Residual-MPPI: Online Policy Customization for Continuous Control

Pengcheng Wang[*1], Chenran Li[*1], Catherine Weaver[1], Kenta Kawamoto[2], Masayoshi Tomizuka[1], Tang Chen[3], Wei Zhan[1]

[1]University of California Berkeley; [2]Sony AI, USA; [3]University of Texas at Austin

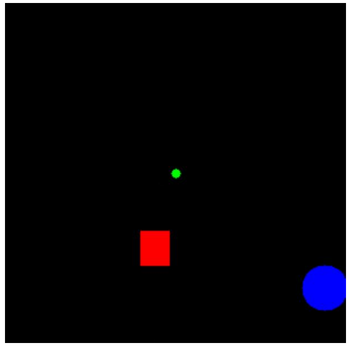
*Equal Contribution

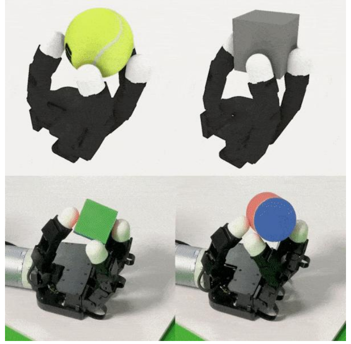
ArXiv: [2407.00898](https://arxiv.org/abs/2407.00898)

ICLR

Motivations

RL/ IL-based Advanced Polices

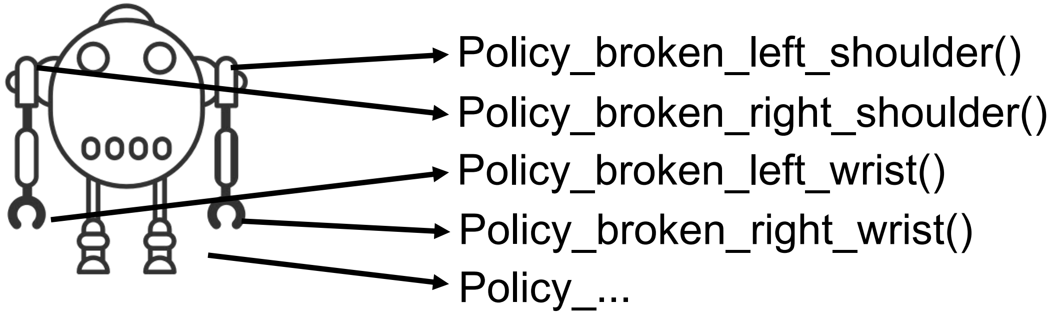

Navigation


Manipulation


Locomotion

Efficient Policy Customization

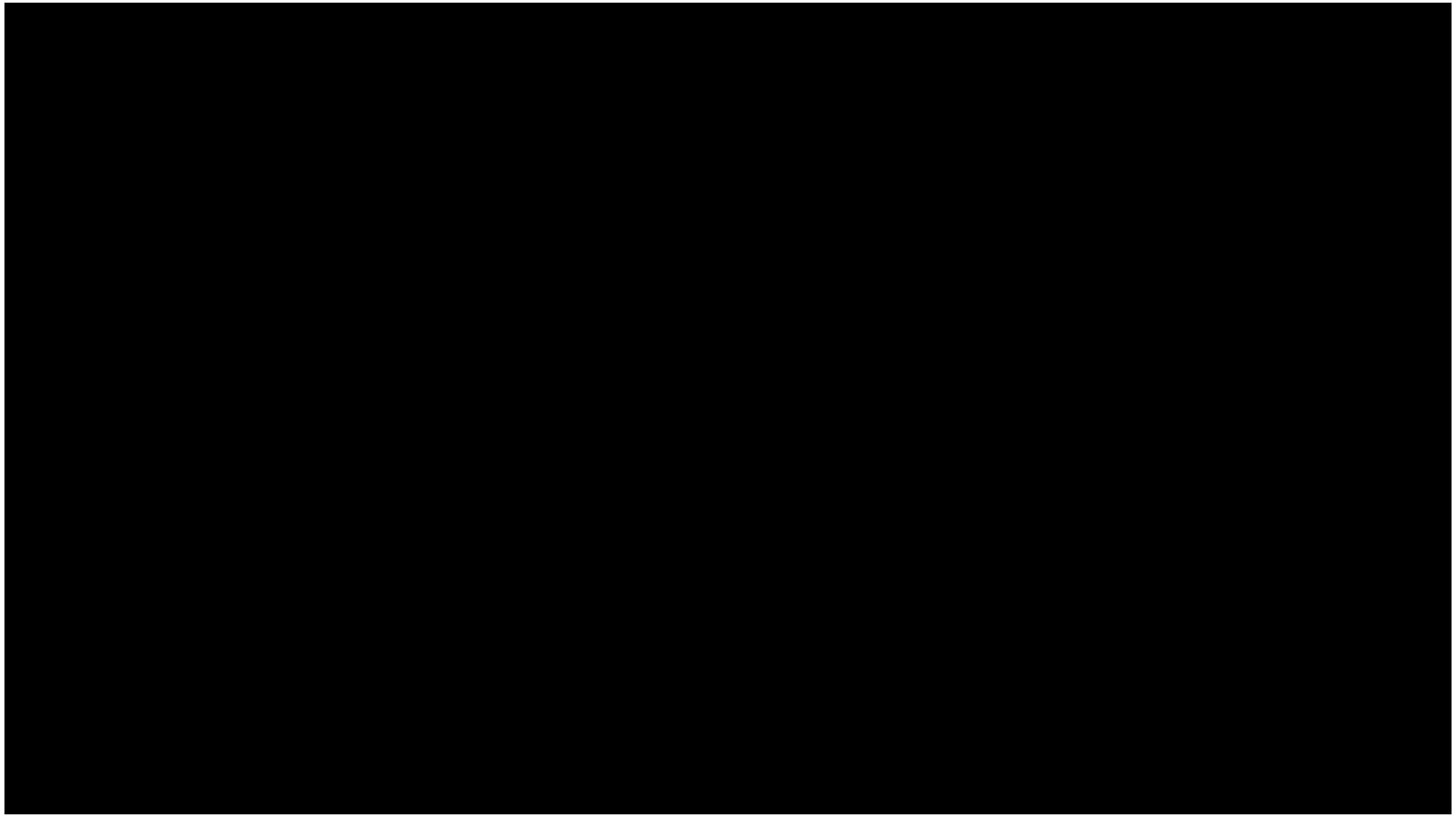
- **Access to original training metrics**
- **“Policies for every disabled joints”**



Online for Efficient Policy Customization!

Example

- GT Sophy 1.0 Behavior



Preliminaries

Residual Q Learning (RQL)

- **Solve Policy Customization**
- **~ Solve MaxEnt Augmented MDP**

$$\hat{\mathcal{M}} = (\mathcal{X}, \mathcal{U}, \omega r + r_R, p)$$

$$\mathcal{M}^{\mathrm{aug}} = (\mathcal{X}, \mathcal{U}, \omega' \log \pi(\boldsymbol{u}|\boldsymbol{x}) + r_R, p)$$

Model Predictive Path Integral (MPPI)

- **Solve MaxEnt MDP Online**

$$S_{\boldsymbol{x}_0}(U) = \sum_{t=0}^{T-1} r(\boldsymbol{x}_t, \boldsymbol{u}_t) + \phi(\boldsymbol{x}_T)$$

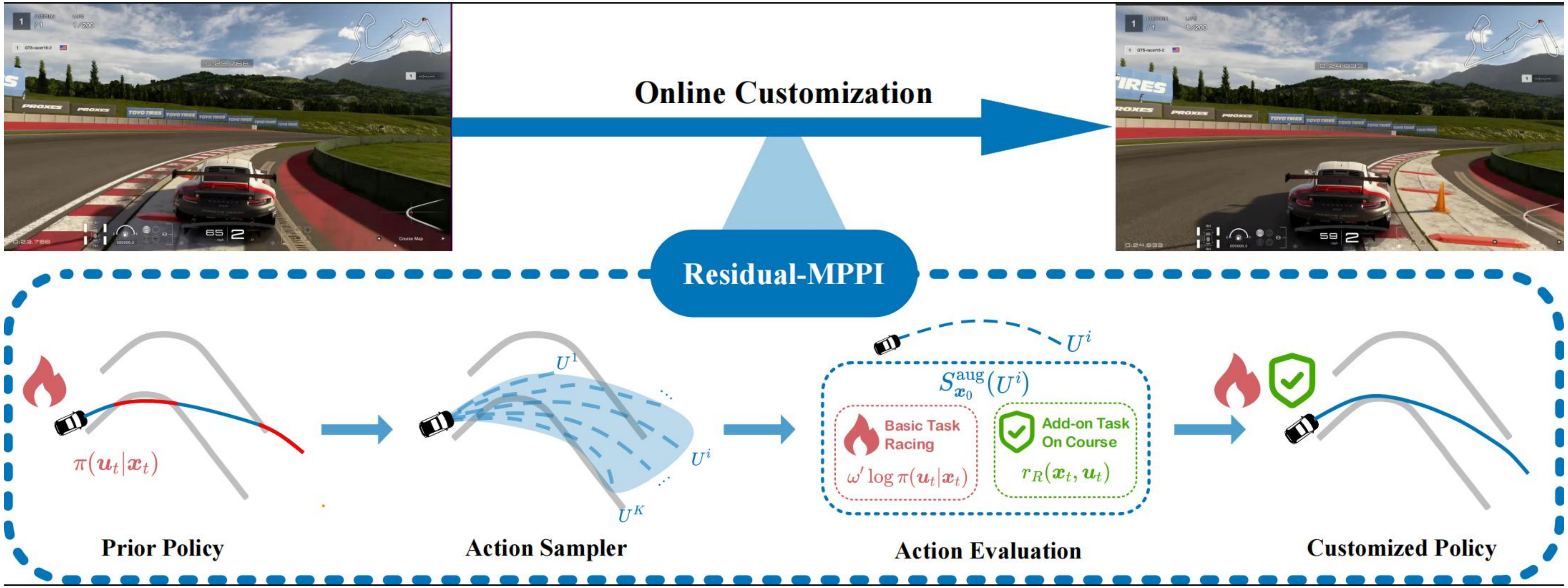
$$q^*(U) = \frac{1}{\eta} \exp\left(\frac{1}{\lambda} S_{\boldsymbol{x}_0}(U)\right) p(\mathcal{E})$$

Methods

- **Residual-MPPI**

$$S^{\text{aug}}_{\boldsymbol{x}_0}(U) = \sum_{t=0}^{T-1} \gamma^t \cdot \left(r_R(\boldsymbol{x}_t, \boldsymbol{u}_t) + \omega' \log \pi(\boldsymbol{u}_t|\boldsymbol{x}_t)\right)$$

- **Planning Loop**



Experiments

MuJoCo Environment

Env.	Policy	Full Task	Basic Task	Add-on Task	
		Total Reward	Basic Reward	$\lvert\bar{\theta}\rvert$	Add-on Reward
Half Cheetah	Prior Policy	1000.7 ± 88.8	2449.8 ± 52.3	0.14 ± 0.00	-1449.1 ± 45.3
	Greedy-MPPI	$\mathbf{1939.9 \pm 134.7}$	2180.9 ± 87.3	$\mathbf{0.02 \pm 0.01}$	$\mathbf{-241.0 \pm 50.3}$
	Full-MPPI	-3595.1 ± 322.7	-1167.3 ± 144.0	0.24 ± 0.03	-2427.7 ± 320.3
	Guided-MPPI	1849.6 ± 151.0	2154.6 ± 95.7	0.03 ± 0.01	-305.0 ± 58.7
	Valued-MPPI	1760.7 ± 478.8	$\mathbf{2201.8 \pm 258.3}$	0.04 ± 0.02	-441.0 ± 222.5
	Residual-MPPI	$\mathbf{1936.2 \pm 109.3}$	2178.6 ± 71.9	$\mathbf{0.02 \pm 0.00}$	$\mathbf{-242.3 \pm 40.5}$
	Residual-SAC (200K)	-265.0 ± 919.0	455.4 ± 678.6	0.07 ± 0.03	720.4 ± 251.8
	Residual-SAC (4M)	2184.5 ± 29.7	2233.7 ± 29.3	0.00 ± 0.00	-49.2 ± 1.7
	Fulltask-SAC	2149.9 ± 28.6	2214.5 ± 27.2	0.01 ± 0.00	-64.5 ± 2.4
Env	Policy	Total Reward	Basic Reward	$\lvert\bar{\theta}\rvert$	Add-on Reward
Swimmer	Prior Policy	-245.2 ± 5.6	345.8 ± 3.2	0.59 ± 0.01	-591.0 ± 5.8
	Greedy-MPPI	$\mathbf{-58.9 \pm 5.4}$	275.8 ± 3.1	$\mathbf{0.33 \pm 0.01}$	$\mathbf{-334.7 \pm 7.4}$
	Full-MPPI	-1686.6 ± 106.7	14.1 ± 6.3	1.70 ± 0.11	-1700.7 ± 106.2
	Guided-MPPI	-149.0 ± 5.6	292.9 ± 3.8	0.44 ± 0.01	-441.9 ± 7.2
	Valued-MPPI	-205.8 ± 6.3	$\mathbf{335.1 \pm 1.6}$	0.54 ± 0.01	-540.9 ± 6.3
	Residual-MPPI	$\mathbf{-60.0 \pm 5.2}$	275.8 ± 3.4	$\mathbf{0.34 \pm 0.01}$	$\mathbf{-335.9 \pm 7.6}$
	Residual-SAC (200K)	-209.0 ± 67.6	2.1 ± 15.5	0.21 ± 0.07	-221.1 ± 72.7
	Residual-SAC (4M)	-10.5 ± 24.1	-1.5 ± 16.9	0.01 ± 0.02	-9.0 ± 16.6
	Fulltask-SAC	-4.2 ± 17.1	2.1 ± 17.6	0.01 ± 0.00	-6.3 ± 3.0

Env.	Policy	Total Reward	Basic Reward	$\bar{z}$	Add-on Reward
Hopper	Prior Policy	7252.7 ± 49.2	3574.5 ± 9.7	1.37 ± 0.00	3678.2 ± 48.3
	Greedy-MPPI	$\mathbf{7367.0 \pm 199.4}$	3553.0 ± 58.4	$\mathbf{1.38 \pm 0.01}$	$\mathbf{3814.0 \pm 156.8}$
	Full-MPPI	20.5 ± 3.0	3.6 ± 0.7	1.24 ± 0.00	16.9 ± 2.4
	Guided-MPPI	6121.3 ± 1590.1	3067.8 ± 679.0	1.35 ± 0.03	3053.4 ± 917.7
	Valued-MPPI	7243.9 ± 75.7	$\mathbf{3562.7 \pm 14.5}$	1.37 ± 0.01	3681.2 ± 74.6
	Residual-MPPI	$\mathbf{7363.0 \pm 254.9}$	3547.6 ± 78.0	$\mathbf{1.38 \pm 0.01}$	$\mathbf{3815.4 \pm 186.4}$
	Residual-SAC (200K)	3543.1 ± 478.9	1019.8 ± 94.3	1.27 ± 0.01	2523.2 ± 405.5
	Residual-SAC (4M)	7682.5 ± 178.2	2310.4 ± 106.8	1.54 ± 0.01	5372.0 ± 75.8
	Fulltask-SAC	7825.3 ± 36.9	2934.5 ± 27.6	1.49 ± 0.00	4890.8 ± 39.6
Env	Policy	Total Reward	Basic Reward	$\bar{v}_y$	Add-on Reward
Ant	Prior Policy	6333.7 ± 753.9	6177.1 ± 703.7	0.16 ± 0.22	156.6 ± 200.5
	Greedy-MPPI	6104.2 ± 1532.0	5092.8 ± 1305.2	$\mathbf{1.01 \pm 0.27}$	$\mathbf{1011.3 \pm 277.7}$
	Full-MPPI	-2767.7 ± 154.0	-2764.4 ± 114.2	-0.00 ± 0.11	-3.3 ± 108.0
	Guided-MPPI	5160.9 ± 1963.0	4999.8 ± 1887.9	0.16 ± 0.22	161.2 ± 217.7
	Valued-MPPI	6437.0 ± 1021.9	$\mathbf{6230.7 \pm 959.0}$	0.21 ± 0.20	206.3 ± 196.3
	Residual-MPPI	$\mathbf{6846.7 \pm 647.8}$	5984.8 ± 541.5	$\mathbf{0.86 \pm 0.19}$	$\mathbf{861.8 \pm 189.8}$
	Residual-SAC (200K)	-1175.5 ± 157.3	-1178.3 ± 156.4	0.00 ± 0.00	2.7 ± 3.9
	Residual-SAC (4M)	6962.9 ± 342.9	5710.2 ± 252.0	1.25 ± 0.13	1252.7 ± 127.3
	Fulltask-SAC	7408.6 ± 312.0	3100.3 ± 184.4	4.31 ± 0.21	4308.3 ± 209.2

Residual-MPPI is Effective and Data-efficient

Experiments

GTS Customization

Table 2: Experimental Results of Residual-MPPI in GTS

Policy	GT Sophy 1.0	Zero-shot MPPI	Few-shot MPPI	Residual-SAC (80K laps)
Lap Time	117.77 ± 0.08	123.34 ± 0.22	122.93 ± 0.14	130.00 ± 0.13
Off-course Steps	93.13 ± 1.98	9.03 ± 3.33	4.43 ± 2.39	0.87 ± 0.78
Policy	**Full-MPPI**	**Guided-MPPI**	**Greedy-MPPI**	**Residual-SAC (2K laps)**
Lap Time	*Failed	*Failed	*Failed	*Failed
Off-course Steps	*Failed	*Failed	*Failed	*Failed

The evaluation results are in the form of mean $\pm$ std over 30 laps. *Failed baseline is not able to finish a complete lap. Valued-MPPI is not available since we only have access to the policy network of GT Sophy 1.0.

Residual-MPPI works in Complex Environment and Policy

Experiments

GTS Demo

- Safer Driving Style
- Advanced Route Selection

Takeaway

Online Principled Customization

= Residual-MPPI

+ Dynamics

+ Add-on Reward

THANK YOU

MSC laboratory

Berkeley
UNIVERSITY OF CALIFORNIA

Paper & Code