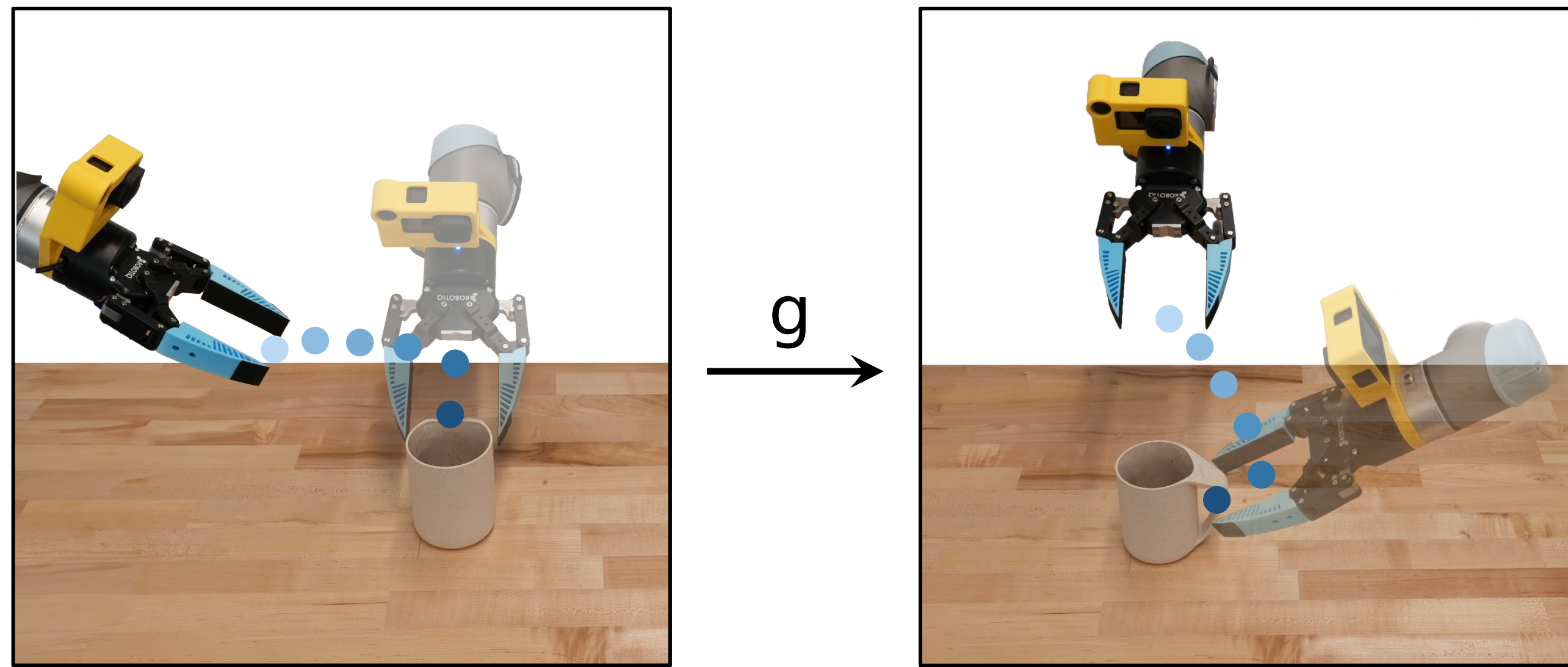


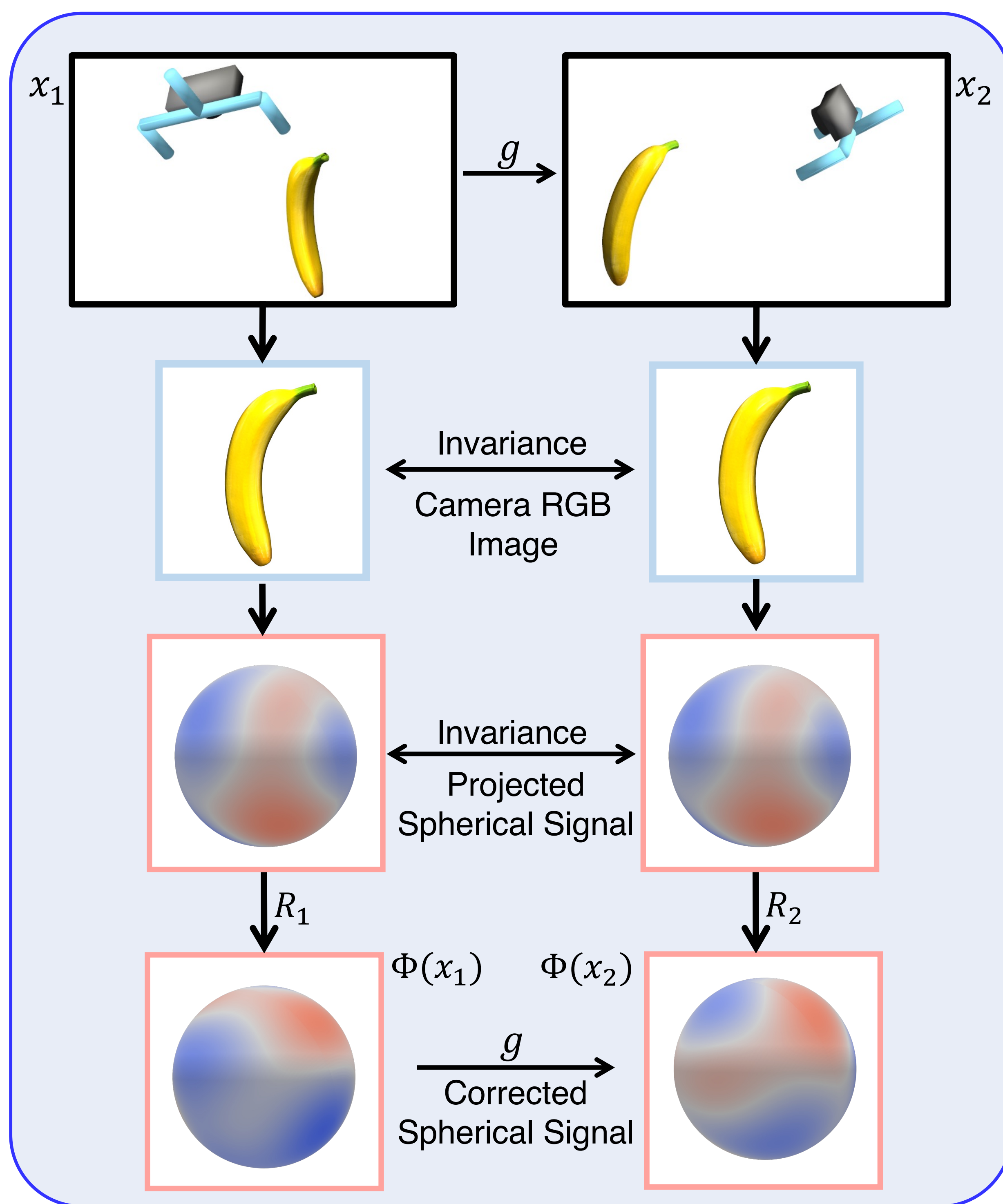


Motivation

Can we achieve $SO(3)$ -equivariance directly from monocular eye-in-hand RGB images to support data-efficient visuomotor policy learning?

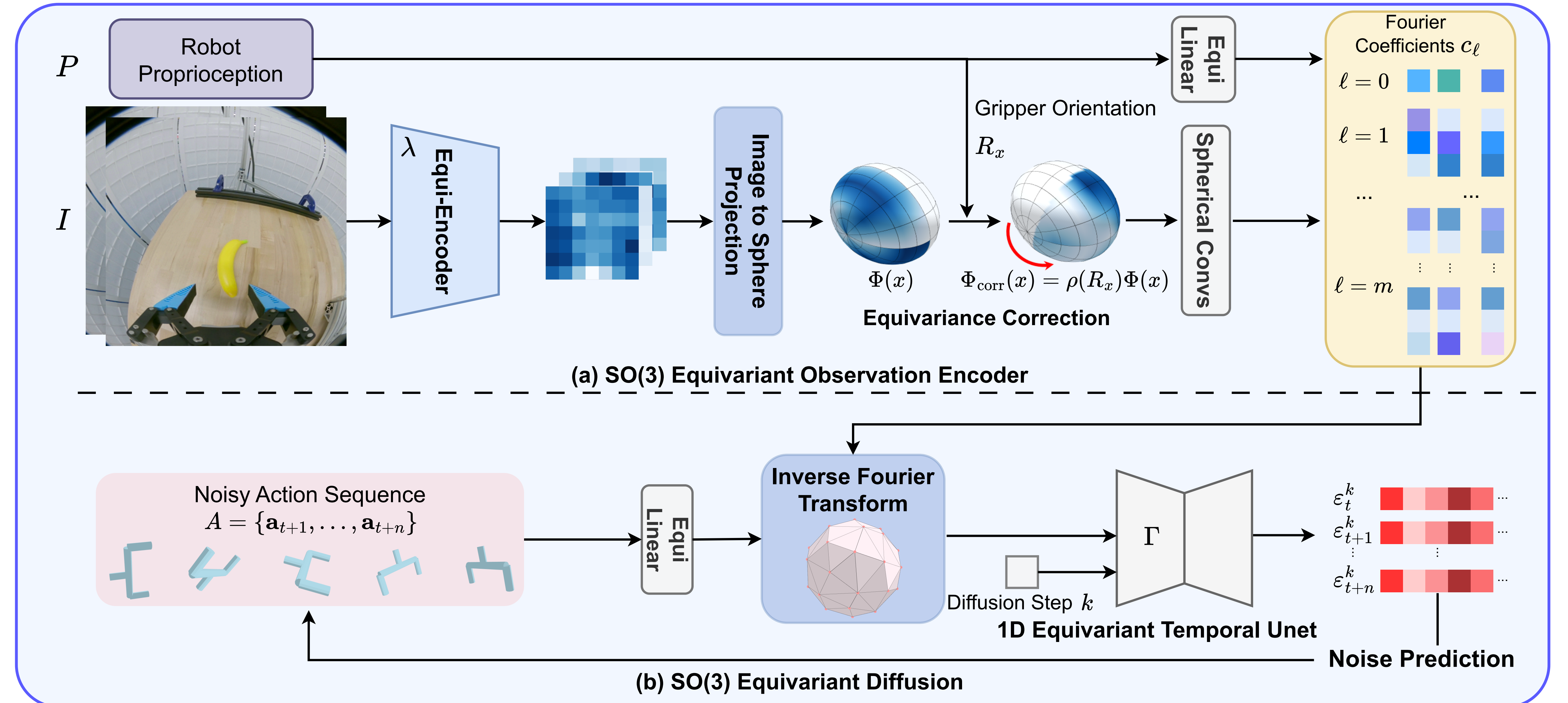


Equivariance Correction



By applying the gripper orientation R as an equivariance correction term, we align these spherical signals to a common world frame, ensuring their equivariant transformation under global scene rotations.

Overview of Image-to-Sphere Policy (ISP)



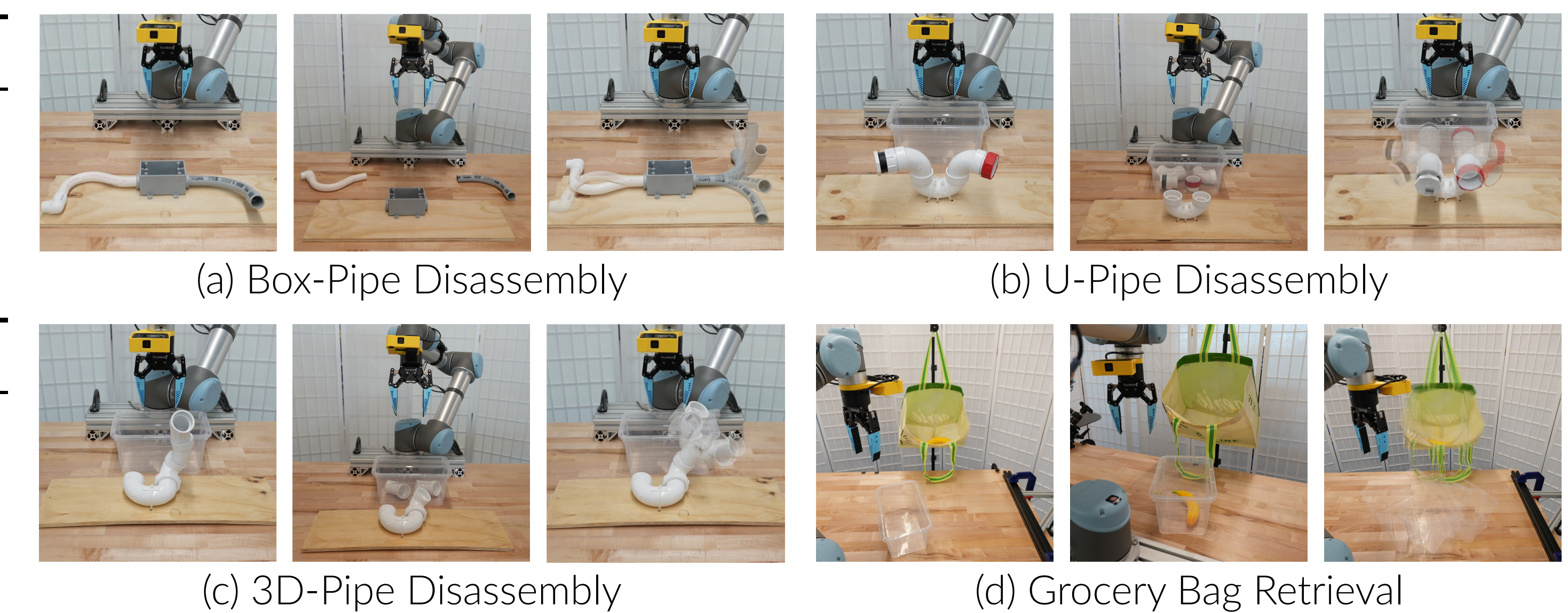
The observation encoder uses spherical projection to map image-extracted features onto a hemisphere and applies spherical convolutions to ensure $SO(3)$ -equivariance, producing the conditioning vector for the diffusion process. The diffusion module is designed as an $SO(3)$ -equivariant function of the conditioning vectors and noisy inputs.

Simulation Results and Robot Experiments

Method	Mean	Stack D1	Stack Three D1	Square D2	Threading D0
ISP (SO(3))	65.2 (+11.6)	99 (+3)	70 (+9)	35 (+23)	90 (+1)
ISP (SO(2))	65.0 (+11.4)	98 (+2)	75 (+14)	32 (+20)	85 (-4)
DiffPo	53.6	91	43	12	77
EquiDiff	53.0	96	61	9	89
ACT	23.0	45	12	3	36

Method	Coffee D2	Three Pc. D0	Hammer Cleanup D1	Mug Cleanup D1
ISP (SO(3))	64 (+11)	71 (-3)	66 (+7)	54 (+3)
ISP (SO(2))	59 (+6)	75 (+1)	71 (+12)	56 (+5)
DiffPo	53	73	59	49
EquiDiff	47	74	59	51
ACT	21	28	35	25

Method	Kitchen D1	Nut Assembly D0	Pick Place D0	Coffee Preparation D1
ISP (SO(3))	75 (+14)	75 (+24)	42 (+6)	41 (+4)
ISP (SO(2))	65 (+4)	74 (+23)	46 (+10)	47 (+10)
DiffPo	61	51	36	37
EquiDiff	55	40	28	27
ACT	21	37	9	8



	Box-Pipe Disassembly	U-Pipe Disassembly	3D-Pipe Disassembly	Grocery Bag Retrieval
# Demos	65	65	65	60
ISP (SO(3))	80% (16/20)	85% (17/20)	75% (15/20)	95% (19/20)
DiffPo	10% (2/20)	65% (13/20)	15% (3/20)	75% (15/20)

Table 1. Performance comparison between ISP and baselines trained with 100 demonstrations on the MimicGen tasks. Table 2. Comparison of ISP and DiffPo in real-world tasks over 20 trials.