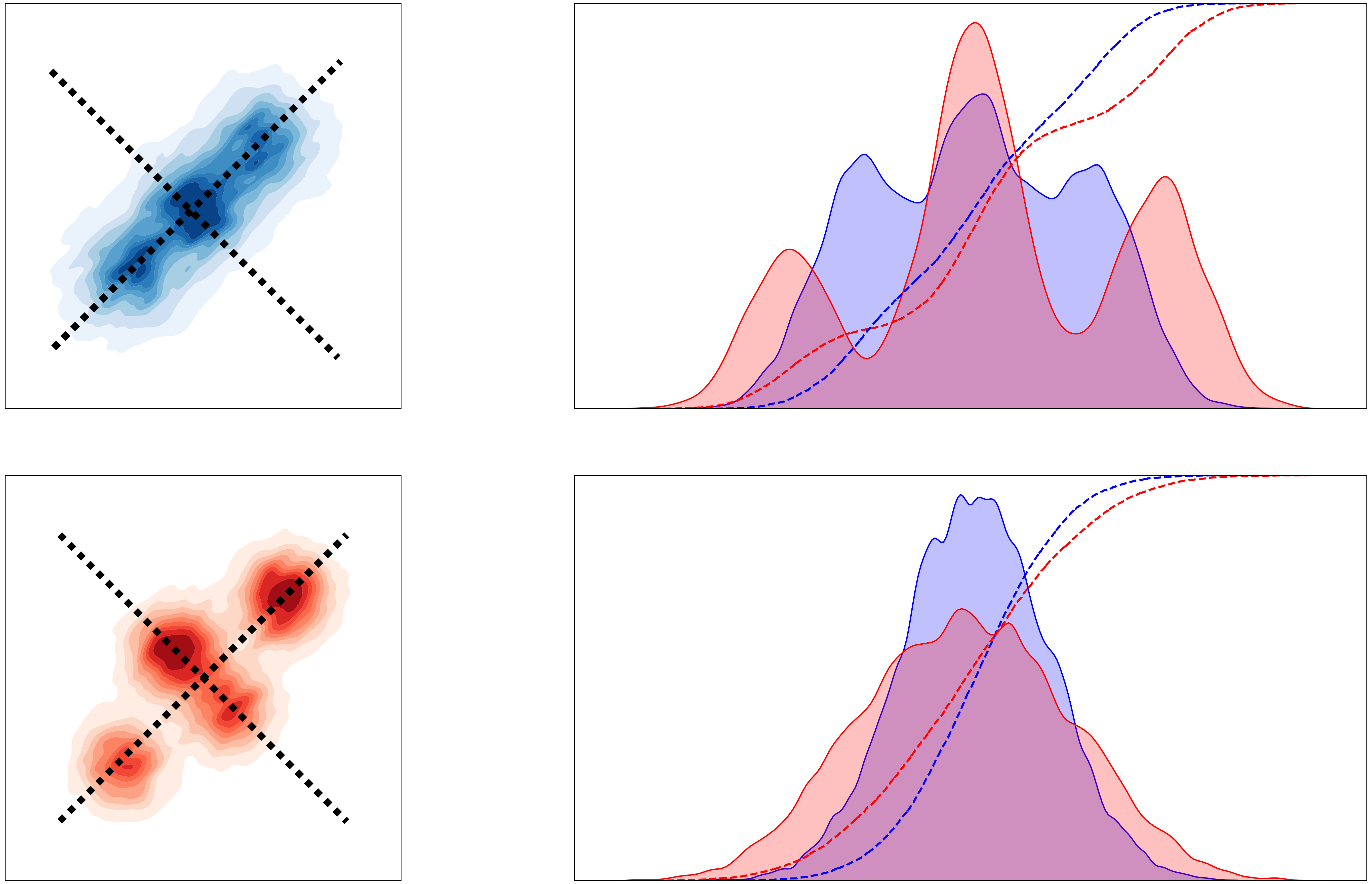


**Leverage advances in computational optimal transport (OT)**

# Exploring computational OT

## Sliced optimal transport



**Leverage closed form of 1-D Wasserstein Distance computation**

$$W_2(\rho, \rho') = \int_0^1 |F_\rho^{-1}(\mu) - F_{\rho'}^{-1}(\mu)|^2 d\mu$$

**Efficient sorting-based computation**

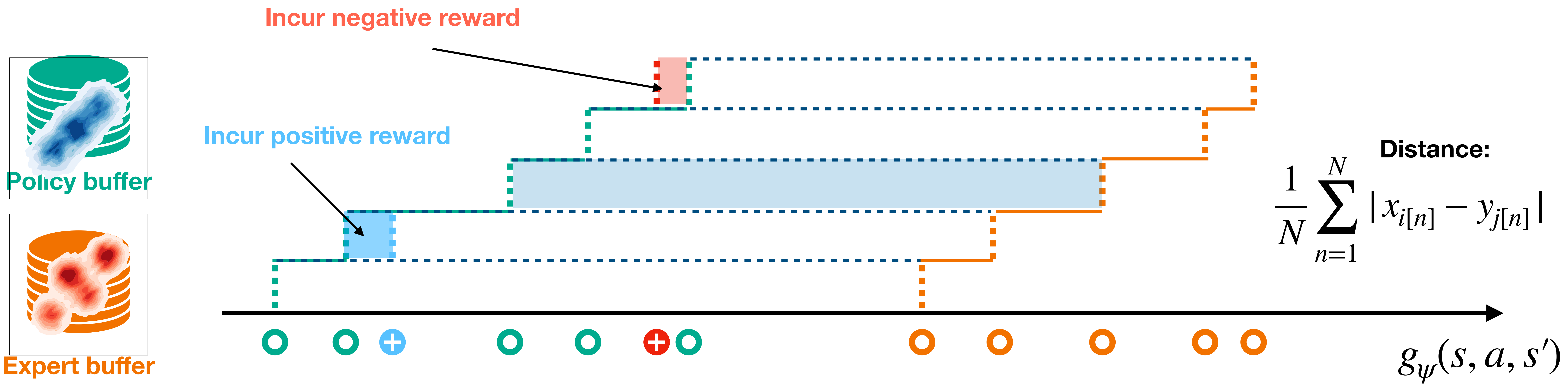
**Sliced Wasserstein Distance**

$$SW_2(\rho, \rho') = \int_{\mathcal{S}^{d-1}} W_2(\psi_\# \rho, \psi_\# \rho') d\psi$$

**Imitation Learning: No access to dynamics**

# Sliced Wasserstein Distances

## Proxy reward structure




**Distance:**

$$\frac{1}{N}\sum_{n=1}^{N} |x_{i[n]} - y_{j[n]}|$$

**Atom replacement reward:** $r_d(x_t; \hat{\rho}_E, \hat{\rho}_\pi) = d(\hat{\rho}_E, \hat{\rho}_\pi) - d(\hat{\rho}_E, \mathrm{rpl}(\hat{\rho}_\pi, x_t))$

**Procedure**

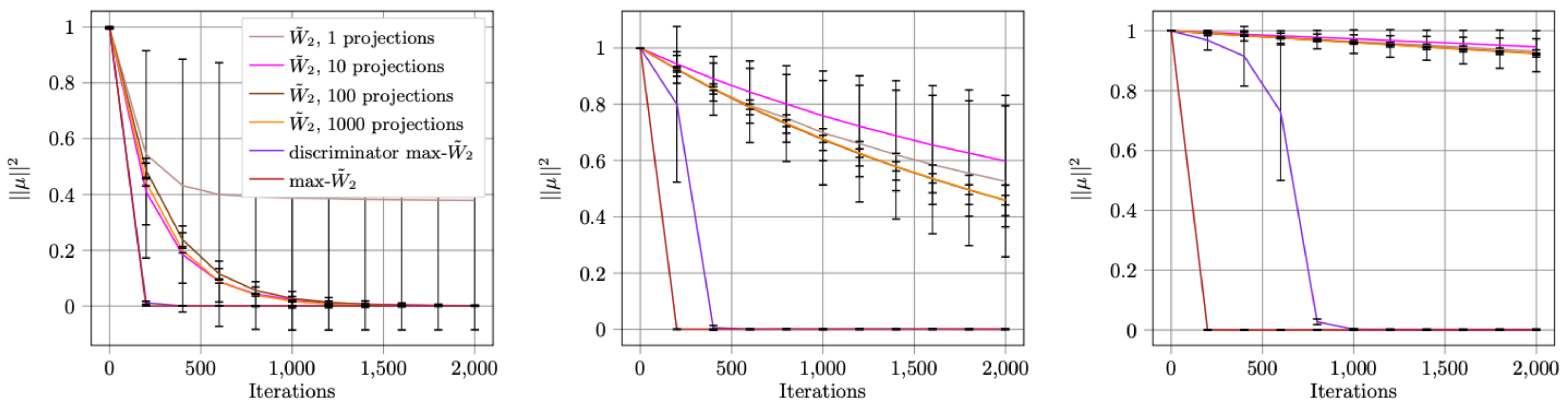
1. Project buffer samples
2. Sample policy and project atom
3. Find closest atom in ranking and measure replacement distance
4. Use measure replacement distance as reward

**Policy optimizing reward reduces the 1-D Wasserstein distance**

# Choice of projections

## Generalized Sliced Wasserstein (GSW) Distances

**Linear projections scale poorly with state space dimensionality**




(a) $d = 10$ (b) $d = 100$ (c) $d = 1000$

*Deshpande et al., 2019*

**Choose most informative (nonlinear) projection**

$$SW_2(\rho, \rho') = \int_{\mathcal{S}^{d-1}} W_2(\psi_\# \rho, \psi_\# \rho') d\psi$$

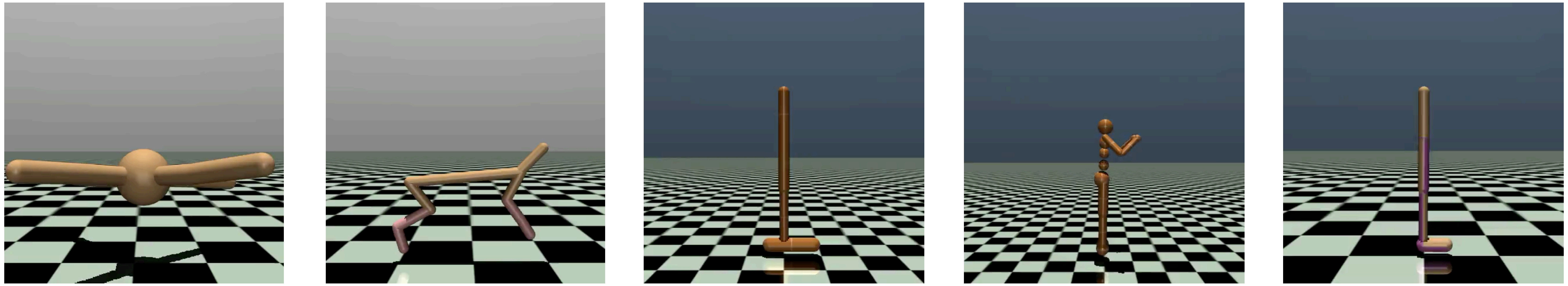
$$MSW_2(\rho, \rho') = \sup_{\psi \in \mathcal{S}^{d-1}} W_2(\psi_\# \rho, \psi_\# \rho')$$

$$MGSW_2(\rho, \rho') = \sup_{\psi \in \Psi} W_2(g_*^{(\psi)} \rho, g_*^{(\psi)} \rho')$$

$$\inf_{\pi \in \Pi} MGSW_2(\hat{\rho}_E, \hat{\rho}_\pi) = \inf_{\pi \in \Pi} \sup_{\psi \in \Psi} W_2(g_*^{(\psi)} \hat{\rho}_E, g_*^{(\psi)} \hat{\rho}_\pi)$$

# Imitation Learning using GSW

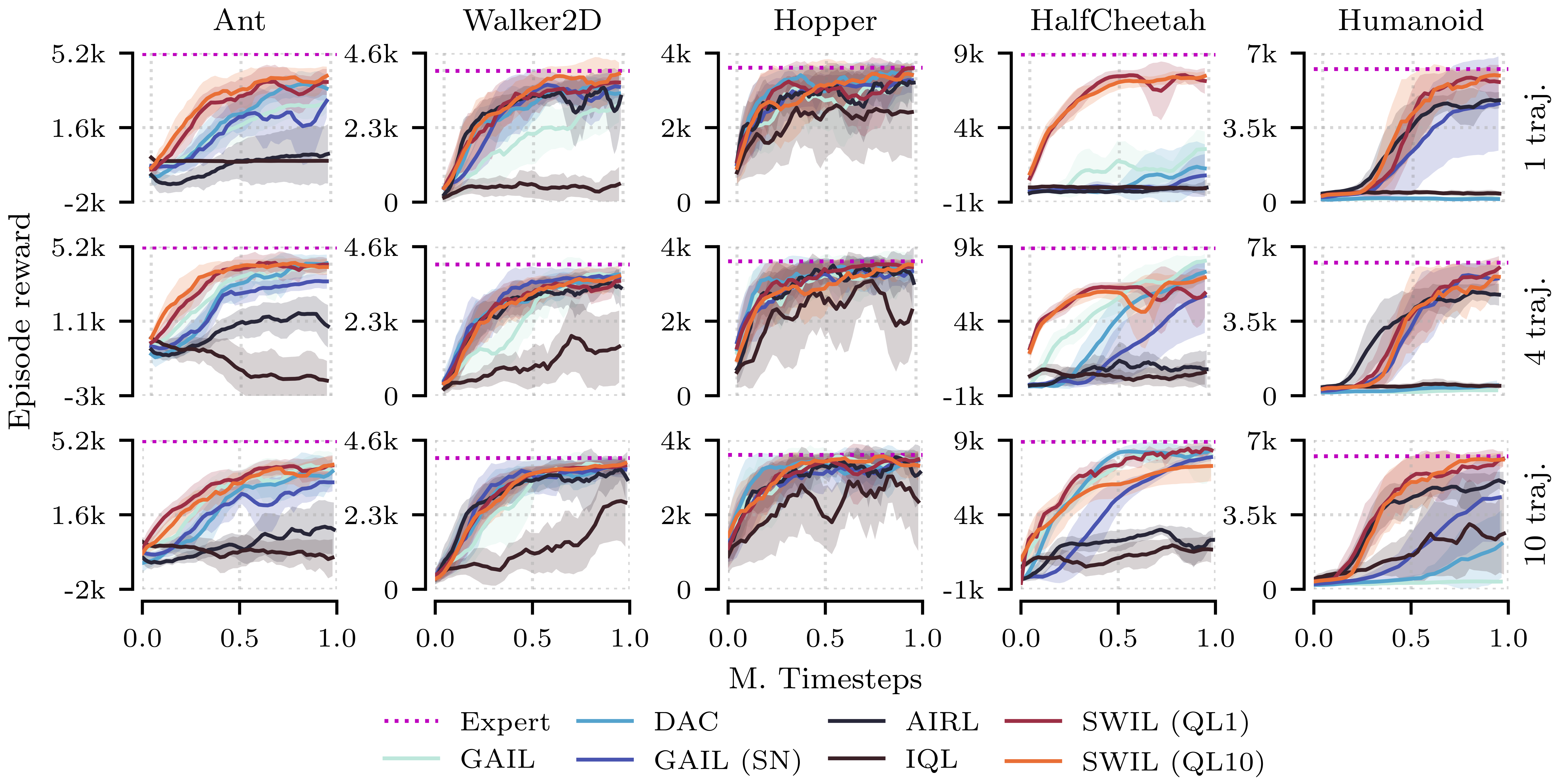
## Recovers expert performance from a single trajectory



| Environment | Ant | HalfCheetah | Hopper | Humanoid | Walker2D |
|---|---|---|---|---|---|
| Expert | 5159.77 | 8900.85 | 3607.17 | 6249.60 | 4063.81 |
| BC | 978.76±10.23 | 221.12±48.71 | 504.56±154.59 | 347.26±32.13 | 299.42±15.63 |
| DAC | 3553.38±1580.64 | 3292.43±1365.91 | 3400.87±247.88 | 169.06±40.84 | 3433.62±362.73 |
| GAIL (SAC) | 4261.13±415.56 | 2569.26±1179.47 | 2866.05±760.34 | 147.39±59.98 | 2774.17±563.45 |
| GAIL (SAC-SN) | 2956.94±697.31 | 804.24±1441.09 | 3263.01±540.13 | 4643.93±2203.33 | 3296.62±417.84 |
| AIRL (SAC) | 361.58±1424.61 | -9.47±382.93 | 3197.09±614.25 | 4778.57±404.82 | 3257.18±867.42 |
| PWIL | 1289.23±697.31 | 1089.76±923.19 | 2890.14±430.12 | 5252.23±156.22 | 2452.17±856.22 |
| GAIL (PPO-SN) | 2143.35±556.98 | 2863.42±1155.38 | 2859.98±1114.94 | 425.11±125.65 | 2648.65±1128.32 |
| SWIL | **4338.65**±555.83 | **7481.79**±769.22 | **3585.28**±66.63 | **5952.09**±315.92 | **3936.35**±365.69 |

# Imitation Learning using GSW

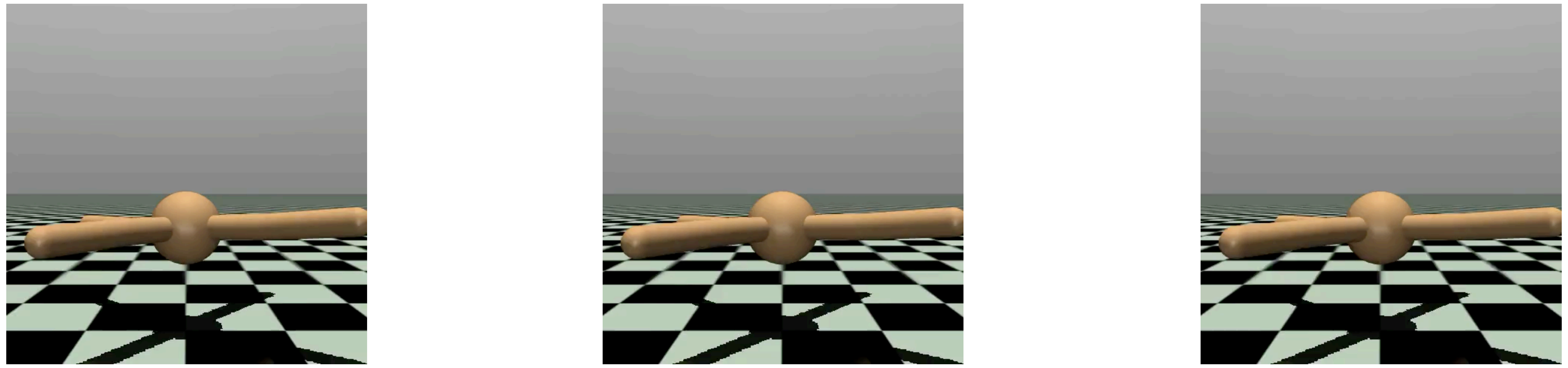
## Varying number of trajectories

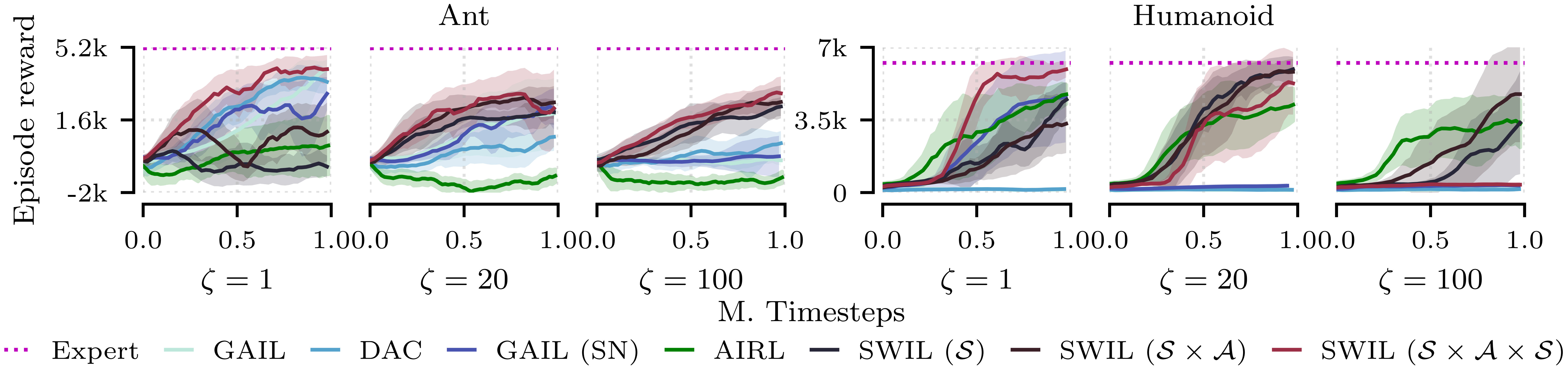


# Imitation Learning using GSW

## Evaluation: data sparse setting

$\zeta = 1$, Return: 4521.4 $\zeta = 20$, Return: 4167.3 $\zeta = 100$, Return: 3810.8

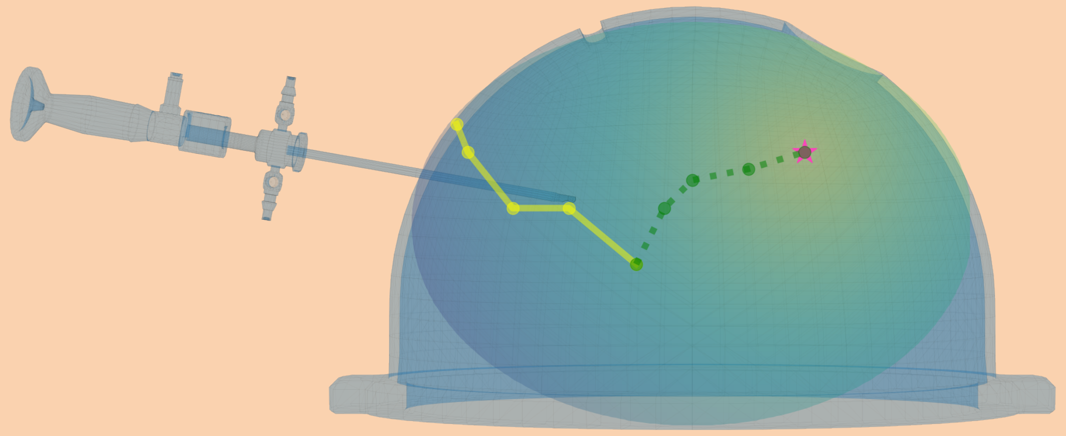


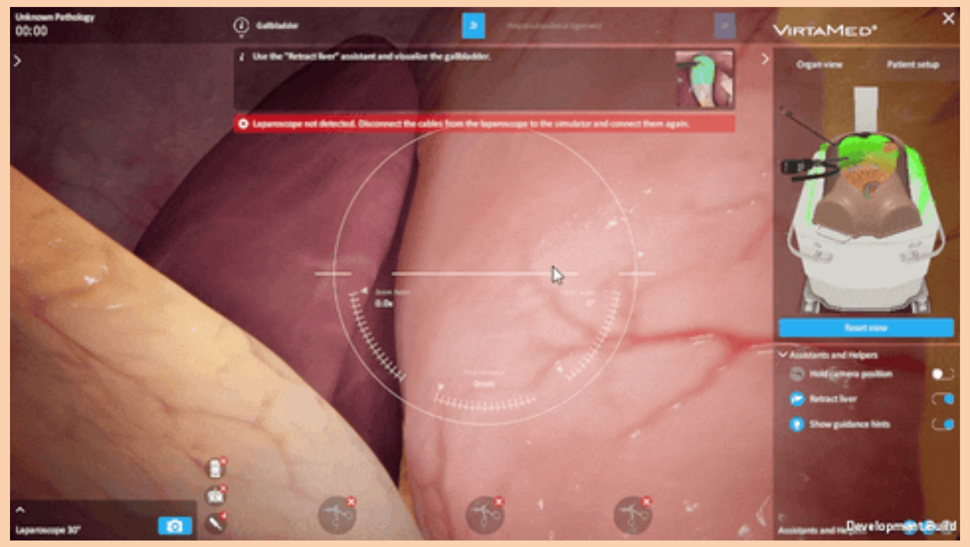



**Performance deteriorates more gracefully under trajectory sparsity**

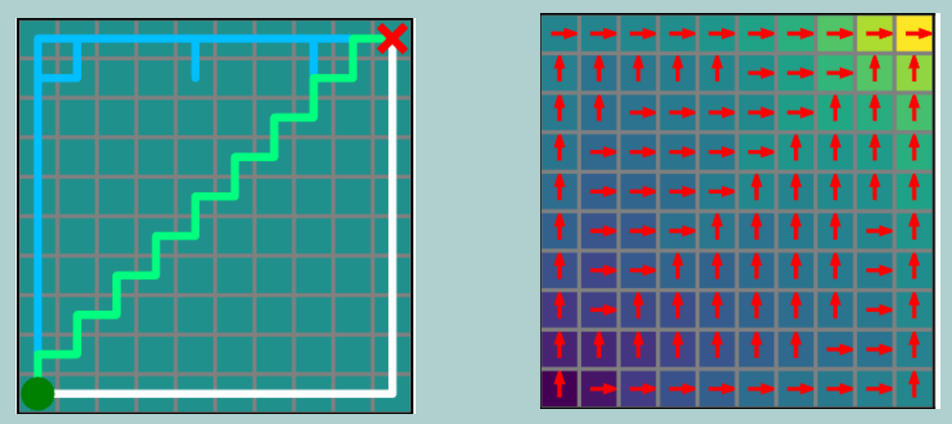
# Conclusion

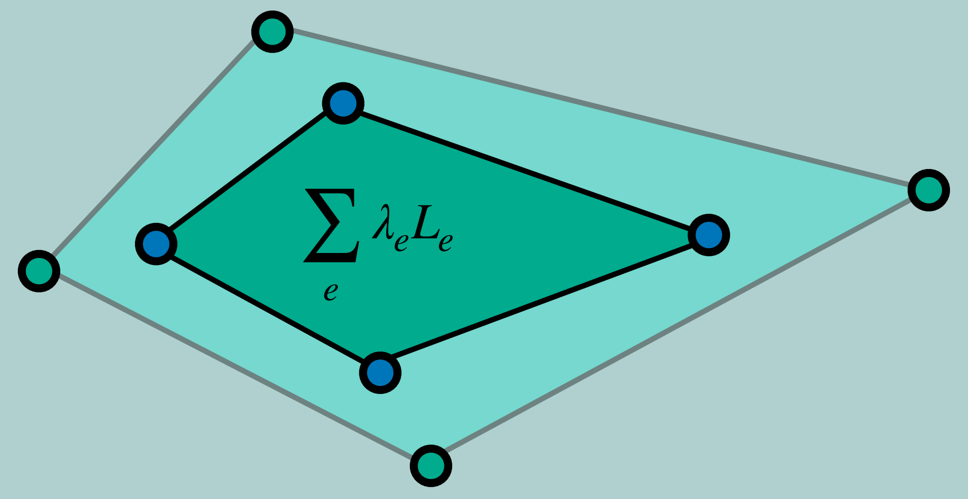
**Application:**
**Algorithmic RL pipeline in Surgical Digital Twins**



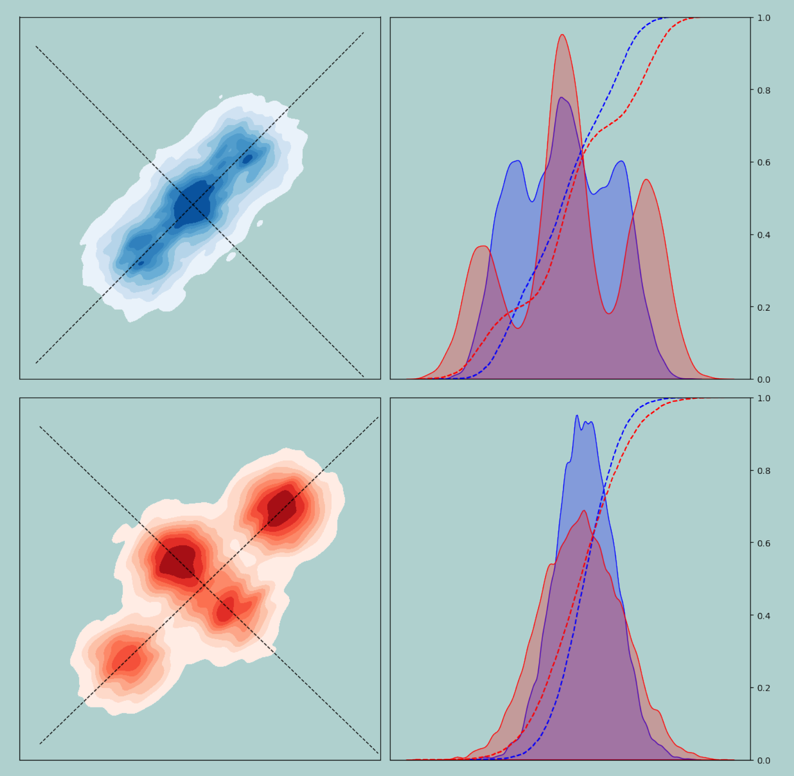


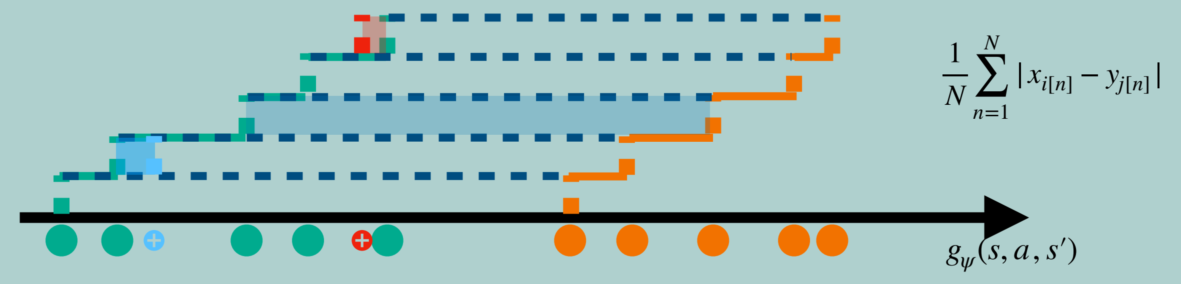
**Method I:**
**Addressing Reward Generalization using Causal Invariance**






**Method II:**
**Addressing Sample Efficiency in Imitation Learning using Sliced Optimal Transport**







Ivan Ovinnikov*, Ami Beuret*,
Flavia Cavaliere, Joachim Buhmann

**FASTRL and beyond: a reinforcement learning pipeline for fundamentals of arthroscopic surgery training**
IJCARS, https://doi.org/10.1007/s11548-024-03116-z


Ivan Ovinnikov, Eugene Bykovets,
Joachim Buhmann

**Learning Causally Invariant Reward Functions from Diverse Demonstrations**
TMLR, in review


Ivan Ovinnikov, Alexander Terenin,
Joachim Buhmann

**Imitation Learning via Generalized Sliced Wasserstein Distances**
Preprint

# Acknowledgements

Prof. Dr. Joachim Buhmann
Prof. Dr. Andreas Krause
Dr. Raimundo Sierra
Prof. Dr. Markus Gross

Rita Klute

Alexander Terenin
Ami Beuret
Eugene Bykovets
Flavia Cavaliere
Imant Daunhawer
Luis Haug

Martina Vitz
Basil Fierz
Francesco Maggiolino

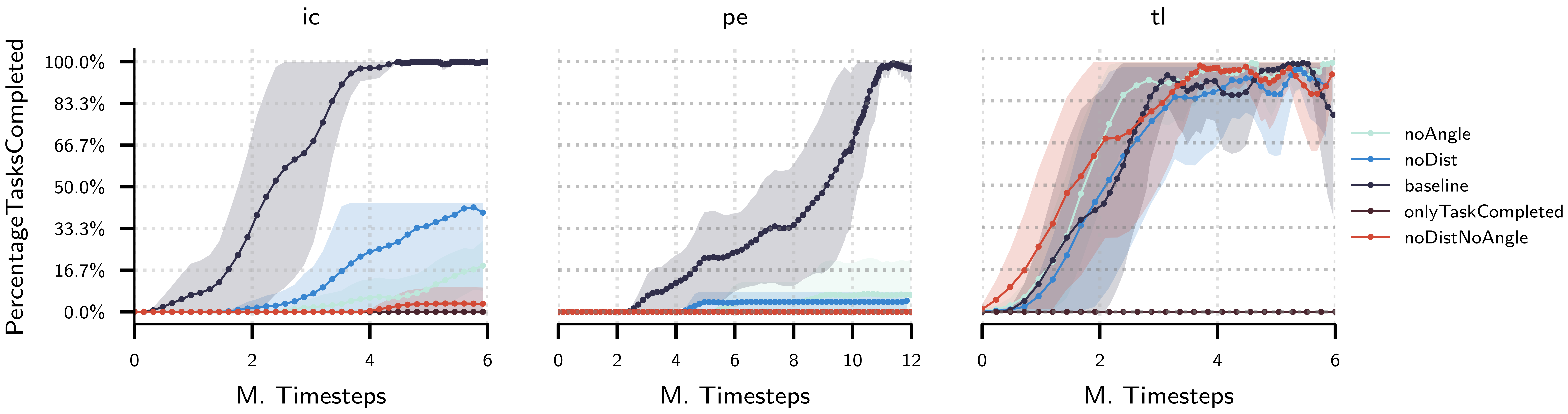
Institute of Machine Learning

Friends and Family

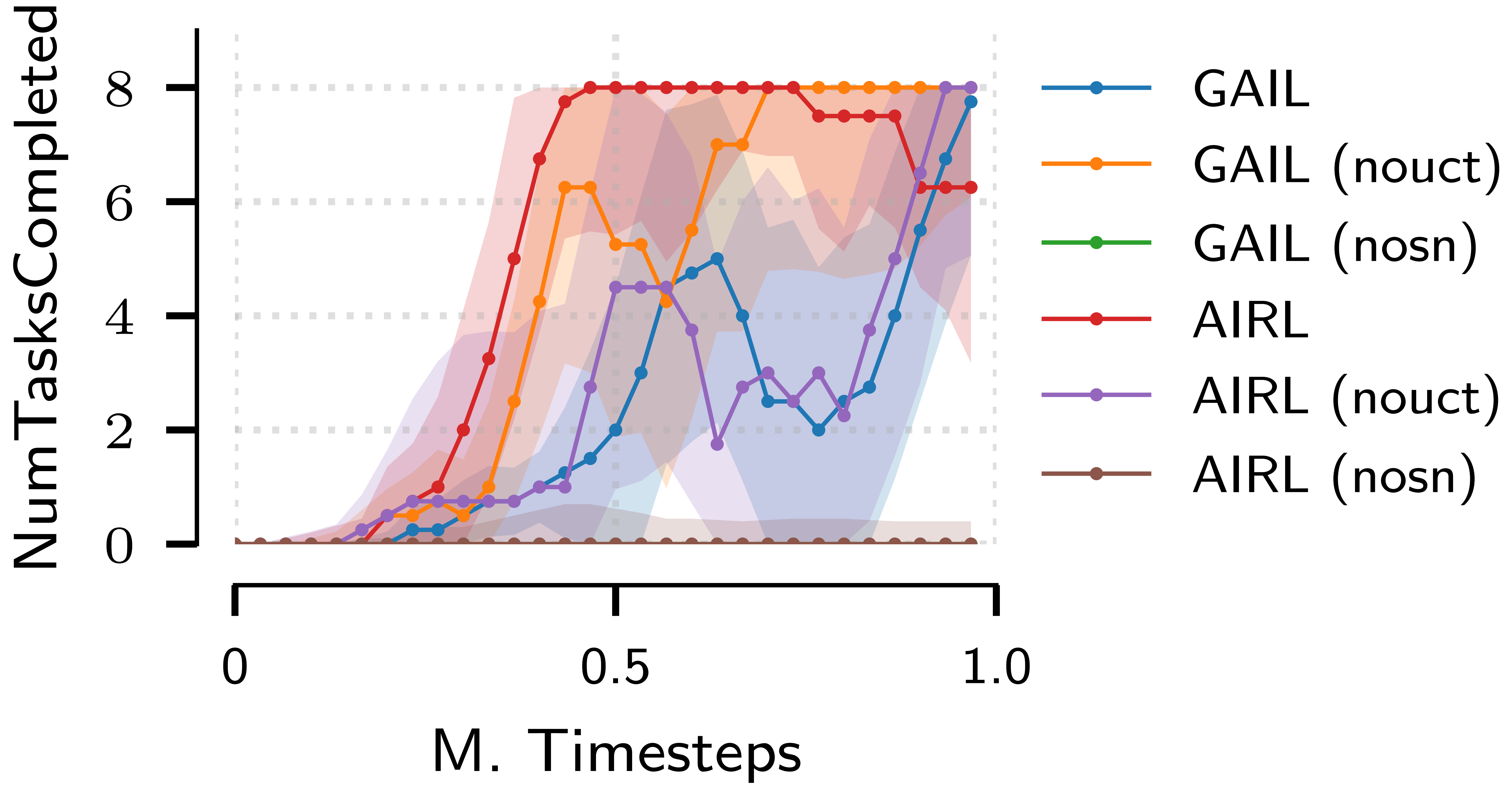
# Questions

# Appendix

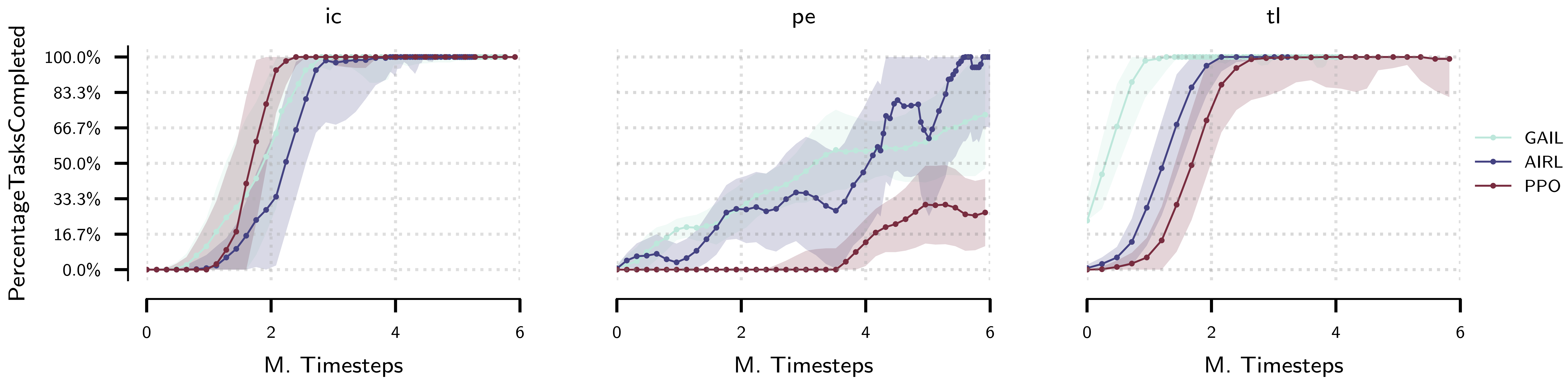
# FASTRL Ablations



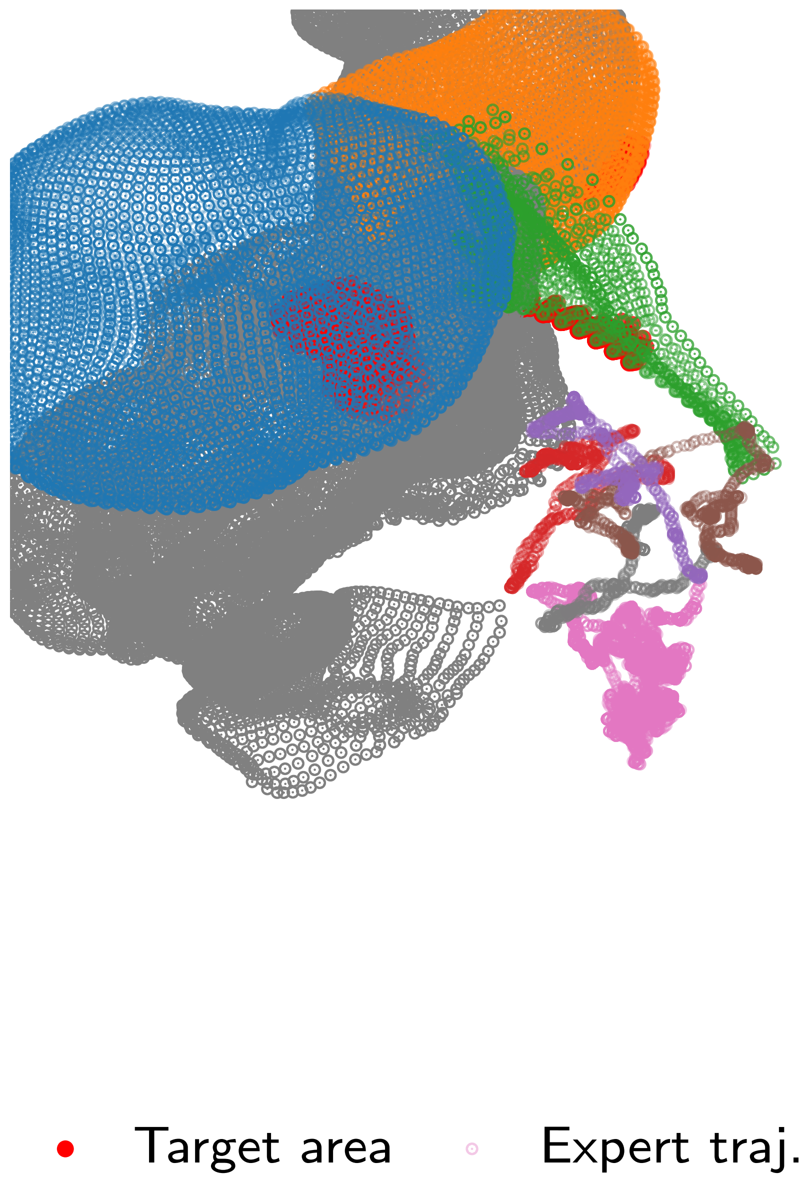
# FASTRL : TraceLines SAC

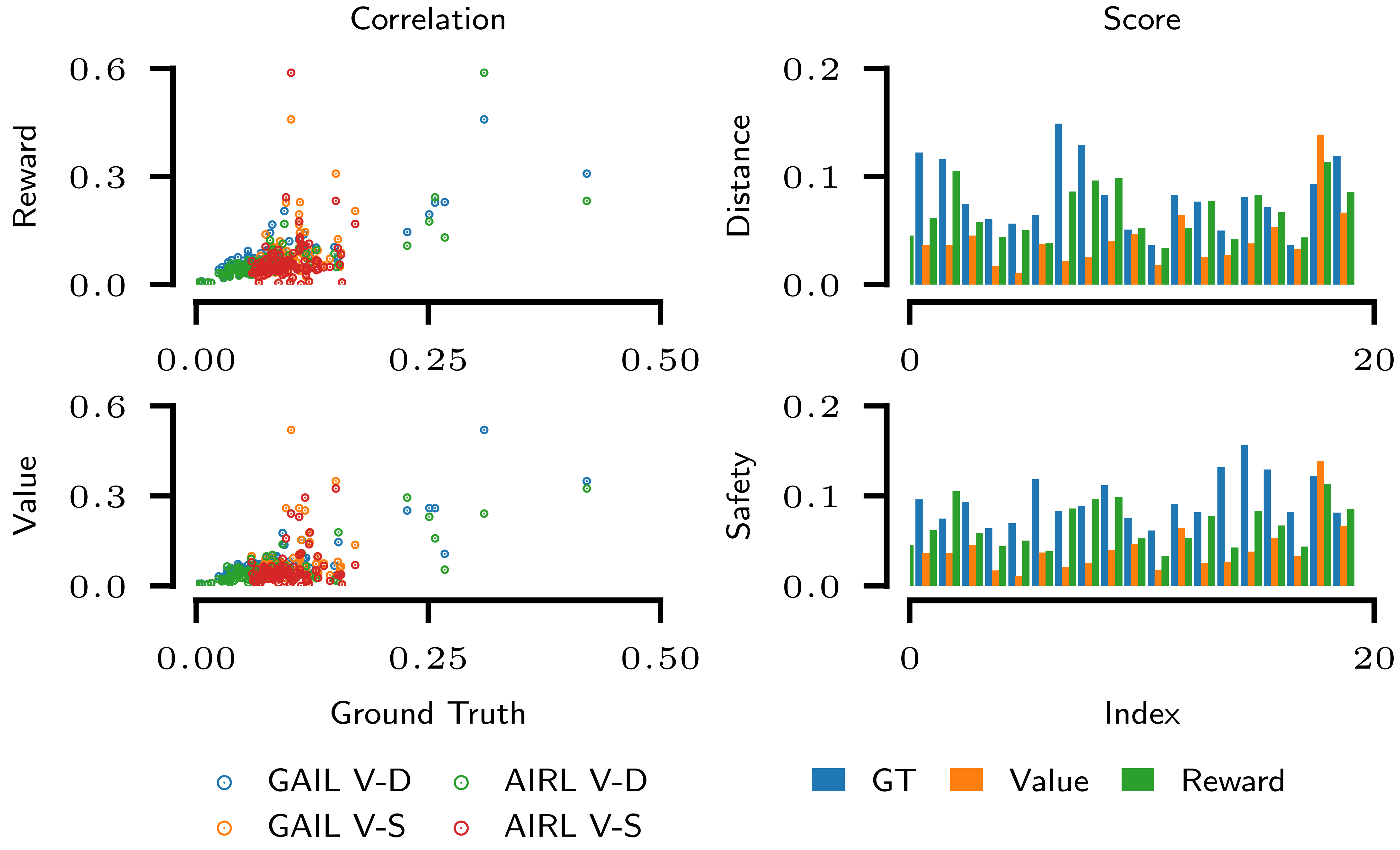


# FASTRL : algorithm comparison



# Laparos: Davos Dataset Evaluation

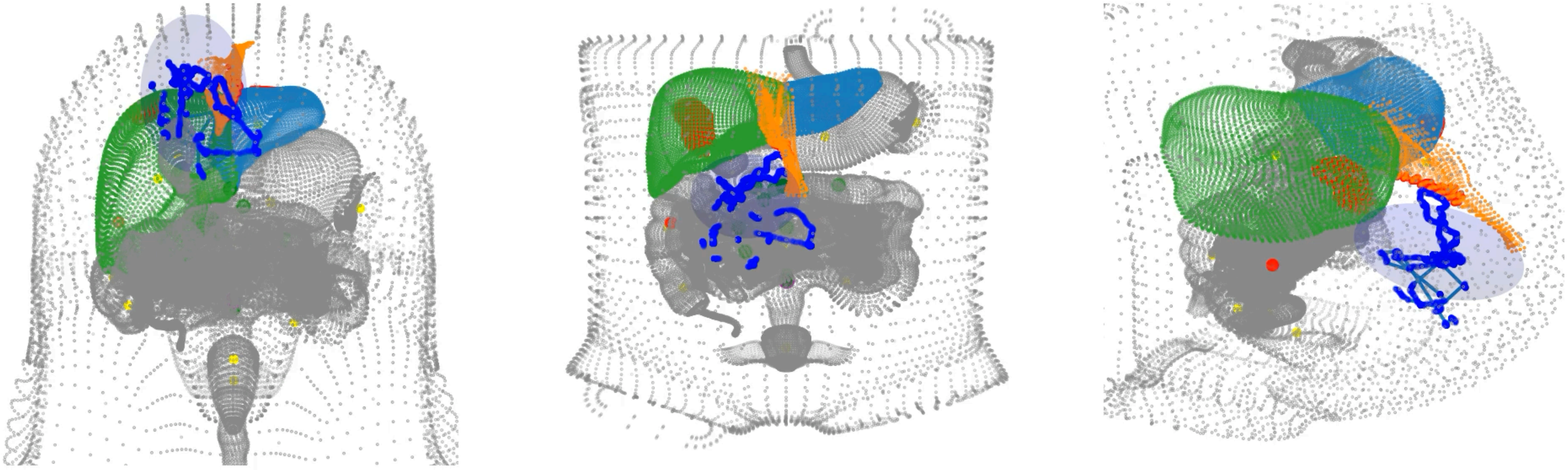



| Model | Reward-Dst. (p) | Reward-Sft. (p) | Value-Dst. (p) | Value-Sft. (p) |
|---|---|---|---|---|
| GAIL | 0.73±0.03 (0.00) | 0.25±0.02 (0.01) | 0.78±0.01 (0.00) | 0.24±0.01 (0.02) |
| AIRL | -0.82±0.01 (0.00) | -0.25±0.01 (0.01) | 0.80±0.00 (0.00) | 0.24±0.00 (0.02) |

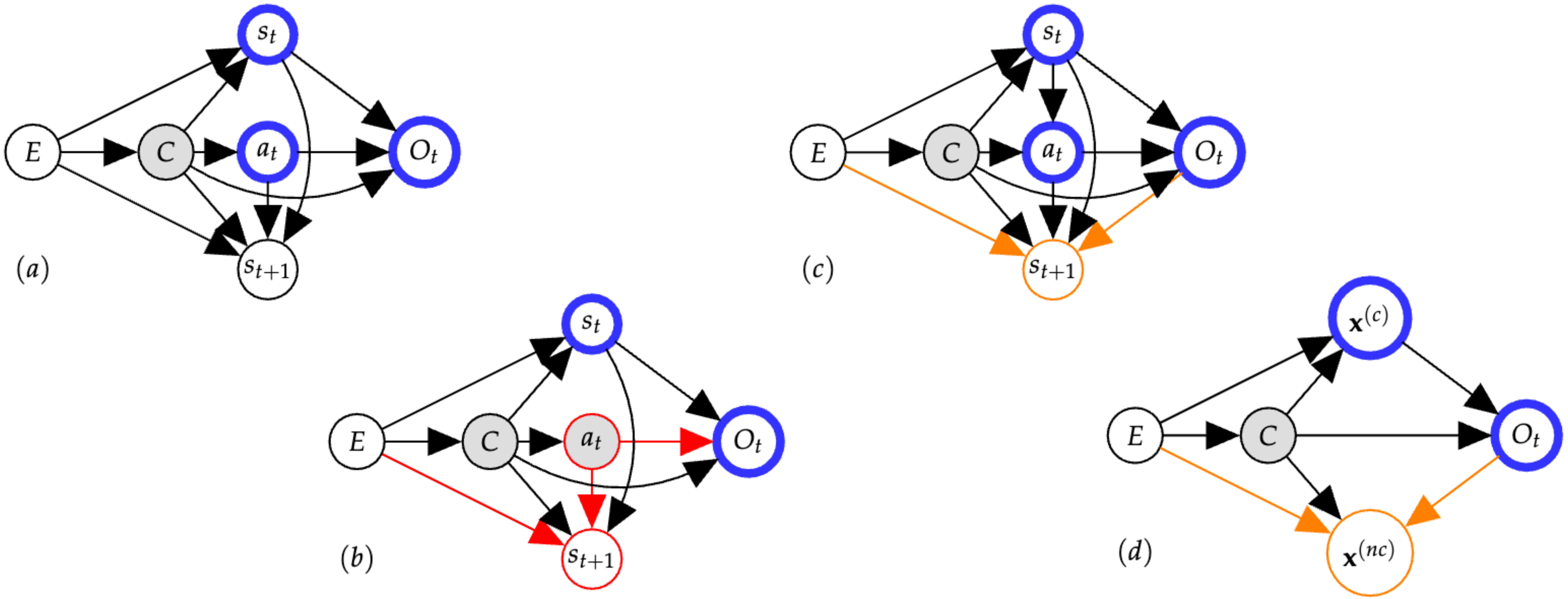
**Table 3**: Correlation coefficients (p-values in brackets) between laparoscopic trajectories evaluated using recovered rewards and two ground truth metrics (*Dst*: total instrument distance and *Sft*: total safety distance).

# SimpleLap

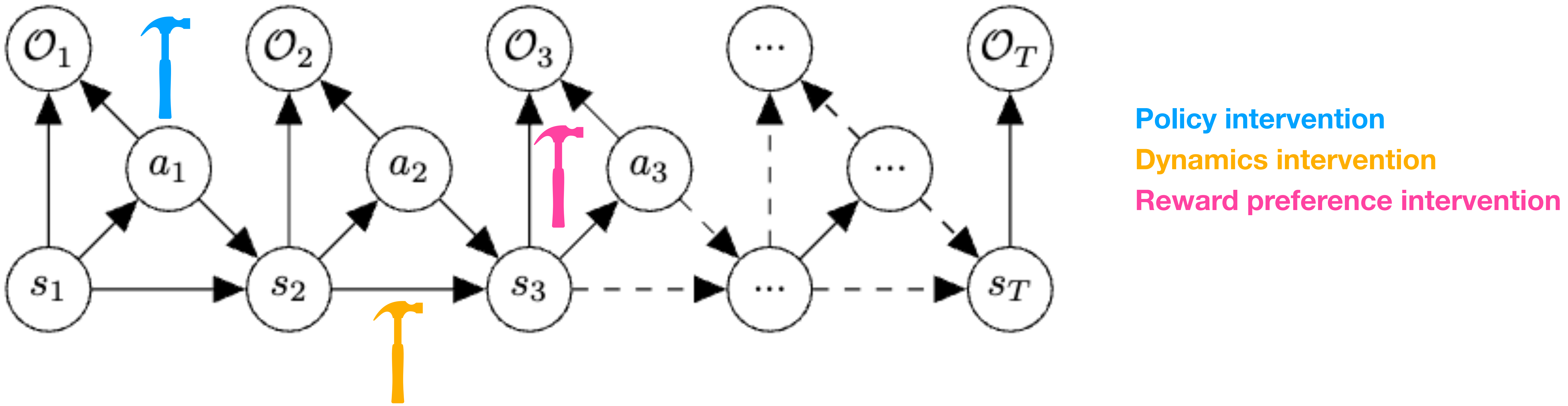


# Transition SCM

## Spurious correlations



# Trajectory SCM




(a) Trajectory Model

$$p(\tau|\mathcal{O}_{1:T}) \propto p(\tau, \mathcal{O}_{1:T}) = p_0(\mathbf{s}_1) \prod_{t=1}^{T} p(\mathcal{O}_t = 1|\mathbf{s}_t, \mathbf{a}_t) p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$$

$$= \left( p_0(\mathbf{s}_1) \prod_{t=1}^{T} p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) \right) \exp\left( \sum_{t=1}^{T} r_\psi(\mathbf{s}_t, \mathbf{a}_t) \right)$$

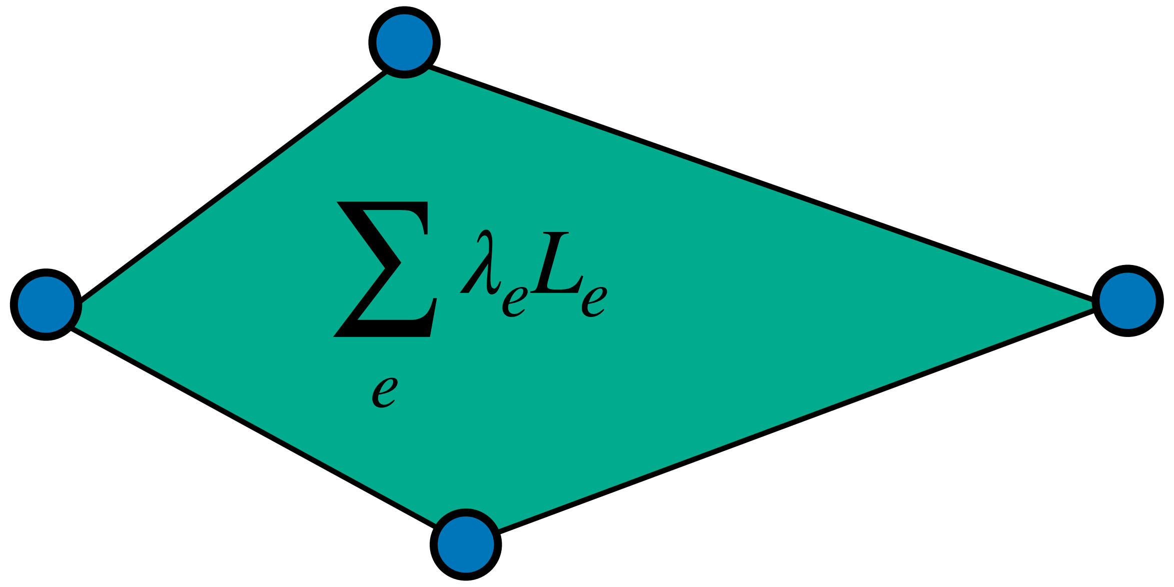
# Reward generalization

## Leveraging diversity

**Goal: recover reward functions which provide meaningful training signal to policies trained across a variety of dynamics**

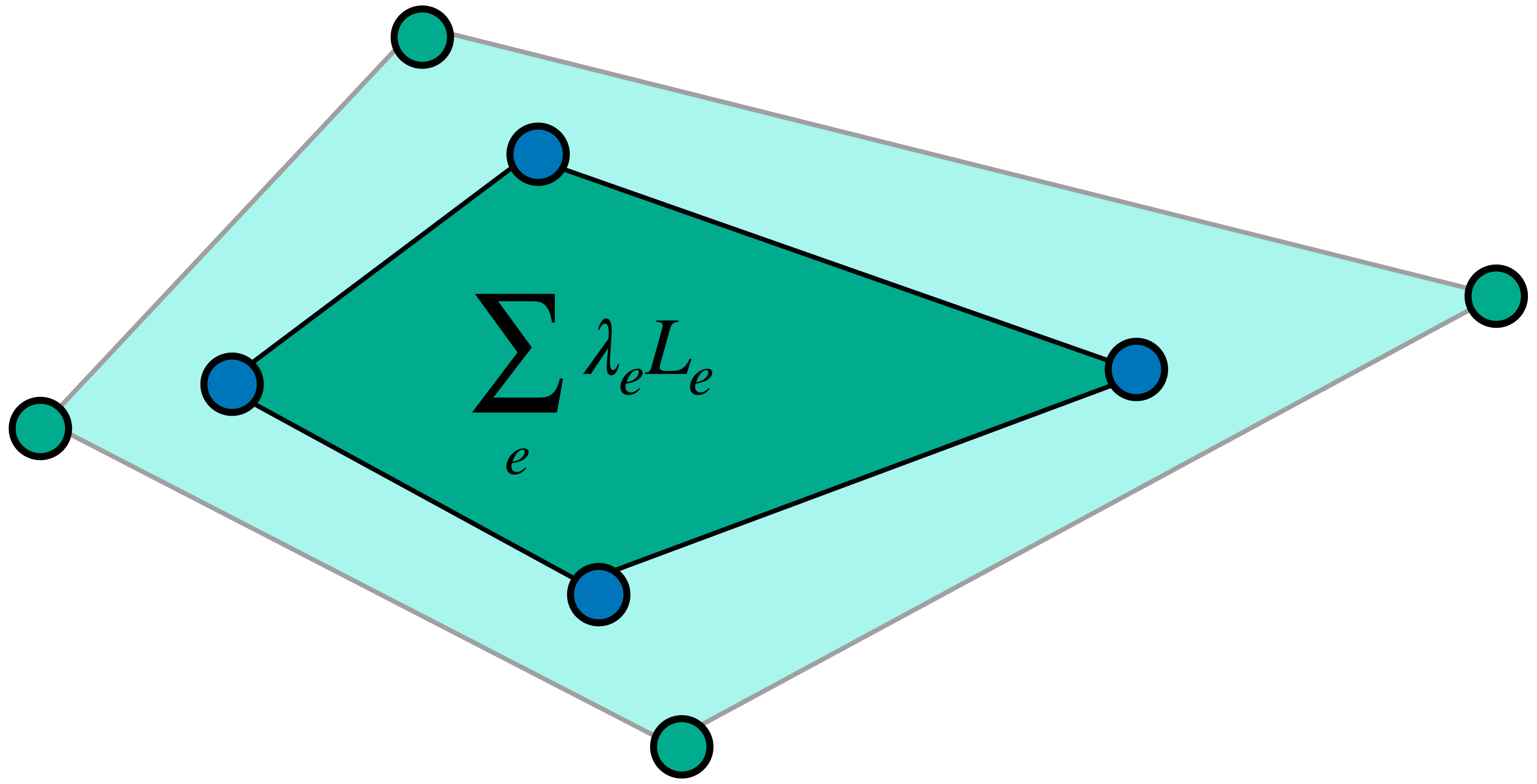
### Distributionally robust optimization

$$\min_{f} \max_{e \in \mathscr{E}_{tr}} \mathbb{E}_{\xi \sim \mathscr{D}_e}[\mathscr{L}_e(f, \xi)]$$



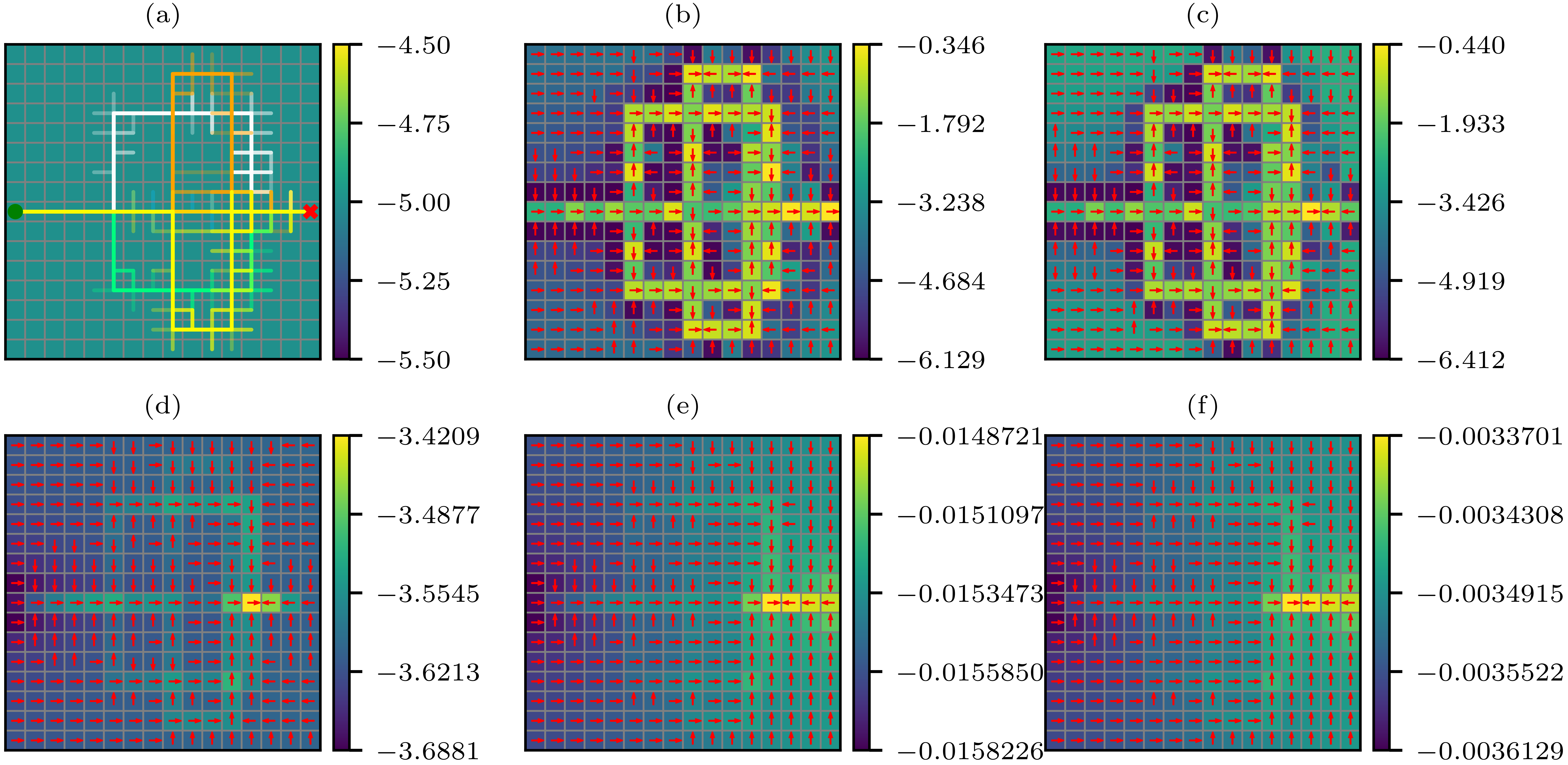

### Causal invariance

$$\min_{f} \mathbb{E}_{\xi \sim \mathscr{D}_e}[\mathscr{L}_e(f, \xi)] \quad \forall e \in \mathscr{E}_{tr}$$

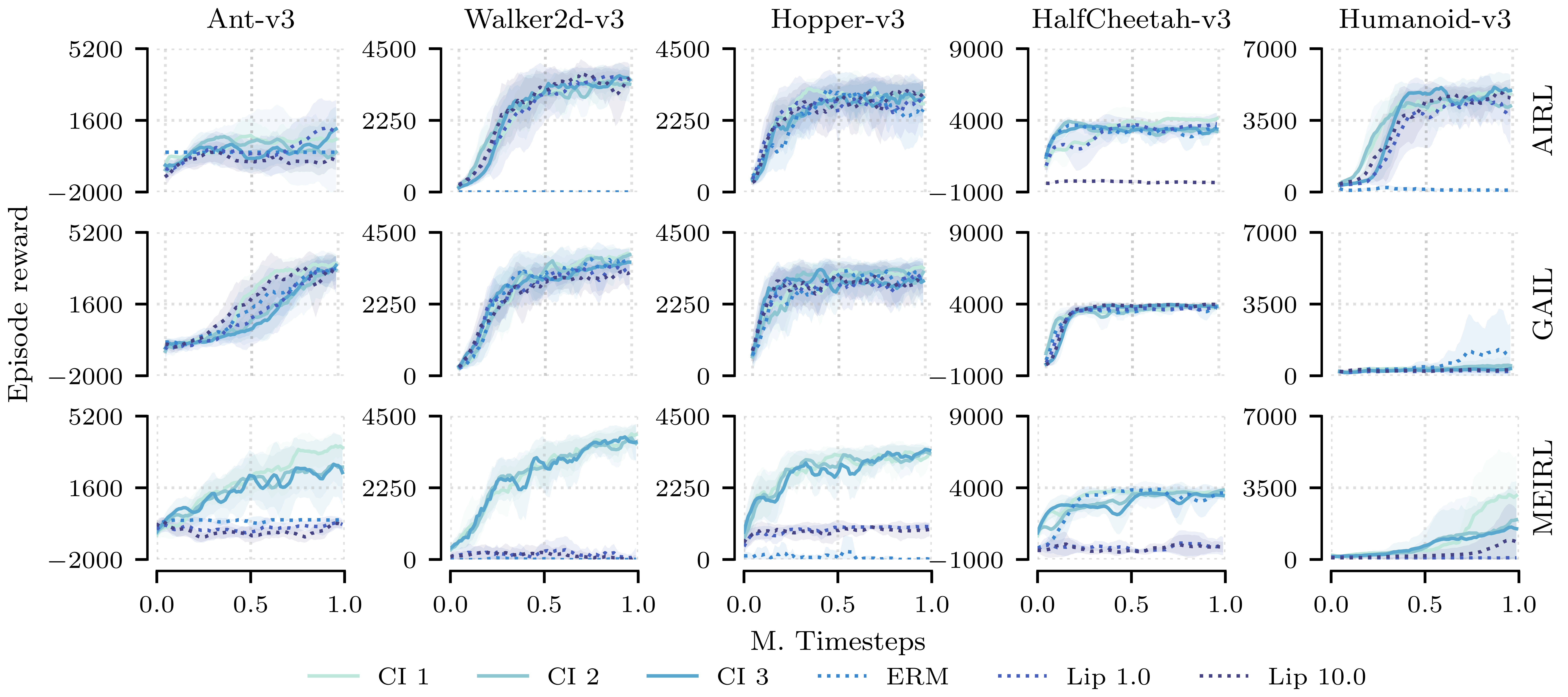



Training setting  Generalization setting

# CI-IRL: gridworld



# CI-IRL: adversarial training



# Table: body mass

| Environment | Ant-v3 | Walker2D-v3 | Hopper-v3 | HalfCheetah-v3 | Humanoid-v3 |
|---|---|---|---|---|---|
| Expert | $3168.49_{\pm 1715.68}$ | $3565.33_{\pm 527.40}$ | $3119.54_{\pm 524.36}$ | $4340.61_{\pm 2020.14}$ | $4774.17_{\pm 2063.52}$ |
| AIRL (ERM) | $580.78_{\pm 1048.73}$ | $-3.29_{\pm 0.71}$ | $77.64_{\pm 88.27}$ | $2046.29_{\pm 460.98}$ | $4451.74_{\pm 1759.31}$ |
| AIRL (Lip) | $1194.04_{\pm 1583.08}$ | $3388.48_{\pm 1586.45}$ | $3382.91_{\pm 234.02}$ | $4388.94_{\pm 726.69}$ | $1788.85_{\pm 1643.00}$ |
| AIRL (CI) | $1880.42_{\pm 935.15}$ | $4162.70_{\pm 517.13}$ | $3334.91_{\pm 221.80}$ | $4477.97_{\pm 532.72}$ | $5107.54_{\pm 119.31}$ |
| GAIL (ERM) | $-746.31_{\pm 468.03}$ | $328.44_{\pm 66.02}$ | $1637.50_{\pm 1419.59}$ | $886.25_{\pm 404.82}$ | $122.62_{\pm 71.53}$ |
| GAIL (Lip) | $220.97_{\pm 524.83}$ | $553.36_{\pm 277.24}$ | $1832.39_{\pm 832.32}$ | $1403.77_{\pm 1282.75}$ | $77.24_{\pm 4.66}$ |
| GAIL (CI) | $230.43_{\pm 565.68}$ | $1172.57_{\pm 539.86}$ | $2636.65_{\pm 1114.94}$ | $2365.55_{\pm 1679.64}$ | $549.63_{\pm 1692.08}$ |
| MEIRL (ERM) | $-66.66_{\pm 112.03}$ | $169.11_{\pm 344.87}$ | $3.22_{\pm 0.22}$ | $-177.39_{\pm 211.43}$ | $55.99_{\pm 3.45}$ |
| MEIRL (Lip) | $-365.41_{\pm 143.70}$ | $917.14_{\pm 132.05}$ | $1045.40_{\pm 54.76}$ | $-335.10_{\pm 84.66}$ | $1001.49_{\pm 1889.60}$ |
| MEIRL (CI) | $153.43_{\pm 1134.46}$ | $2520.24_{\pm 994.27}$ | $2351.07_{\pm 679.37}$ | $1371.59_{\pm 1469.01}$ | $3099.51_{\pm 2411.21}$ |

Table 4.1: **Policy rollout results using ground truth reward for perturbed MuJoCo environments** after being trained for 1M timesteps using the rewards recovered from the different discriminators in section 4.3.2. Here, the *body mass* parameter is perturbed with a noise magnitude of $\varepsilon = 0.2$. The results are averaged over 10 rollouts and obtained by training the model using five different random seeds.

# Table: actuator control range

| Environment | Ant | Walker2D | Hopper | HalfCheetah | Humanoid |
|---|---|---|---|---|---|
| Expert | $3168.49_{\pm 1715.68}$ | $3565.33_{\pm 527.40}$ | $3119.54_{\pm 524.36}$ | $4340.61_{\pm 2020.14}$ | $4774.17_{\pm 2063.52}$ |
| AIRL (ERM) | $1279.87_{\pm 1281.66}$ | $-0.41_{\pm 3.07}$ | $37.62_{\pm 62.71}$ | $2536.90_{\pm 212.27}$ | $357.19_{\pm 135.66}$ |
| AIRL (Lip) | $809.60_{\pm 1425.76}$ | $2779.73_{\pm 1303.91}$ | $2784.28_{\pm 510.50}$ | $4175.49_{\pm 918.92}$ | $312.24_{\pm 90.05}$ |
| AIRL (CI) | $2166.58_{\pm 1471.70}$ | $3897.58_{\pm 831.24}$ | $2884.34_{\pm 130.60}$ | $4470.17_{\pm 731.81}$ | $2730.60_{\pm 982.13}$ |
| GAIL (ERM) | $-641.93_{\pm 284.65}$ | $1180.57_{\pm 1413.21}$ | $491.27_{\pm 565.63}$ | $1862.67_{\pm 1026.74}$ | $1219.94_{\pm 1784.26}$ |
| GAIL (Lip) | $-92.74_{\pm 363.44}$ | $1672.06_{\pm 1263.95}$ | $2028.07_{\pm 1004.96}$ | $3638.51_{\pm 1164.42}$ | $93.01_{\pm 24.01}$ |
| GAIL (CI) | $2486.00_{\pm 2078.11}$ | $2660.74_{\pm 866.36}$ | $2985.44_{\pm 280.70}$ | $3979.90_{\pm 2494.31}$ | $2986.70_{\pm 2389.78}$ |
| MEIRL (ERM) | $-10.63_{\pm 3.35}$ | $-3.83_{\pm 0.45}$ | $3.17_{\pm 0.27}$ | $-36.66_{\pm 489.57}$ | $58.40_{\pm 0.15}$ |
| MEIRL (Lip) | $-411.65_{\pm 244.20}$ | $832.80_{\pm 311.20}$ | $1073.22_{\pm 138.00}$ | $-261.74_{\pm 164.78}$ | $1064.41_{\pm 2011.91}$ |
| MEIRL (CI) | $133.50_{\pm 969.39}$ | $2286.80_{\pm 1040.59}$ | $2551.59_{\pm 1131.76}$ | $3303.82_{\pm 2332.89}$ | $3058.78_{\pm 2286.41}$ |

Table 4.3: **Policy rollout results using ground truth reward for perturbed MuJoCo environments** after being trained for 1M timesteps using the rewards recovered from the different discriminators in section 4.3.2. Here, the *actuator control range* parameter is perturbed with a noise magnitude of $\varepsilon = 0.2$. The results are averaged over 10 rollouts and obtained by training the model using five different random seeds.

# Table: geometry friction

| Environment | Ant | Walker2d | Hopper | HalfCheetah | Humanoid |
|---|---|---|---|---|---|
| Expert | $3168.49_{\pm 1715.68}$ | $3565.33_{\pm 527.40}$ | $3119.54_{\pm 524.36}$ | $4340.61_{\pm 2020.14}$ | $4774.17_{\pm 2063.52}$ |
| AIRL (ERM) | $603.00_{\pm 909.86}$ | $-3.87_{\pm 0.44}$ | $149.17_{\pm 151.96}$ | $2141.85_{\pm 942.19}$ | $4507.23_{\pm 659.04}$ |
| AIRL (Lip) | $283.73_{\pm 1294.15}$ | $3429.17_{\pm 372.29}$ | $3311.25_{\pm 128.82}$ | $4659.44_{\pm 533.32}$ | $1432.57_{\pm 952.87}$ |
| AIRL (CI) | $1434.01_{\pm 1530.65}$ | $4167.15_{\pm 721.26}$ | $3288.86_{\pm 149.76}$ | $4737.72_{\pm 749.52}$ | $4756.93_{\pm 368.15}$ |
| GAIL (ERM) | $-421.27_{\pm 752.40}$ | $910.12_{\pm 951.80}$ | $939.05_{\pm 959.38}$ | $1563.17_{\pm 1245.51}$ | $871.61_{\pm 1002.21}$ |
| GAIL (Lip) | $-172.69_{\pm 196.92}$ | $1065.02_{\pm 1672.12}$ | $2541.08_{\pm 906.67}$ | $4795.59_{\pm 1018.65}$ | $89.51_{\pm 16.87}$ |
| GAIL (CI) | $1148.32_{\pm 1938.45}$ | $2395.43_{\pm 1282.70}$ | $3068.00_{\pm 459.50}$ | $4037.08_{\pm 983.32}$ | $3385.58_{\pm 2279.28}$ |
| MEIRL (ERM) | $-103.10_{\pm 188.90}$ | $-3.43_{\pm 0.15}$ | $3.36_{\pm 0.19}$ | $191.23_{\pm 630.81}$ | $56.41_{\pm 2.24}$ |
| MEIRL (Lip) | $-252.29_{\pm 184.32}$ | $865.84_{\pm 308.25}$ | $1128.52_{\pm 125.06}$ | $-284.92_{\pm 204.61}$ | $991.70_{\pm 1871.72}$ |
| MEIRL (CI) | $-191.76_{\pm 912.46}$ | $2546.37_{\pm 1073.22}$ | $2425.38_{\pm 1070.78}$ | $1724.44_{\pm 2057.07}$ | $2155.55_{\pm 2281.06}$ |

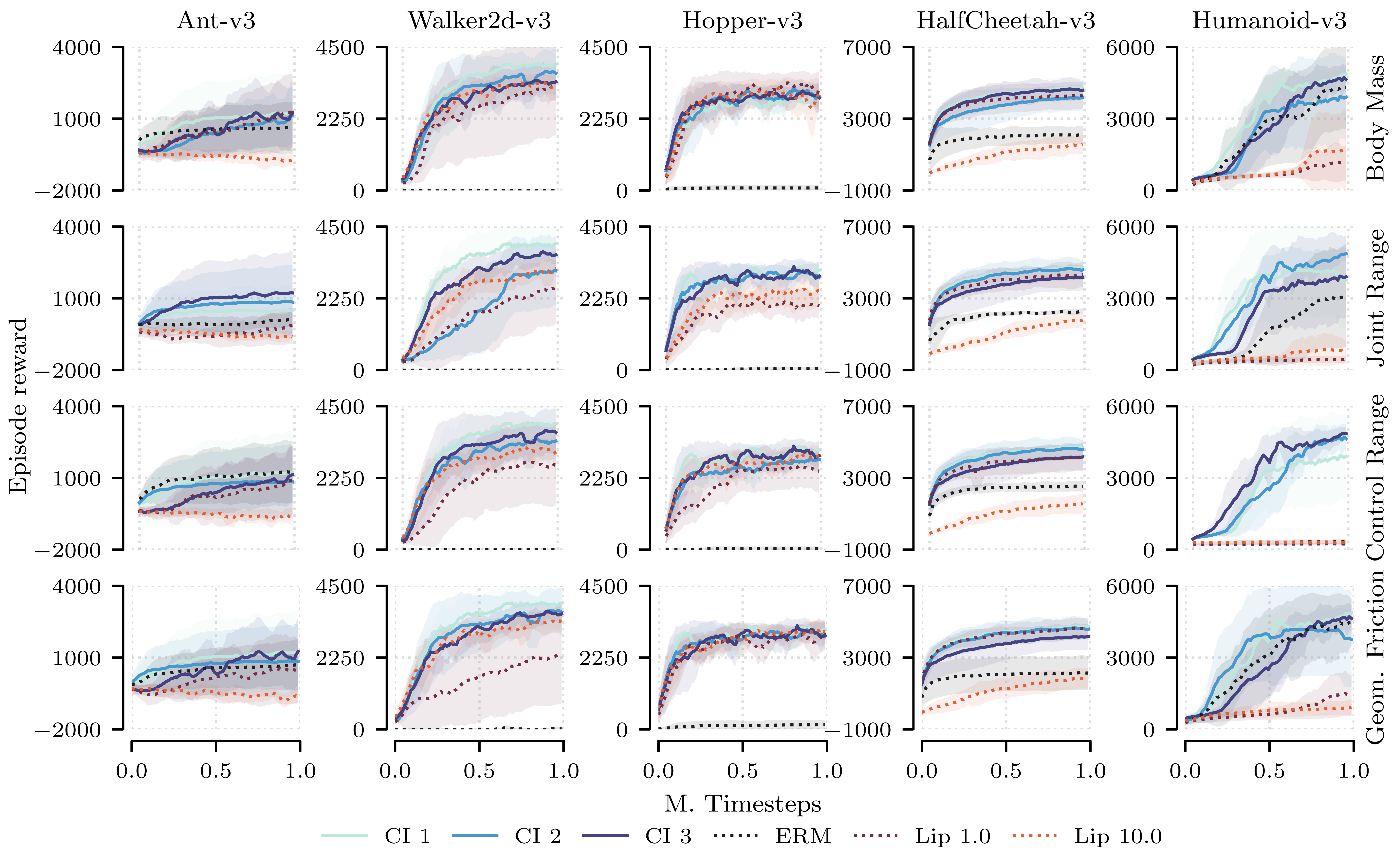
Table 4.4: **Policy rollout results using ground truth reward for perturbed MuJoCo environments** after being trained for 1M timesteps using the rewards recovered from the different discriminators in section 4.3.2. Here, the *geometry friction* parameter is perturbed with a noise magnitude of $\varepsilon = 0.2$. The results are averaged over 10 rollouts and obtained by training the model using five different random seeds.

# Table: joint range

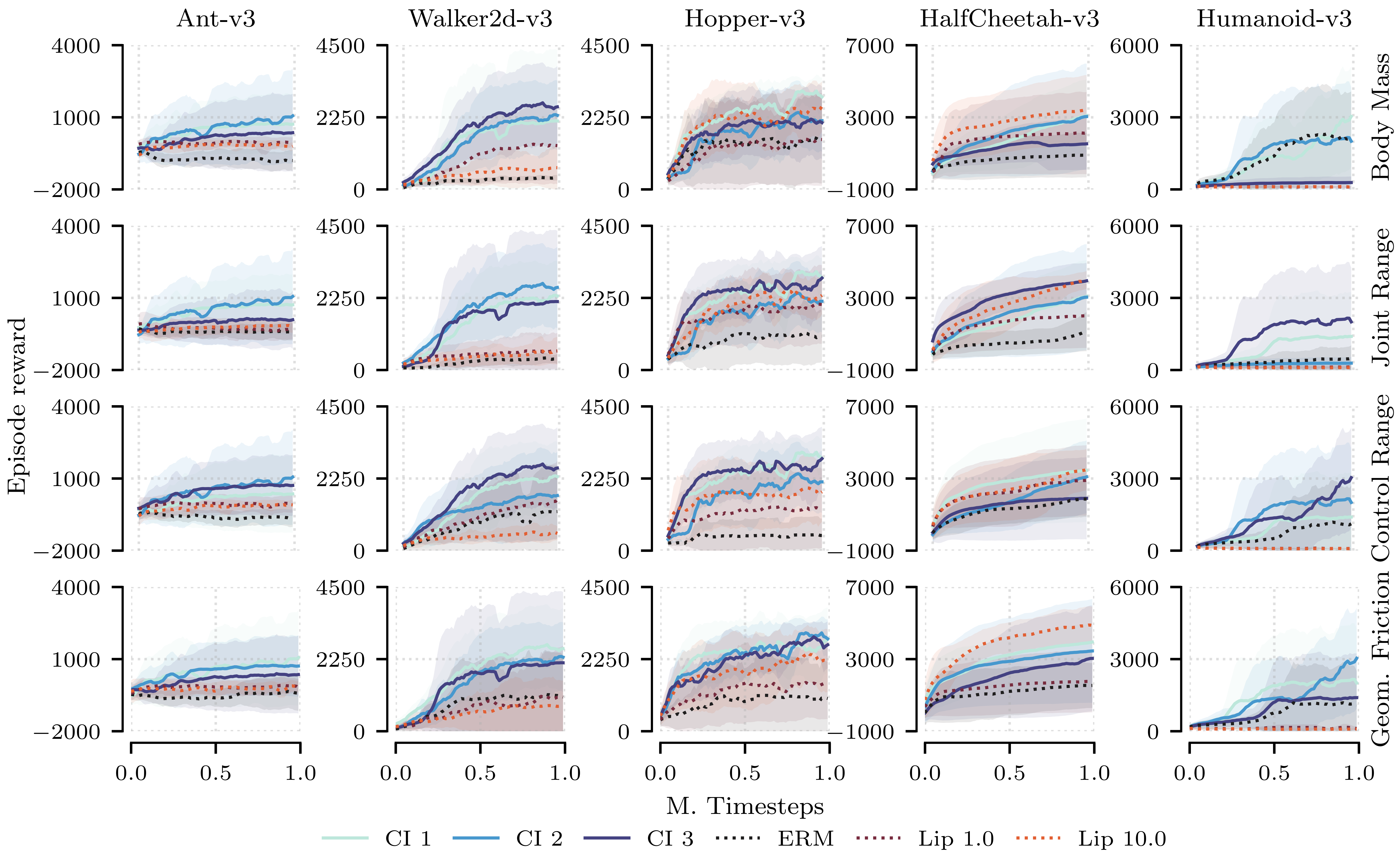
| Environment | Ant | Walker2d | Hopper | HalfCheetah | Humanoid |
|---|---|---|---|---|---|
| Expert | $3168.49_{\pm 1715.68}$ | $3565.33_{\pm 527.40}$ | $3119.54_{\pm 524.36}$ | $4340.61_{\pm 2020.14}$ | $4774.17_{\pm 2063.52}$ |
| AIRL (ERM) | $-18.73_{\pm 255.23}$ | $-3.50_{\pm 1.66}$ | $48.82_{\pm 76.24}$ | $2310.93_{\pm 118.75}$ | $3261.17_{\pm 2272.68}$ |
| AIRL (Lip) | $-213.89_{\pm 738.75}$ | $3202.09_{\pm 185.98}$ | $2544.28_{\pm 445.86}$ | $4293.70_{\pm 666.33}$ | $710.88_{\pm 356.04}$ |
| AIRL (CI) | $155.62_{\pm 875.01}$ | $3670.21_{\pm 599.00}$ | $2906.77_{\pm 490.33}$ | $4653.89_{\pm 762.09}$ | $4022.43_{\pm 671.37}$ |
| GAIL (ERM) | $-486.05_{\pm 388.06}$ | $359.23_{\pm 254.85}$ | $1047.76_{\pm 871.98}$ | $1160.78_{\pm 1134.26}$ | $508.86_{\pm 601.99}$ |
| GAIL (Lip) | $-208.97_{\pm 252.99}$ | $577.40_{\pm 550.86}$ | $2339.87_{\pm 465.97}$ | $4168.13_{\pm 472.22}$ | $122.49_{\pm 82.43}$ |
| GAIL (CI) | $1021.26_{\pm 1845.56}$ | $3479.99_{\pm 1242.39}$ | $2976.49_{\pm 417.33}$ | $5581.43_{\pm 1442.39}$ | $2170.96_{\pm 2425.51}$ |
| MEIRL (ERM) | $-57.30_{\pm 93.23}$ | $-3.69_{\pm 0.18}$ | $3.17_{\pm 0.52}$ | $347.55_{\pm 790.54}$ | $54.70_{\pm 1.92}$ |
| MEIRL (Lip) | $-554.25_{\pm 82.05}$ | $622.45_{\pm 464.25}$ | $1136.95_{\pm 209.97}$ | $-297.84_{\pm 71.23}$ | $1014.42_{\pm 1915.24}$ |
| MEIRL (CI) | $-337.17_{\pm 1310.57}$ | $2292.94_{\pm 1521.43}$ | $2800.37_{\pm 666.09}$ | $1650.50_{\pm 1683.38}$ | $2935.90_{\pm 2262.13}$ |

Table 4.2: **Policy rollout results using ground truth reward for perturbed MuJoCo environments** after being trained for 1M timesteps using the rewards recovered from the different discriminators in section 4.3.2. Here, the *joint range* parameter is perturbed with a noise magnitude of $\varepsilon = 0.2$. The results are averaged over 10 rollouts and obtained by training the model using five different random seeds.

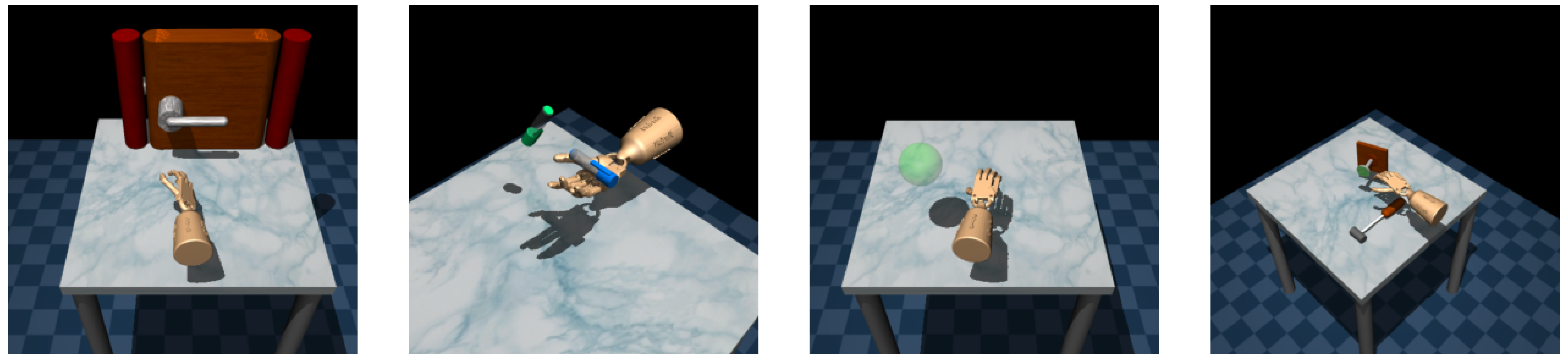
# AIRL training dynamics I

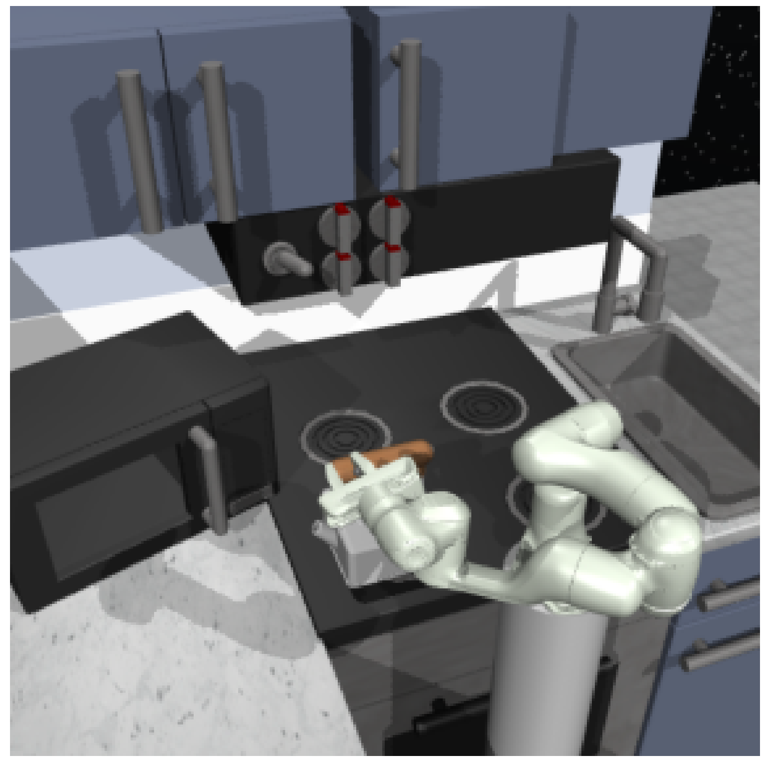


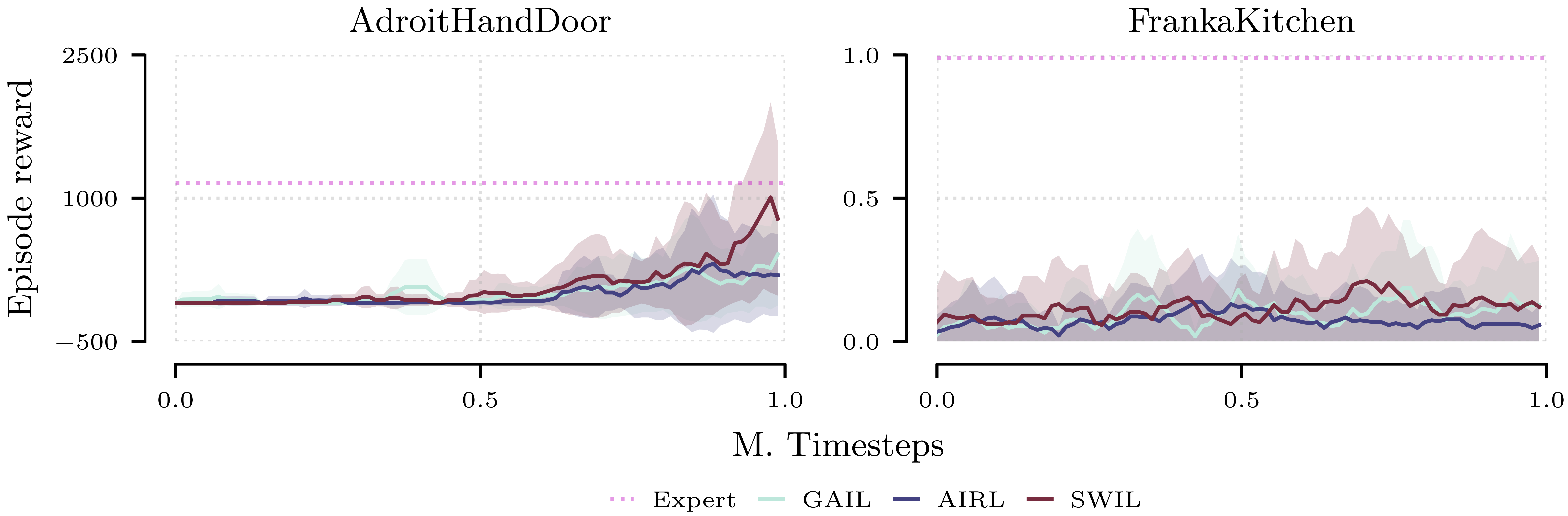
# GAIL Training Dynamics I



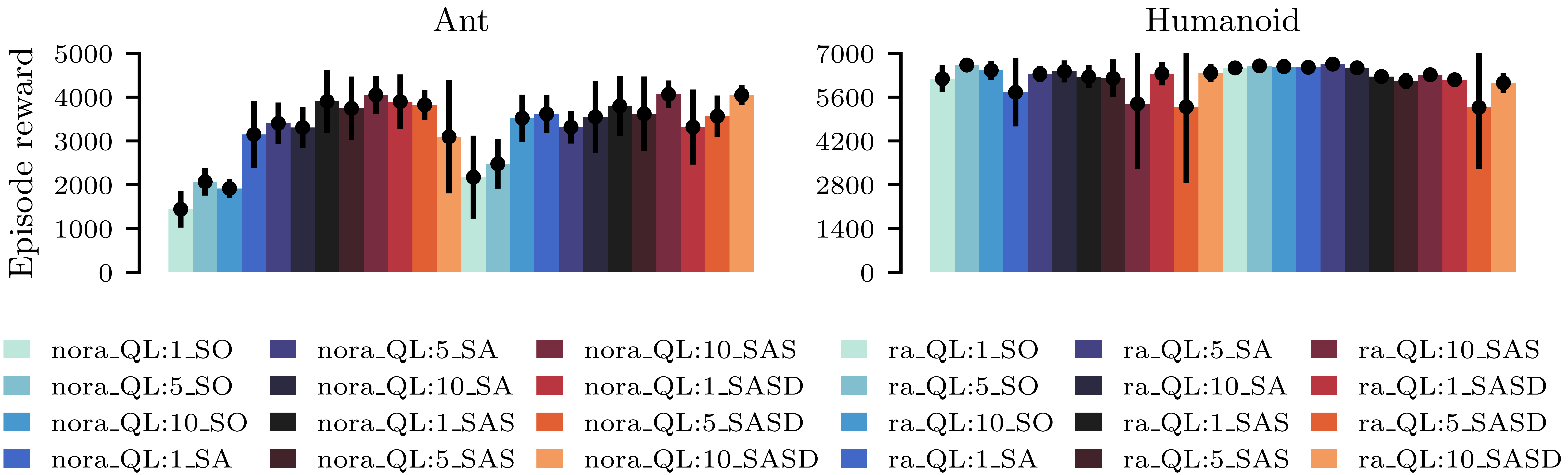
# Adroit / Franka Results



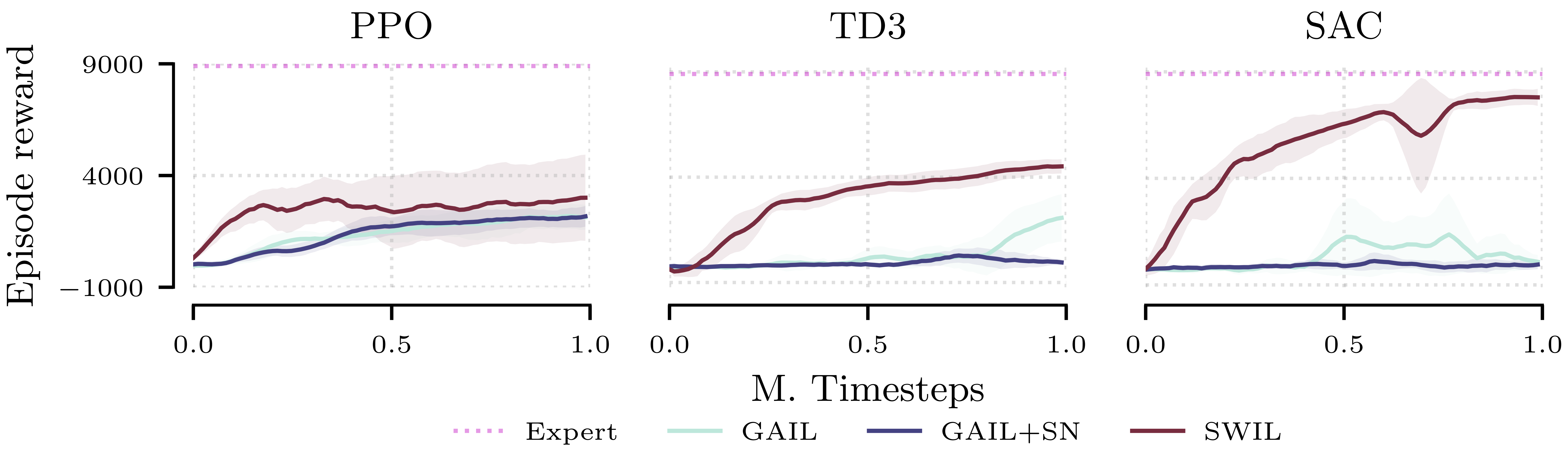


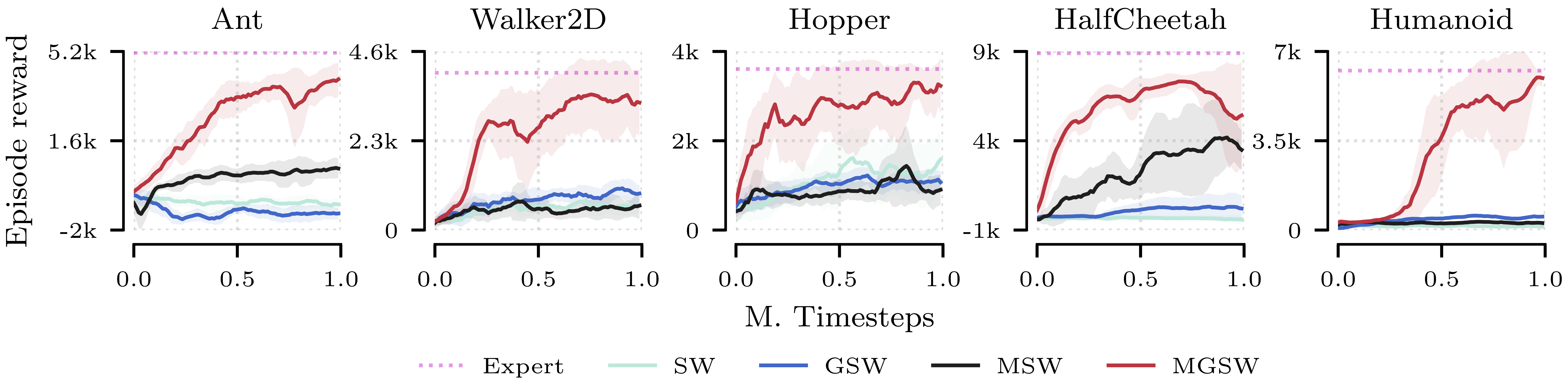


# Ablation SWIL design choices



# SWIL: comp RL / slicing ablation



# SWIL: Algorithm

**Algorithm 1** Sliced Wasserstein Imitation Learning (SWIL)

1: **require** Reinforcement learning algorithm with a policy improvement step for $\pi_\theta$.
2: **input** Expert trajectories $(\xi_i)_{i=1}^N$, initial policy $\pi_\theta$ (and any other initial state needed by the reinforcement learning algorithm), initial $\mathcal{MGSW}_2$-critic $g^{(\psi)} : X \to \mathbb{R}_+$, policy and expert replay buffers with batch size $B$.
3: **while** below maximum number of iterations **do**
4: Compute a rollout $(x_R^{(t)})_{t=1}^T$ using the policy $\pi_\theta$, add each element to the policy replay buffer.
5: Sample a minibatch of state-action pairs $(x_\pi^{(b)})_{b=1}^B$ and $(x_E^{(b)})_{b=1}^B$ from the policy and expert replay buffers and compute $\mathcal{G}_\pi = (g^{(\psi)}(x_\pi^{(b)}))_{b=1}^B$ and $\mathcal{G}_E = (g^{(\psi)}(x_\pi^{(b)}))_{b=1}^B$, which we implicitly interpret as empirical measures.
6: For each $t$, replace the closest atom in $\mathcal{G}_\pi$ with $x_t$, obtaining $\mathcal{G}_R(x_t) = \mathrm{rpl}(\mathcal{G}_\pi, x_t)$.
7: Update $g^{(\psi)}$ via gradient-based supervised learning using the cross-entropy loss $\mathcal{L} = \sum_{b=1}^B \log g^{(\psi)}(x_\pi^{(b)}) - \sum_{b=1}^B (1 - \log g^{(\psi)}(x_E^{(b)}))$ with respect to the sampled minibatches.
8: Perform a reinforcement learning step using the rewards $r_{\mathcal{MGSW}_2} = W_2(\mathcal{G}_E, \mathcal{G}_\pi) - W_2(\mathcal{G}_E, \mathcal{G}_R(x_t))$ for each $t = 1, .., T$.
9: **return** learned policy $\pi_\theta$.

# CI-IRL: Algorithm 1

**Algorithm 1** CI regularized Feature Matching IRL (CI-FMIRL)

**Input:** Expert trajectories $\mathcal{D}_E^e$ assumed to be obtained from multiple experts *by intervening on* $p(\xi|\psi, \varphi)$
**Init:** Initialize reward estimate $r_\psi$ and state feature network $\varphi_\theta$
**for** setting $e$ in $\{1, ..., \mathcal{E}_{tr}\}$ **do**
  **while** $r_{\psi,\varphi}$ not converged **do**
    Compute feature matching gradient $\nabla_\psi \mathcal{L}(\psi, \varphi; e) = \mathbb{E}_{\mathcal{D}_E^e}[\varphi(\xi)] - \mathbb{E}_{p(\xi|\psi)}[\varphi(\xi))$ and *causal invariance* penalty gradient $\nabla_\varphi \mathbb{D}(\psi, \varphi; e)$ and backpropagate the weighted sum through feature network $\varphi_\theta(s)$
    Compute policy $\pi_{r_{\psi,\varphi}}$ using value iteration on the reward estimate $r_{\psi,\varphi}$
  **end for**
**end for**
**Return:** Trained reward $r_{\varphi,\psi}$

# CI-IRL: Algorithm 2

**Algorithm 2** CI regularized Adversarial IRL (CI-AIRL)

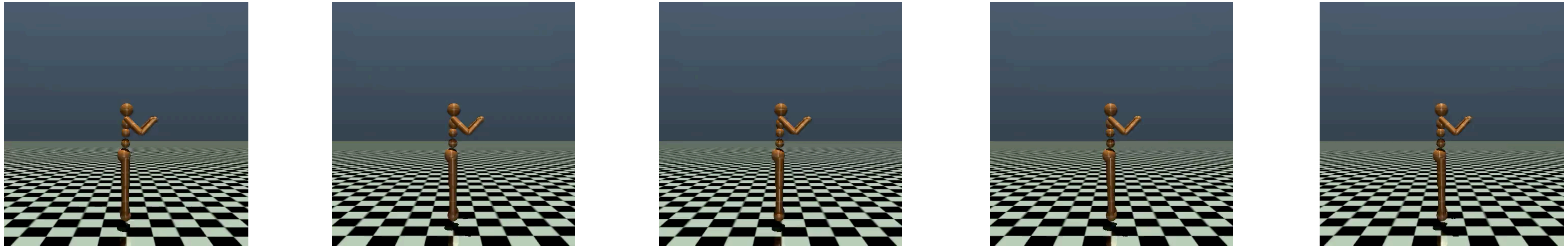
**Input:** Expert trajectories $\mathcal{D}_E^e$ assumed to be obtained from multiple experts *by intervening on* $p(\xi|\psi,\varphi)$
**Init:** Initialize actor-critic $\pi_\theta, \nu_\vartheta$ and discriminator $g_{\xi,\varphi}$
**for** setting $e$ in $\{1, ..., \mathcal{E}_{tr}\}$ **do**
  Collect trajectory buffer $\mathcal{D}_\pi = \{\xi_i\}_{i \leq |\mathcal{D}_\pi|}$ by executing the policy $\pi_\theta$
  Update $g_{\varphi,\theta}(s,a)$ via binary logistic regression by maximizing

$$\mathcal{L}(\varphi,\psi;e) = \mathcal{L}_{\mathrm{BCE}}(\xi,\varphi,\psi;e) + \lambda||\nabla_{\psi|\psi=1.0}\mathcal{L}_{\mathrm{BCE}}(\xi,\varphi,\psi;e)||^2$$

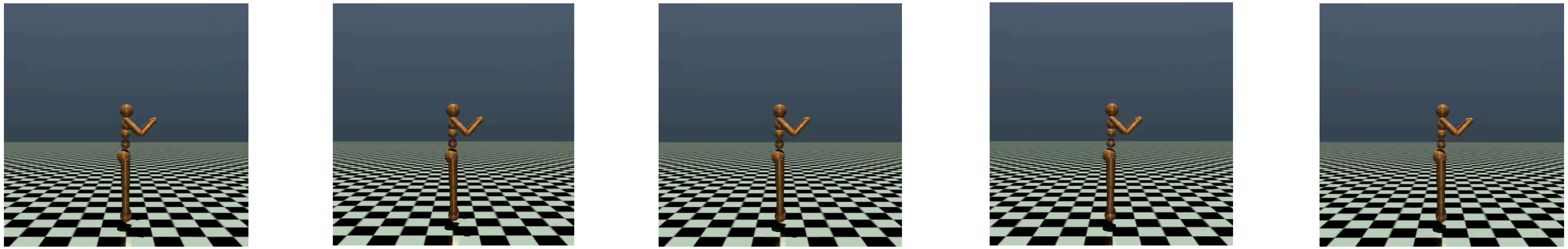
  using dataset tuple $(\mathcal{D}_E^e, \mathcal{D}_\pi)$
  Update actor-critic $(\pi_\theta, \nu_\vartheta)$ w.r.t. the reward function of the *regularized discriminator* using the *soft-actor-critic* RL procedure
**end for**
**Return:** Trained reward $r_{\varphi,\psi}$ and actor-critic $\pi_\theta, \nu_\vartheta$

# More silly walks

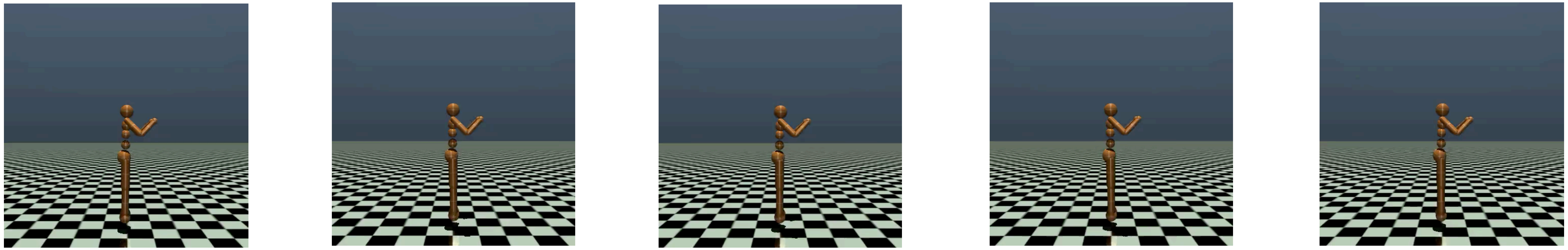
**ERM**



**Lip**



**CI**



# Posterior Agreement in Soft Actor Critic