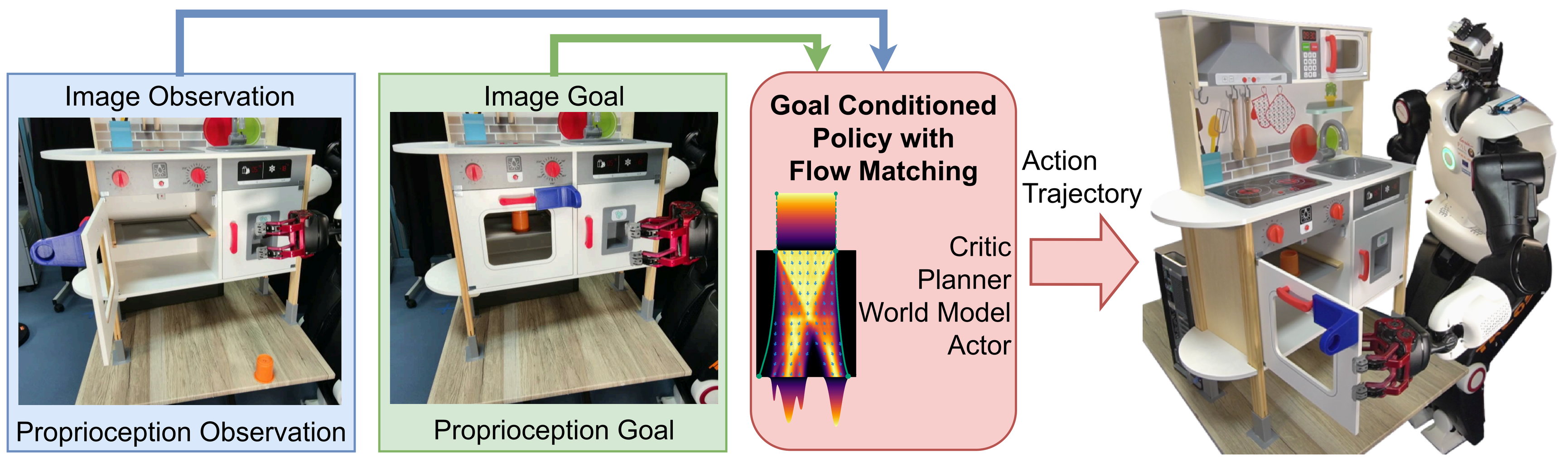
# Extremum Flow Matching for Offline Goal Conditioned Reinforcement Learning

Quentin Rouxel[1,2], Clemente Donoso[1], Fei Chen[2], Serena Ivaldi[1], Jean-Baptiste Mouret[1]

[1]Inria, CNRS, Université de Lorraine, France.
[2]Department of Mechanical and Automation Engineering, T-Stone Robotics Institute, The Chinese University of Hong Kong, Hong Kong.

## Project Website

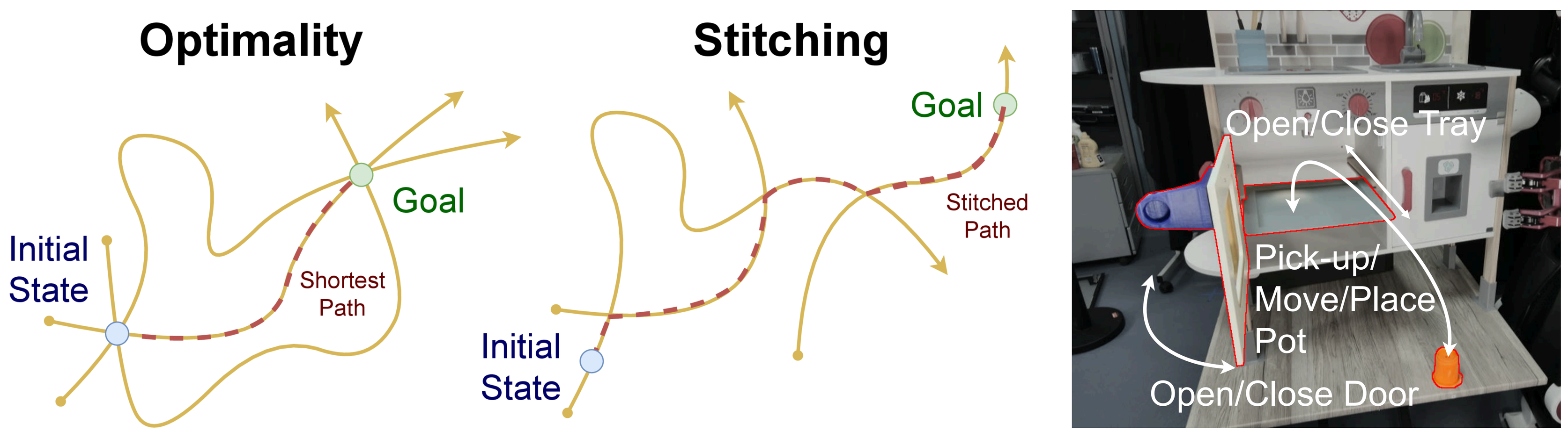
**https://hucebot.github.io/extremum_flow_matching_website/**




## Objectives: Scale-Up Imitation Learning with Play Data

- Build a **generalist policy** capable of complex, long-horizon manipulation tasks.
- Use **goal-conditioned** imitation learning with generative methods.
  No simulation, no reward design, or task labels required.
- Learn from **play data**: open-ended, diverse, exploratory demonstrations without specific tasks or goals.
- Play data is easier and cheaper to collect, enabling scalable training across multiple tasks and environments.

## Challenges of Learning with Play Data

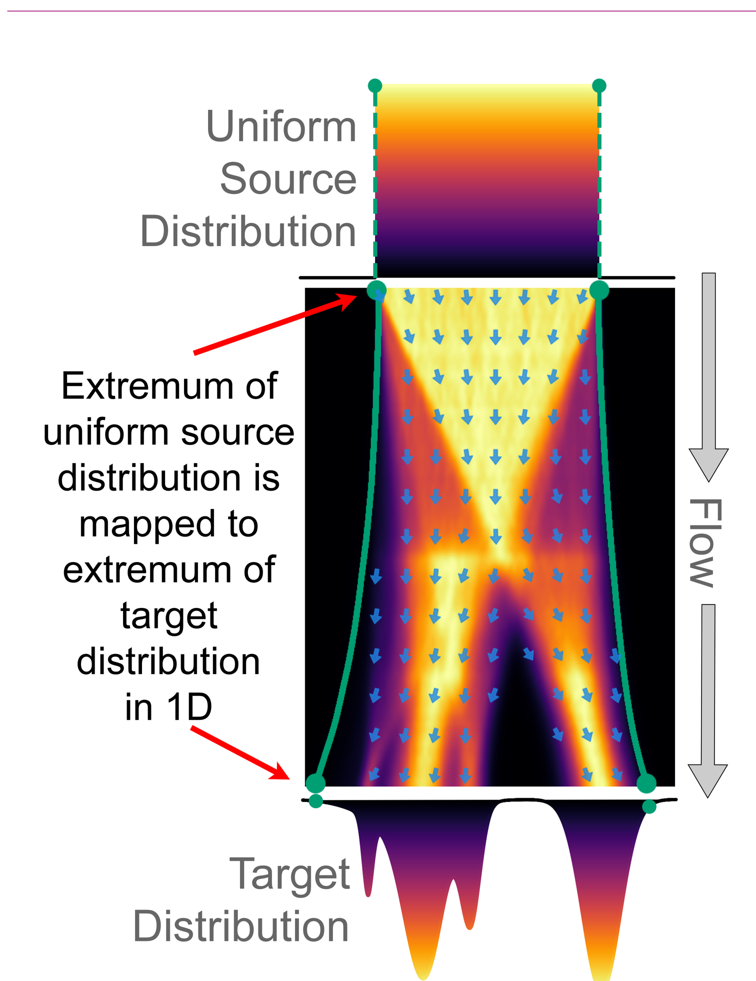



- **Optimality:** Play data contains both direct and inefficient paths to reach goals.
  Agents must identify and prefer the most efficient actions.
- **Stitching:** Full paths to specific distant goals are rarely demonstrated in play data.
  Agents must learn to piece together meaningful segments to reach long-horizon targets.

## Main Ideas

- Introduce **Extremum Flow Matching** to address optimality by estimating minimum and maximum of conditional distributions.
- Propose several **goal-conditioned imitation and offline reinforcement learning** agents based on Flow Matching.
- Evaluate agents on **OGBench**, analyze the impact of data collection strategies, and validate on the real **Talos humanoid robot**.

## Extremum Flow Matching: Min/Max of Conditional Distributions




- Estimates the **minimum or maximum** of a distribution using **Flow Matching**.
- Leverages Flow Matching's unique properties: **deterministic transport** and support for **arbitrary source distributions**, unlike Diffusion.
- Serves as a principled **alternative to Expectile Regression** for offline reinforcement learning.
- Extends to **multi-dimensional distributions** using a structured approach similar to the **conditioning-on-returns** framework.

### Extremum Flow Matching

**Multi-dimensional distributions decomposition**:

$$\boldsymbol{x} = (z, \boldsymbol{y}) \text{ with } z \in \mathbb{R} \text{ and } \boldsymbol{y} \in \mathbb{R}^{n-1}$$

$$\mathcal{P}(\boldsymbol{x}) = \mathcal{P}(z, \boldsymbol{y}) = \mathcal{P}(z)\mathcal{P}(\boldsymbol{y}|z)$$

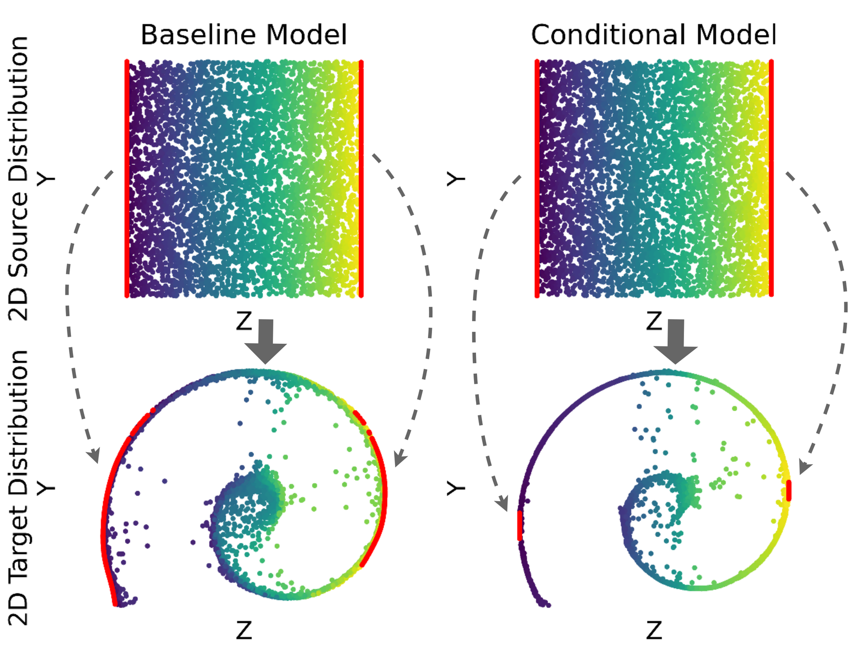
**Objective:** min/max of $z$: $\mathcal{P}(z, \boldsymbol{y}|z = \max \mathcal{P}(z))$

**Training Generative Models with Flow Matching:**

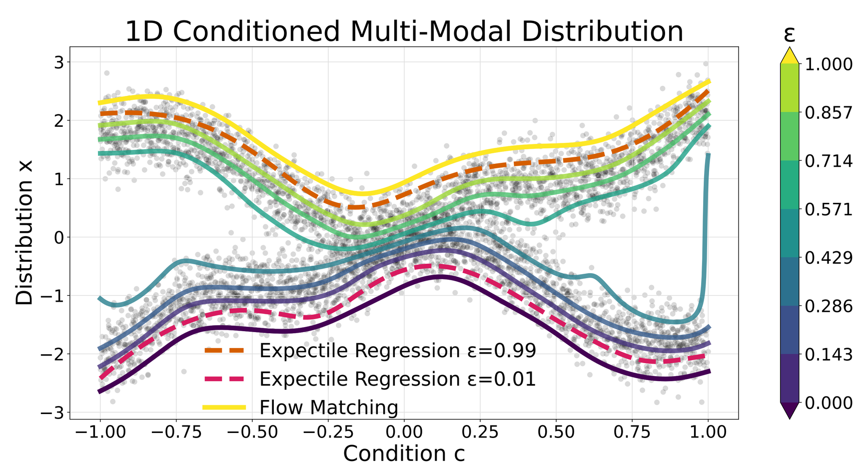
$$F_1 : \mathcal{U}(0,1) \mapsto z, \quad F_2 : \mathcal{P}^{\mathsf{src}}|z \mapsto \boldsymbol{y},$$

**Two Steps Inference:**

$$\tilde{z} = F_1(0) \text{ or } F_1(1), \quad \tilde{\boldsymbol{y}} = F_2(\mathcal{P}^{\mathsf{src}}|\tilde{z}).$$




Extrema of unconditioned 2D distributions




Expectile Regression vs. Extremum Flow Matching

## Goal Conditioned Agents with Extremum Flow Matching

Imitation learning and reinforcement learning agents are composed of **multiple interacting components**. We propose a **family of algorithms** that combine these core modules in different ways with Flow Matching:

- **Critic** : $\bullet \mapsto d$ estimates expected return (distance in time-step to goal) given observation and goal
- **Planner** : $\bullet \mapsto \boldsymbol{\tau}^o$ generates sub-goal or trajectory of future observations toward goal
- **Actor** : $\bullet \mapsto \boldsymbol{\tau}^a$ (inverse dynamic) produces action or trajectory of actions to reach goal
- **World** : $\boldsymbol{\tau}^a, \bullet \mapsto \boldsymbol{\tau}^o$ (world model) predicts environment's dynamics (trajectory of future observation) from actions

We use **dataset with trajectories formalism:** $(\boldsymbol{o}_k, \boldsymbol{\tau}^o_k, \boldsymbol{\tau}^a_k, d, \boldsymbol{g})$

| Name | Training | Inference | Comment |
|---|---|---|---|
| **FM-GC** | Actor : $\mathcal{P}^{\mathsf{src}}_{\tau_a}\|\boldsymbol{o}, \boldsymbol{g} \mapsto \boldsymbol{\tau}^a$ | $\tilde{\boldsymbol{\tau}}^a = \text{Actor}(\mathcal{P}^{\mathsf{src}}_{\tau_a}\|\boldsymbol{o}, \boldsymbol{g})$ | Baseline goal conditioned with Flow Matching |
| **FM-AC** | Critic : $\mathcal{U}(0,1)\|\boldsymbol{o}, \boldsymbol{g} \mapsto d$<br>Actor : $\mathcal{P}^{\mathsf{src}}_{\tau_a}\|\boldsymbol{o}, \boldsymbol{g}, d \mapsto \boldsymbol{\tau}^a$ | $\tilde{d} = \text{Critic}(0\|\boldsymbol{o}, \boldsymbol{g})$<br>$\tilde{\boldsymbol{\tau}}^a = \text{Actor}(\mathcal{P}^{\mathsf{src}}_{\tau_a}\|\boldsymbol{o}, \boldsymbol{g}, \tilde{d})$ | Actor conditioned, inspired by GCIQL |
| **FM-PC** | Critic : $\mathcal{U}(0,1)\|\boldsymbol{o}, \boldsymbol{g} \mapsto d$<br>Planner : $\mathcal{P}^{\mathsf{src}}_{\tau_o}\|\boldsymbol{o}, \boldsymbol{g}, d \mapsto \boldsymbol{\tau}^o$<br>Actor : $\mathcal{P}^{\mathsf{src}}_{\tau_a}\|\boldsymbol{o}, \boldsymbol{\tau}^o \mapsto \boldsymbol{\tau}^a$ | $\tilde{d} = \text{Critic}(0\|\boldsymbol{o}, \boldsymbol{g})$<br>$\tilde{\boldsymbol{\tau}}^o = \text{Planner}(\mathcal{P}^{\mathsf{src}}_{\tau_o}\|\boldsymbol{o}, \boldsymbol{g}, \tilde{d})$<br>$\tilde{\boldsymbol{\tau}}^a = \text{Actor}(\mathcal{P}^{\mathsf{src}}_{\tau_a}\|\boldsymbol{o}, \tilde{\boldsymbol{\tau}}^o)$ | Planner conditioned, inspired by HIQL |
| **FM-PS** | Critic : $\mathcal{U}(0,1)\|\boldsymbol{o}, \boldsymbol{g} \mapsto d$<br>Planner : $\mathcal{P}^{\mathsf{src}}_{\tau_o}\|\boldsymbol{o} \mapsto \boldsymbol{\tau}^o$<br>Actor : $\mathcal{P}^{\mathsf{src}}_{\tau_a}\|\boldsymbol{o}, \boldsymbol{\tau}^o \mapsto \boldsymbol{\tau}^a$ | $T^o = \{\boldsymbol{\tau}^o \| \boldsymbol{\tau}^o \sim \text{Planner}(\mathcal{P}^{\mathsf{src}}_{\tau_o}\|\boldsymbol{o})\}$<br>$\tilde{\boldsymbol{\tau}}^o = \underset{\boldsymbol{\tau}^o \in T^o}{\text{argmin}}\ \text{Critic}(0\|\boldsymbol{\tau}^o_{-1}, \boldsymbol{g})$<br>$\tilde{\boldsymbol{\tau}}^a = \text{Actor}(\mathcal{P}^{\mathsf{src}}_{\tau_a}\|\boldsymbol{o}, \tilde{\boldsymbol{\tau}}^o)$ | Planner rejection sampling, inspired by Diffusion Veteran |
| **FM-AS** | Critic : $\mathcal{U}(0,1)\|\boldsymbol{o}, \boldsymbol{g} \mapsto d$<br>Actor : $\mathcal{P}^{\mathsf{src}}_{\tau_a}\|\boldsymbol{o} \mapsto \boldsymbol{\tau}^a$<br>World : $\mathcal{P}^{\mathsf{src}}_{\tau_o}\|\boldsymbol{o}, \boldsymbol{\tau}^a \mapsto \boldsymbol{\tau}^o$ | $T^a = \{\boldsymbol{\tau}^a \| \boldsymbol{\tau}^a \sim \text{Actor}(\mathcal{P}^{\mathsf{src}}_{\tau_a}\|\boldsymbol{o})\}$<br>$\tilde{\boldsymbol{\tau}}^a = \underset{\boldsymbol{\tau}^a \in T^a}{\text{argmin}}\ \text{Critic}(0\|\boldsymbol{\tau}^o_{-1}, \boldsymbol{g})$<br>where $\boldsymbol{\tau}^o = \text{World}(\mathcal{P}^{\mathsf{src}}_{\tau_o}\|\boldsymbol{o}, \boldsymbol{\tau}^a)$ | Actor rejection sampling with world model |

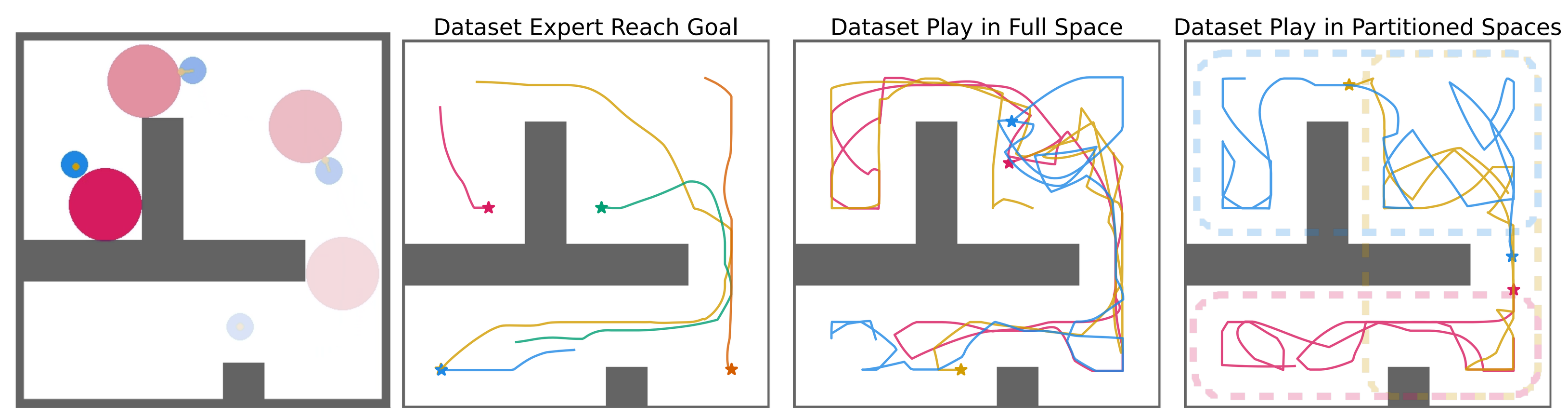
**Reinforcement Learning Recursive Bootstrap:** augment dataset to address **stitching**:

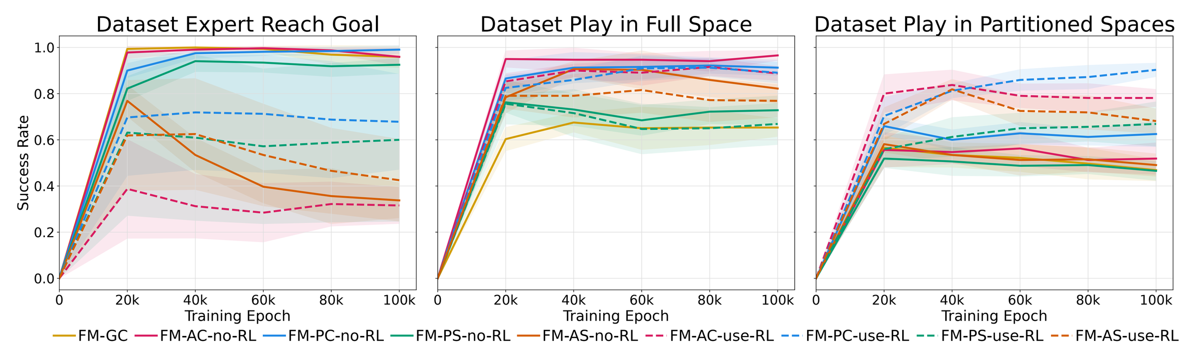
$$(\boldsymbol{o}, \boldsymbol{\tau}^o, \boldsymbol{\tau}^a, d + \text{Critic}(\epsilon_g|\boldsymbol{g}, \boldsymbol{g}'), \boldsymbol{g}'), \text{ with } \epsilon_g \sim \mathcal{U}(0, r_g), r_g \in [0,1],$$

## Comparison on OGBench Benchmark

| | | no-RL | | | | use-RL | | | | OGBench | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **OGBench Dataset** | **FM-GC** | **FM-AC** | **FM-PC** | **FM-PS** | **FM-AS** | **FM-AC** | **FM-PC** | **FM-PS** | **FM-AS** | **GCBC** | **GCIVL** | **GCIQL** | **QRL** | **CRL** | **HIQL** |
| pointmaze-large-navigate-v0 | 66 | 60 | 60 | 31 | 29 | **89** | **89** | 67 | 64 | 29 | 45 | 34 | 86 | 39 | 58 |
| pointmaze-large-stitch-v0 | 39 | 37 | 23 | 15 | 14 | 40 | 40 | 42 | 44 | 7 | 12 | 31 | **84** | 0 | 13 |
| antmaze-large-navigate-v0 | 7 | 5 | 5 | 15 | 1 | 7 | 22 | 34 | 15 | 24 | 16 | 34 | 75 | 83 | **91** |
| antmaze-large-stitch-v0 | 1 | 0 | 0 | 6 | 3 | 0 | 3 | 18 | 7 | 3 | 18 | 7 | 18 | 11 | **67** |
| cube-double-play-v0 | **69** | 32 | 13 | 22 | 14 | 2 | 1 | 12 | 16 | 1 | 36 | 40 | 1 | 10 | 6 |
| scene-play-v0 | 53 | 52 | 32 | 42 | 43 | 7 | 16 | 40 | **55** | 5 | 42 | 51 | 5 | 19 | 38 |
| puzzle-4x4-play-v0 | 1 | 0 | 3 | 22 | **48** | 0 | 1 | 14 | 38 | 0 | 13 | 26 | 0 | 0 | 7 |

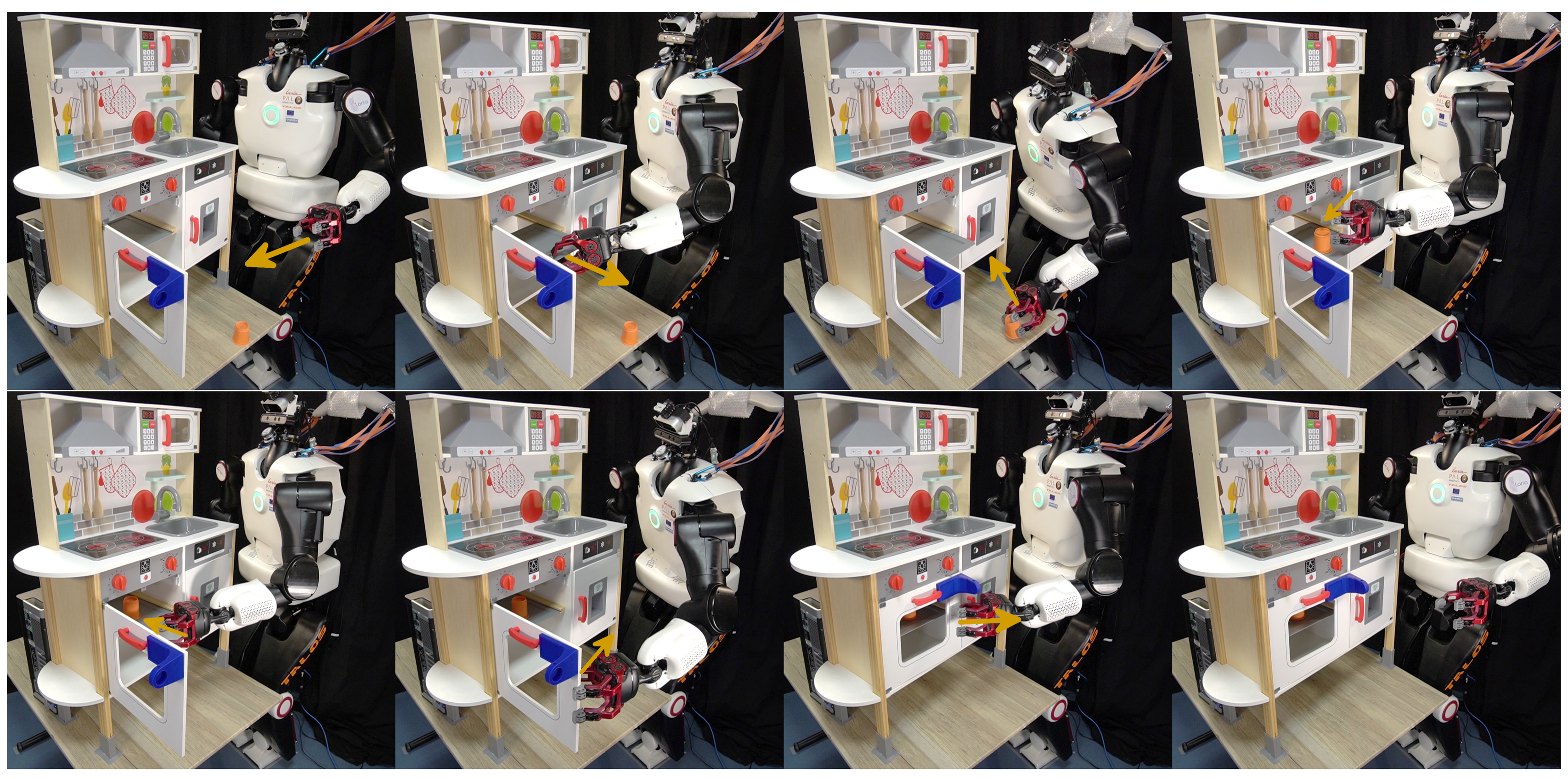
## Impact of Demonstration Behaviors




| Agent | Expert | Full | Partitioned |
|---|---|---|---|
| **FM-GC** | **96**±5 | 65±4 | 47±5 |
| **FM-AC-no-RL** | **96**±4 | **97**±2 | 52±7 |
| **FM-PC-no-RL** | **99**±2 | 91±3 | 62±5 |
| **FM-PS-no-RL** | 92±4 | 73±6 | 47±4 |
| **FM-AS-no-RL** | 34±8 | 82±6 | 49±5 |
| **FM-AC-use-RL** | 32±8 | 89±3 | 78±4 |
| **FM-PC-use-RL** | 68±21 | 89±4 | **90**±3 |
| **FM-PS-use-RL** | 60±35 | 67±9 | 67±10 |
| **FM-AS-use-RL** | 42±18 | 77±7 | 68±5 |

## Vision-Based Manipulation with Talos Humanoid Robot



## Main Results

- Successfully solve multi-step, long-horizon manipulation learned from suboptimal non-expert play data.
- Algorithm performance is highly sensitive to dataset properties, especially the collection policy.
- No single agent consistently outperforms others across all settings.

## References

[1] Q. Rouxel et al. "Flow Matching Imitation Learning for Multi-Support Manipulation". In: *2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids)*. IEEE. 2024.

[2] Q. Rouxel et al. "Multi-Contact Whole-Body Force Control for Position-Controlled Robots". In: *IEEE RA-L* (2024).