ICML
International Conference
On Machine Learning

Robust Reward Alignment via Hypothesis Space Batch Cutting

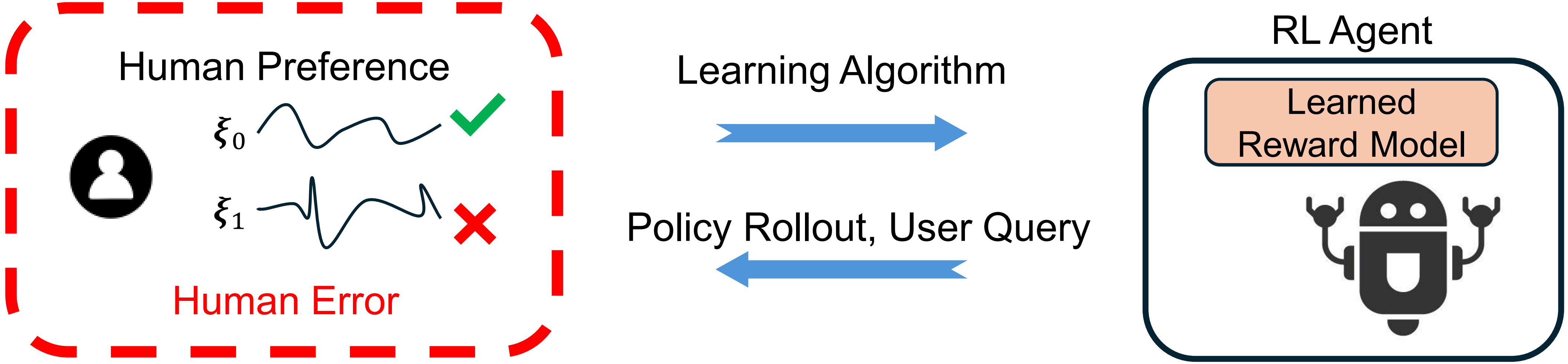
Zhixian Xie, Haode Zhang, Yizhe Feng, Wanxin Jin

Arizona State University

Intelligent Robotics & Interactive Systems

Shanghai Jiao Tong University

Motivation: Preference-Based RL Suffers from Erroneous Feedback



- B-Pref Benchmark: Up to 20% downgrade with 10% error rate[1]

Recent effort: Filtering False Feedback → Prior Knowledge, Computation Overhead

Target:

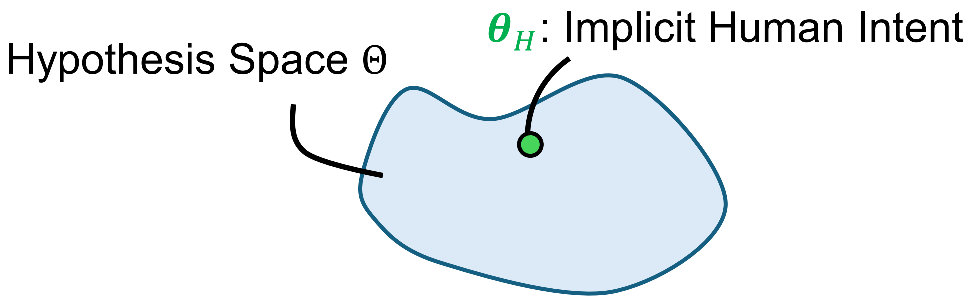
1. **Certifiable human data efficiency**
2. **Provable robustness**

Without explicitly identifying the error.

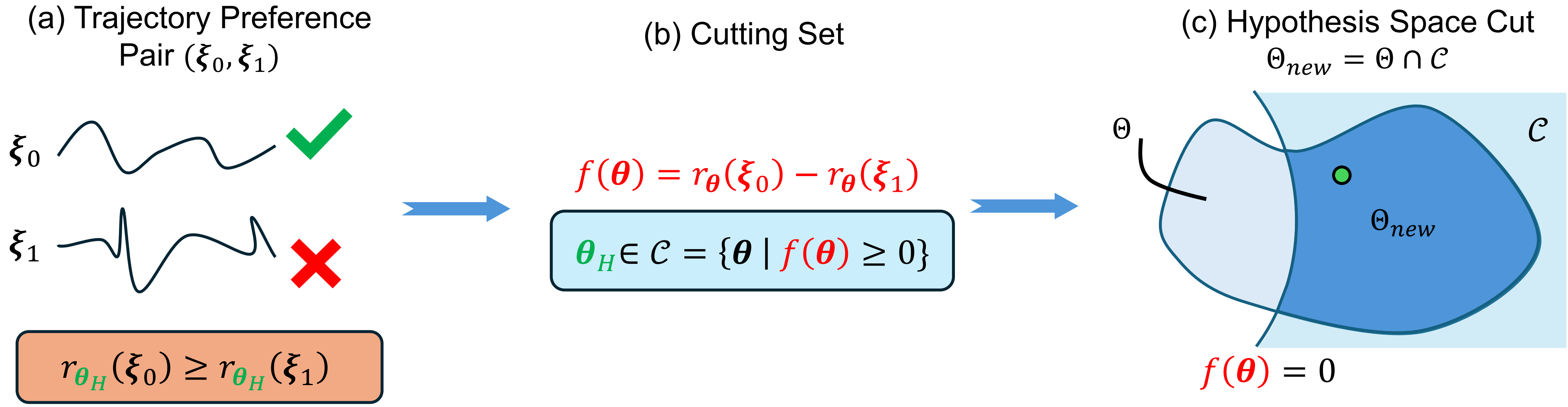
1: Lee, Kimin, et al. "B-pref: Benchmarking preference-based reinforcement learning." arXiv preprint arXiv:2111.03026 (2021).

Preferences as Hypothesis Space Cuts

Hypothesis Space: The space of reward function (parameters), containing implicit human intent.



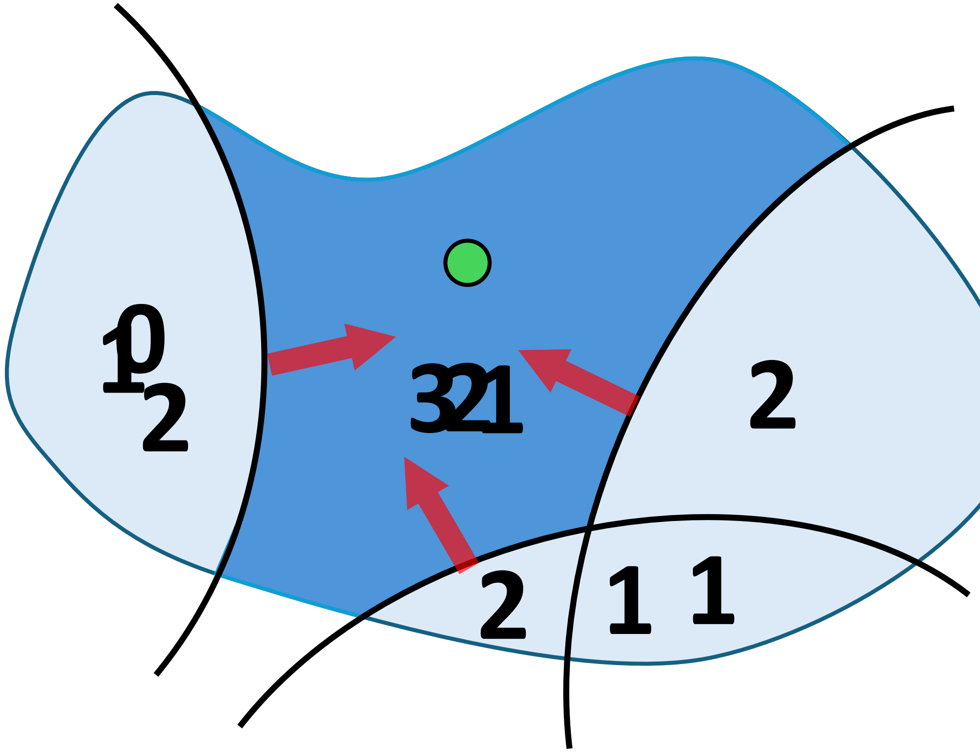
Searching for $\boldsymbol{\theta}_H$: Use preferences to induce cuts and remove the hypothesis space inconsistent with preference!



Batched Cutting as Voting

Use a batch of preference to vote, update the hypothesis space base on vote function:

$$V(\boldsymbol{\theta}) = \#\text{ of satisfied cuts for given }\boldsymbol{\theta}$$

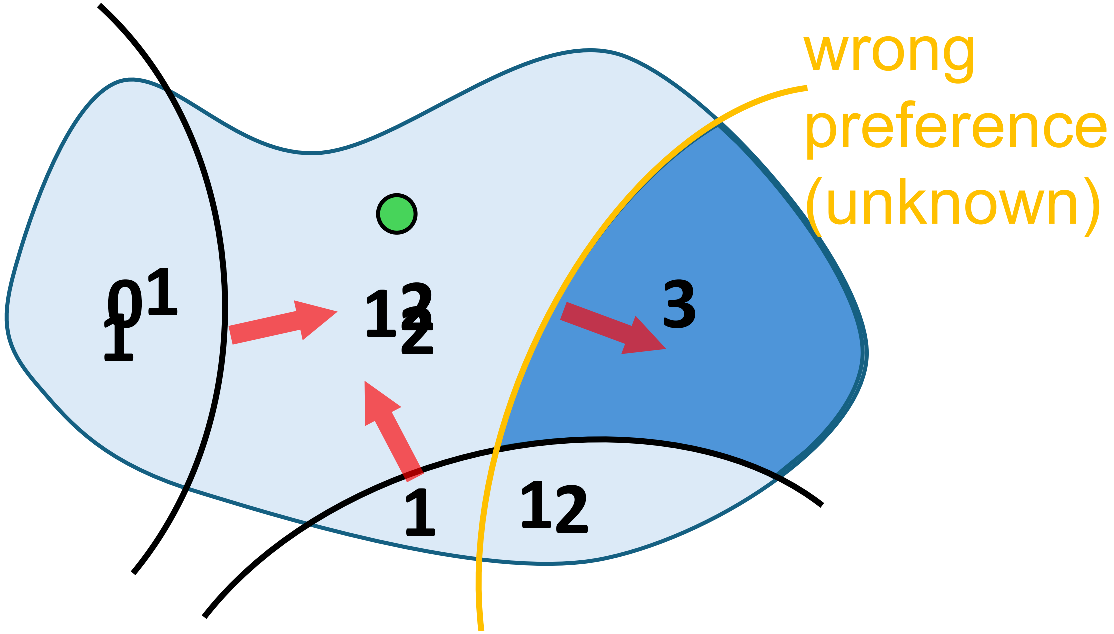


Update the hypothesis space using threshold $V(\boldsymbol{\theta}) \geq 3$!

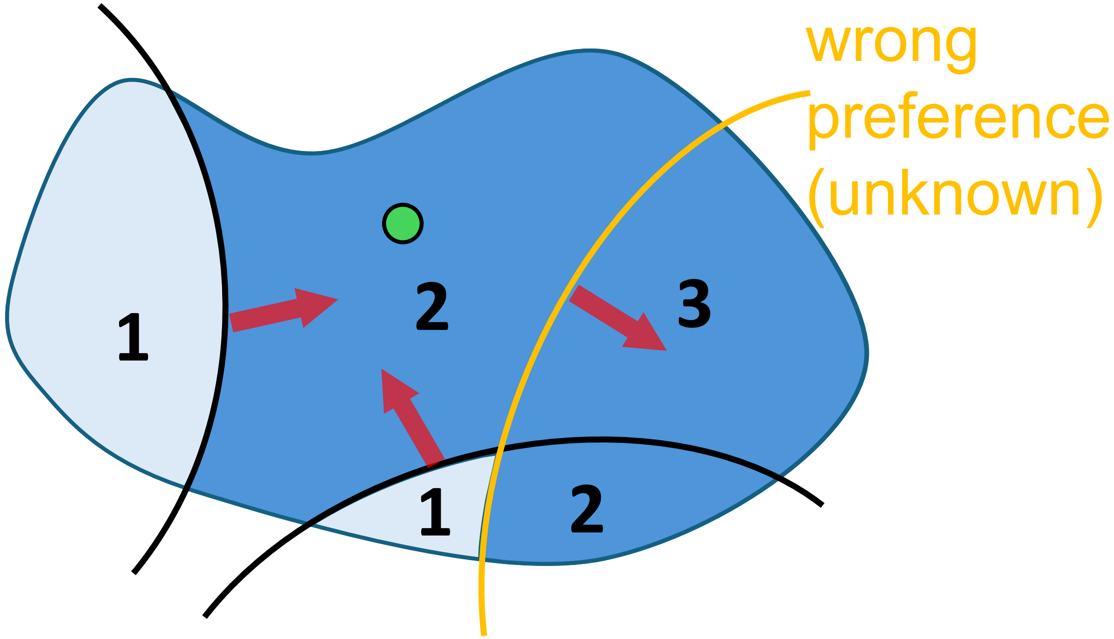
Key Idea: Batched Cutting with Conservativeness

Use Lower threshold to perform conservative update:

Threshold $V(\boldsymbol{\theta}) \geq 3$: Cut out $\boldsymbol{\theta}_H$



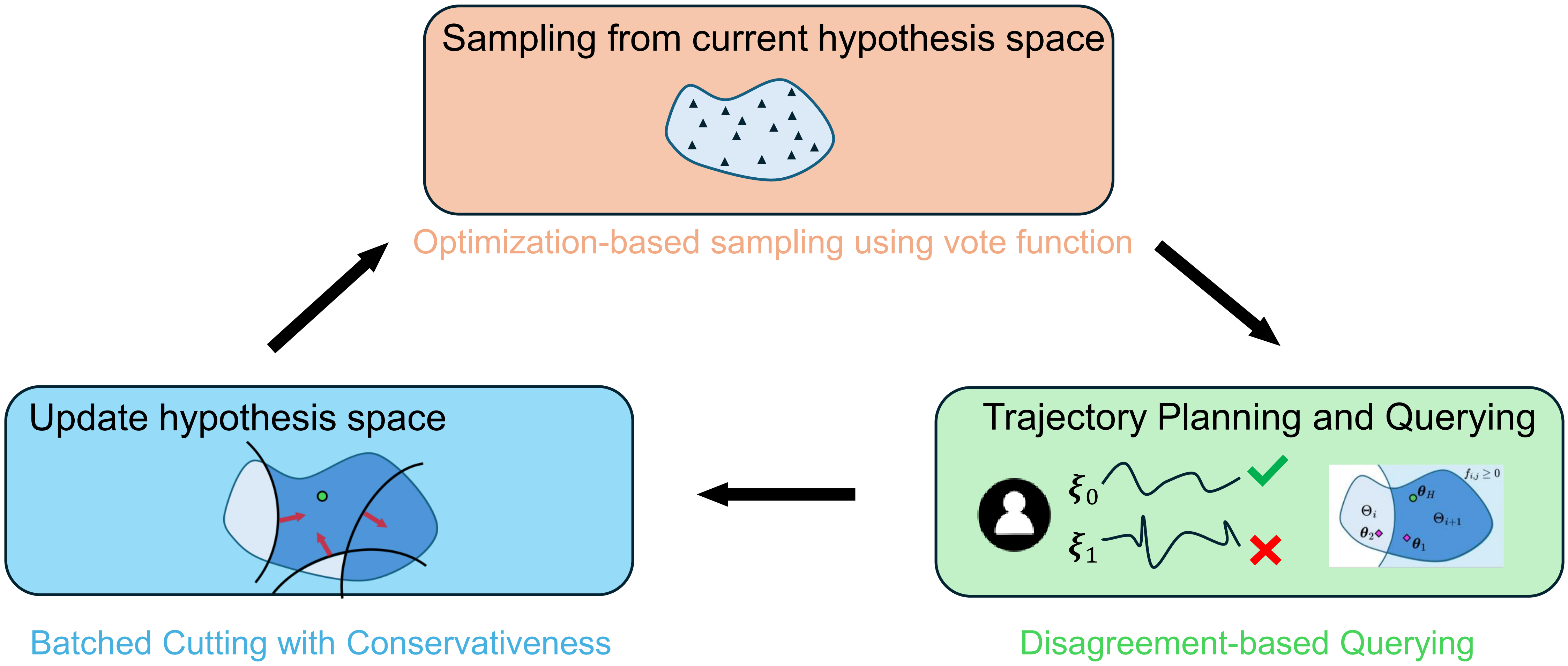
Threshold $V(\boldsymbol{\theta}) \geq 2$: Preserve $\boldsymbol{\theta}_H$ in update



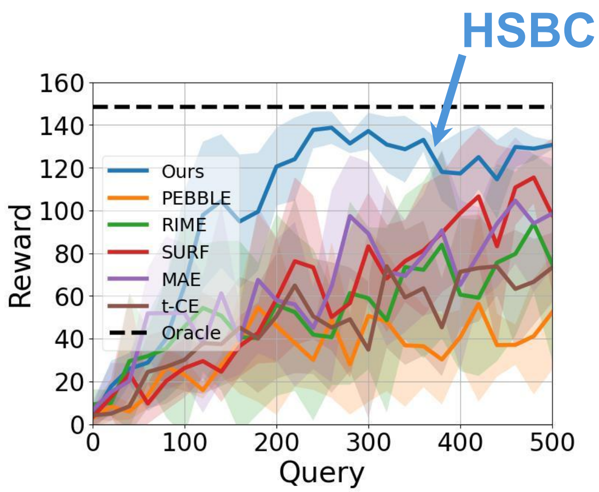
With batch size *N*: $V(\boldsymbol{\theta}) \geq (1-\gamma)N$, conservativeness $0 < \gamma < 1$

Provably Robustness, Error Agnostic!

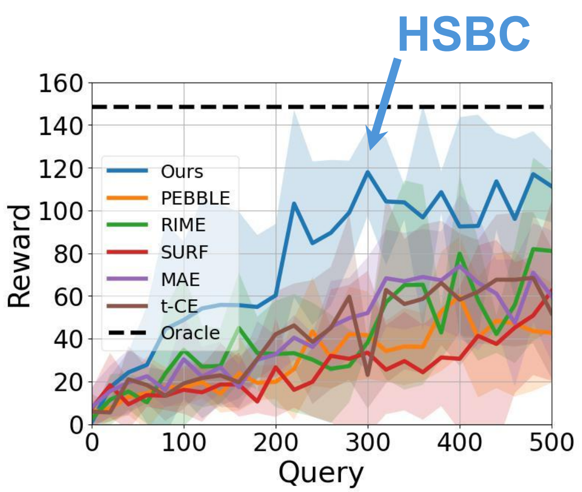
Hypothesis Space Batch Cutting (HSBC) Overview



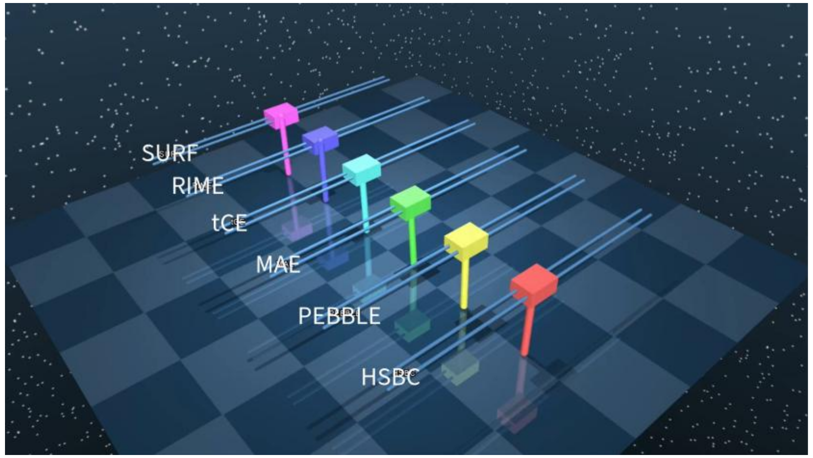
Result



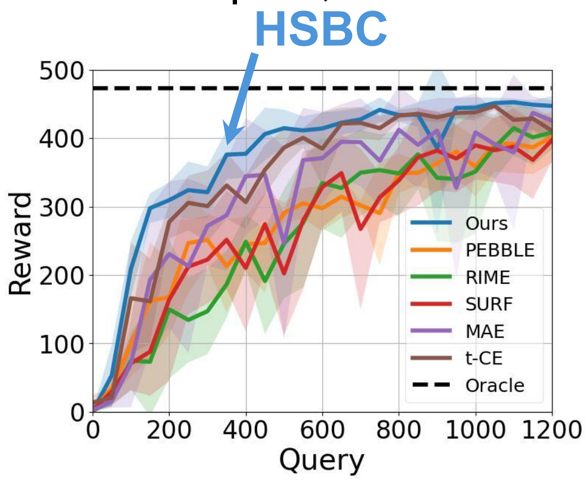
Cartpole, 20% error



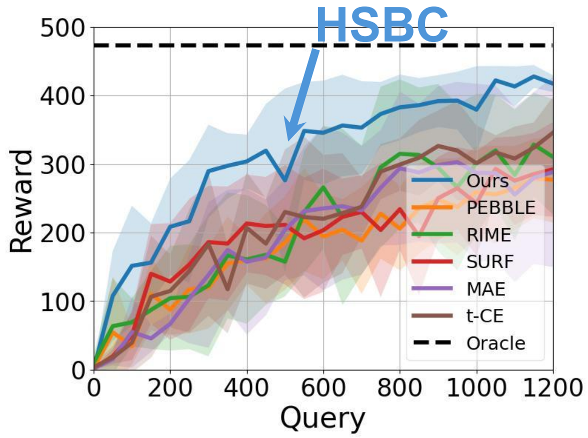
Cartpole, 30% error



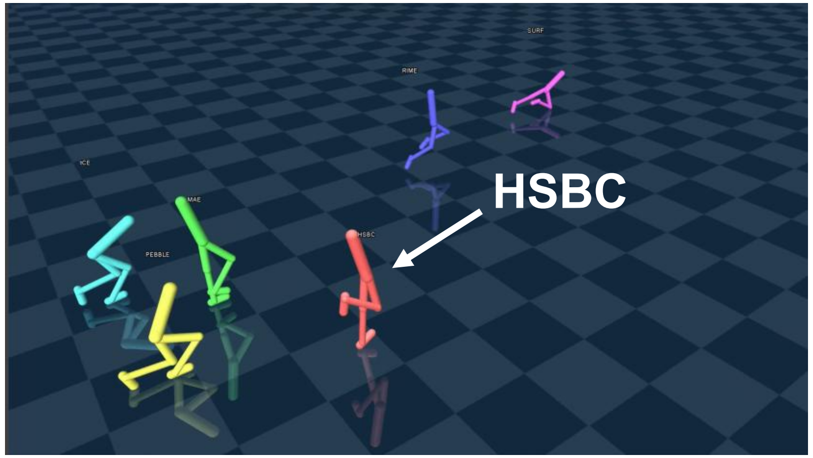
Cartpole, 30% error



Walker, 20% error



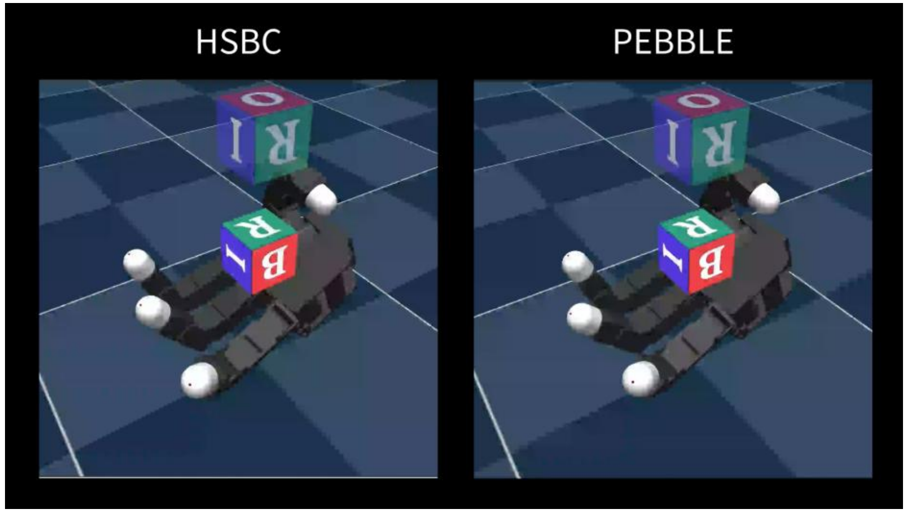
Walker, 30% error



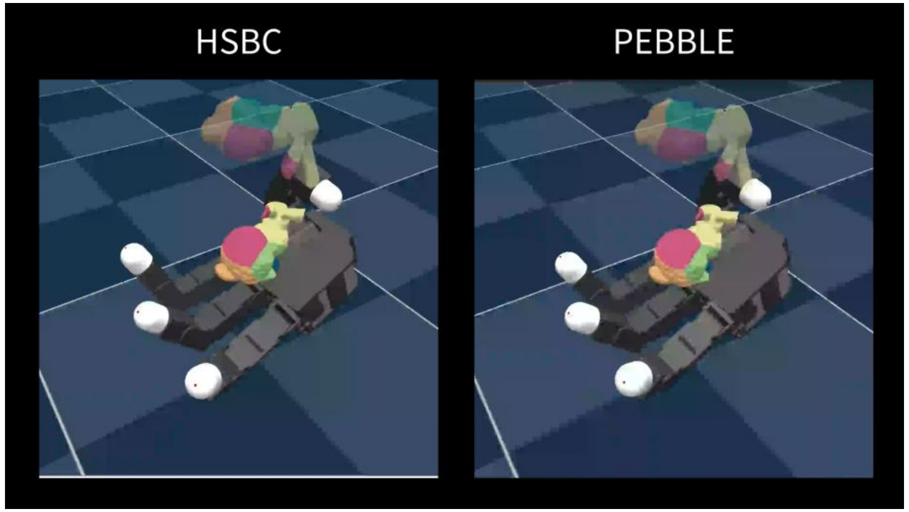
Walker, 30% error

Result

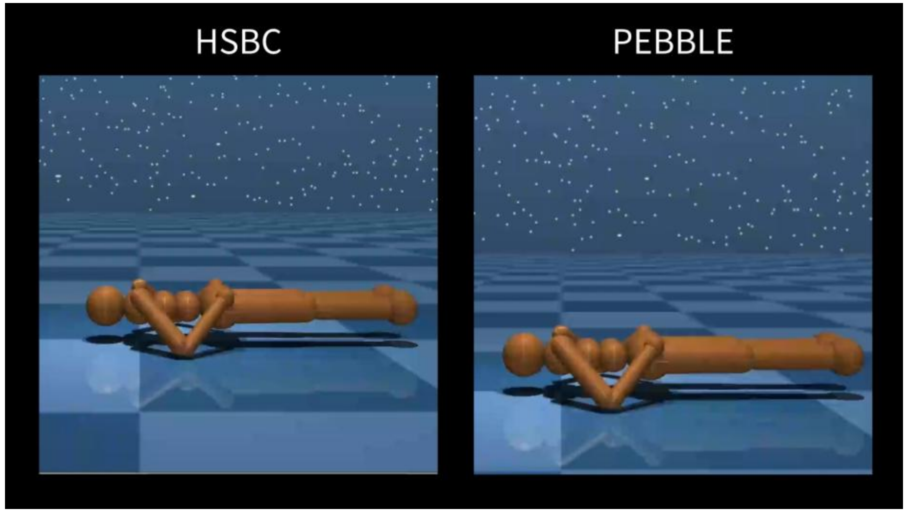
Manipulation: Cube



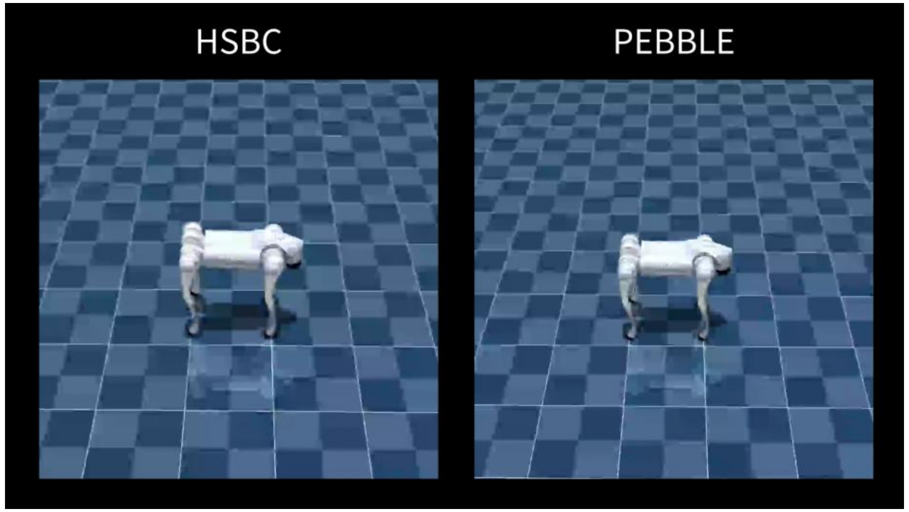
Manipulation: Bunny



Humanoid



Go2



Thanks for Listening!

Project Page