Highlight

YONSEI UNIVERSITY · Berkeley UNIVERSITY OF CALIFORNIA · Massachusetts Institute of Technology

Diffusion-EDFs: Bi-equivariant Denoising Generative Modeling on SE(3) for Visual Robotic Manipulation

Hyunwoo Ryu[1], Jiwoo Kim[1], Hyunseok An[1], Junwoo Chang[1], Joohwan Seo[2], Taehan Kim[3], Yubin Kim[4], Chaewon Hwang[1], Jongeun Choi[1,2], Roberto Horowitz[2]

[1]Yonsei University, [2]University of California, Berkeley, [3]Samsung Research, [4]Massachusetts Institute of Technology

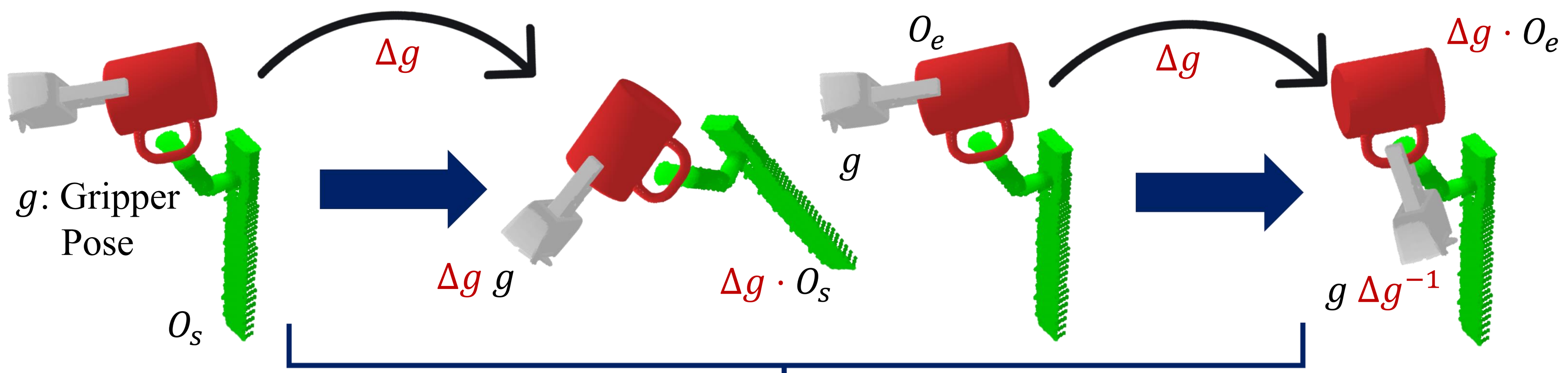
Project Page

CVPR JUNE 17-21, 2024 SEATTLE, WA

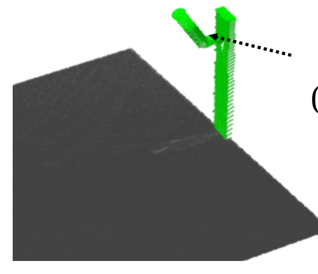
SE(3)-Equivariant Robotic Manipulation

Left-equivariance: gripper follows the transformation of the placement target

Right-equivariance: gripper compensates the transformation of the grasp



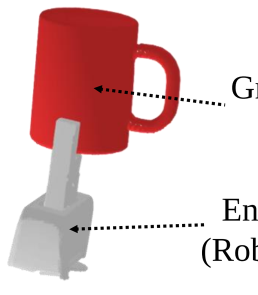
Bi-equivariance



O_s: Scene Point Cloud

Benefits of SE(3)-equivariance in Robot Learning

- Data-efficient (Only 5~10 demonstrations are enough)
- Generalizable (Previously unseen poses, instances, distractors)



O_e: Grasp Point Cloud

Theory: Bi-equivariant Diffusion Process on SE(3)

$$P_t(g_t|O_s,O_e) = \int dg_0\ P_{t|0}(g_t|g_0,O_s,O_e)P_0(g_0|O_s,O_e)$$

$P_{t|0}(g_t|g_0,O_s,O_e)$: Diffusion Kernel

$P_t(g_t|O_s,O_e)$: Noised Marginal Distribution $g_t \in SE(3)$: Diffused Pose

$P_0(g_0|O_s,O_e)$: Target Distribution $g_0 \in SE(3)$: Target Pose

Bi-equivariance of Probability Distribution Function (PDF) on SE(3)

$$P(g|O_s,O_e) = P(\Delta g\ g|\Delta g\cdot O_s,O_e) = P(g\ \Delta g^{-1}|O_s,\Delta g\cdot O_e) \quad \forall g,\Delta g \in SE(3)$$

Core Equation: Bi-equivariance of SE(3) Score Functions

If $P(g|O_s,O_e)$ is bi-equivariant, the following hold for $\boldsymbol{s}(g|O_s,O_e) = \nabla_{SE(3)}\log P(g|O_s,O_e)$

1. Left Invariance: $\boldsymbol{s}(\Delta g\ g|\Delta g\cdot O_s,O_e) = \boldsymbol{s}(g|O_s,O_e)$
2. Right Equivariance: $\boldsymbol{s}(g\ \Delta g^{-1}|O_s,\Delta g\cdot O_e) = [\mathrm{Ad}_{\Delta g}]^{-T}\boldsymbol{s}(g|O_s,O_e)$

$$Ad_g^{-T} = \begin{bmatrix} R & \emptyset \\ [\boldsymbol{p}]^\wedge R & R\end{bmatrix}$$

$$[\boldsymbol{p}]^\wedge = \begin{bmatrix} 0 & -p_3 & p_2 \\ p_3 & 0 & -p_1 \\ -p_2 & p_1 & 0\end{bmatrix}$$

$\boldsymbol{s}(g|O_s,O_e) = \nabla_{SE(3)}\log P(g|O_s,O_e)$: Score Function of P

$\nabla = (\nabla_{\mathbb{R}^3}, \nabla_{SO(3)})$

$\nabla_G = (\mathcal{L}_1,\mathcal{L}_2,\cdots,\mathcal{L}_{\dim(G)})$: Lie-derivatives of group G

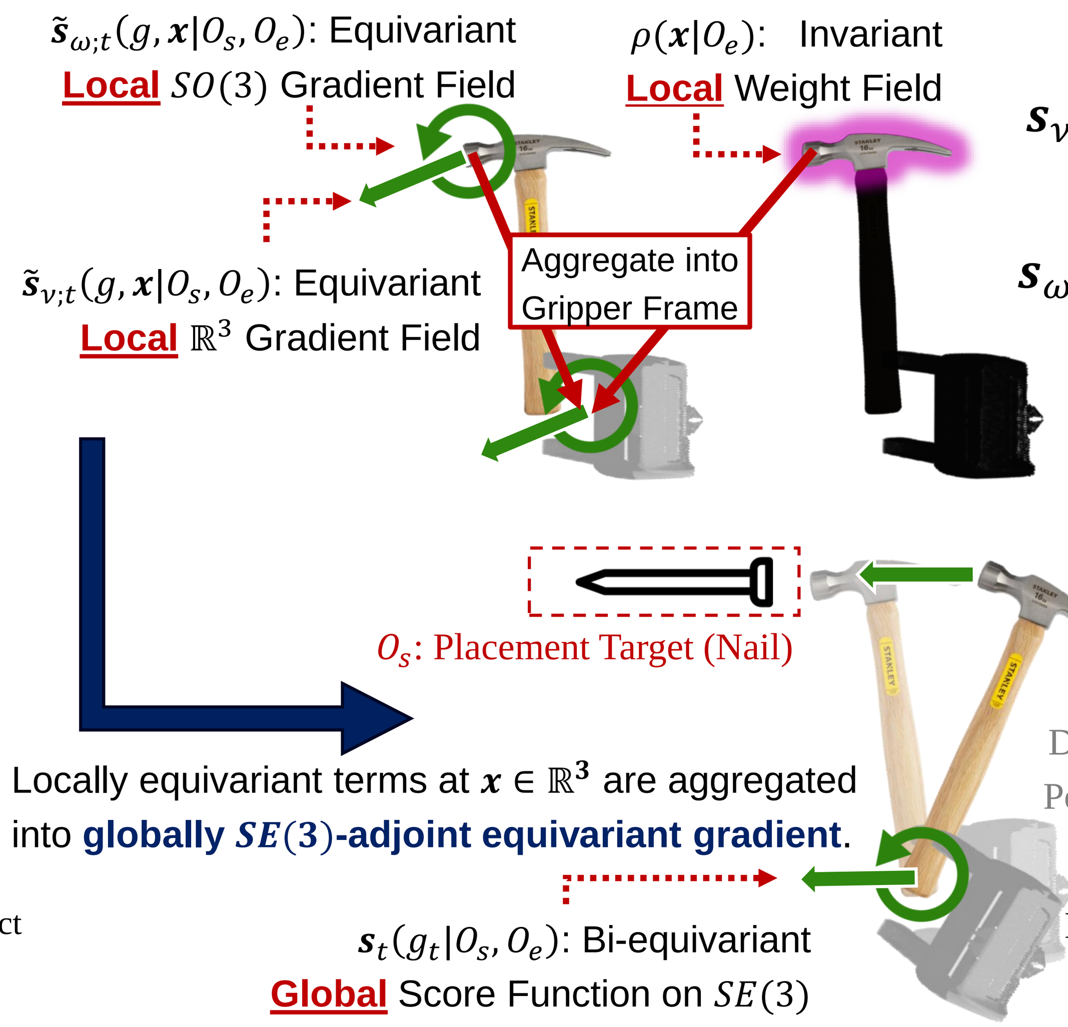
$\mathcal{L}_i$: Lie-derivative along i-th Lie algebra of G

Method: Bi-equivariant Score-based Model on SE(3)

$$\nabla_{SE(3)}\log P_t(g|O_s,O_e) = \boldsymbol{s}_t(g|O_s,O_e) = \begin{bmatrix}\boldsymbol{s}_{\nu;t}(g|O_s,O_e) \\ \boldsymbol{s}_{\omega;t}(g|O_s,O_e)\end{bmatrix}$$

Translational gradient

Rotational gradient



Locally equivariant terms at $\boldsymbol{x}\in\mathbb{R}^3$ are aggregated into **globally $SE(3)$-adjoint equivariant gradient.**

$\boldsymbol{s}_t(g_t|O_s,O_e)$: Bi-equivariant **Global** Score Function on $SE(3)$

$$\boldsymbol{s}_{\nu;t}(g|O_s,O_e) = \int_{\mathbb{R}^3} d^3\boldsymbol{x}\ \rho(\boldsymbol{x}|O_e)\tilde{\boldsymbol{s}}_{\nu;t}(g,\boldsymbol{x}|O_s,O_e)$$

Spin Term

$$\boldsymbol{s}_{\omega;t}(g|O_s,O_e) = \int_{\mathbb{R}^3} d^3\boldsymbol{x}\ \rho(\boldsymbol{x}|O_e)\tilde{\boldsymbol{s}}_{\omega;t}(g,\boldsymbol{x}|O_s,O_e)$$

Orbital Term

$$+\int_{\mathbb{R}^3} d^3\boldsymbol{x}\ \rho(\boldsymbol{x}|O_e)\ \boldsymbol{x}\wedge\tilde{\boldsymbol{s}}_{\nu;t}(g,\boldsymbol{x}|O_s,O_e)$$

Cross Product

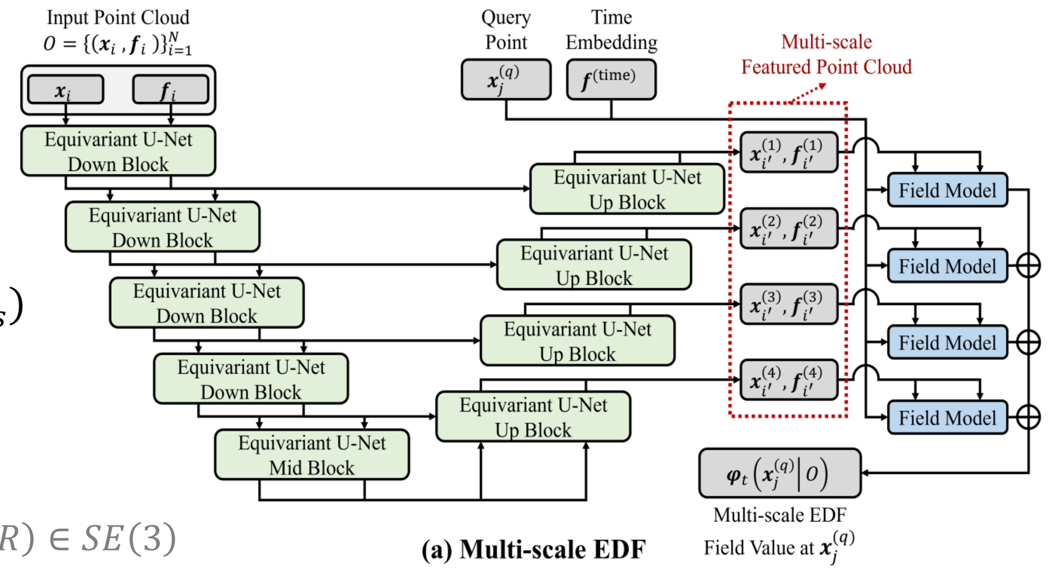
***Remark:** A Purely translational term at the local regions could result in a global rotational effect at the gripper frame through the **orbital term**.

Model Architecture

- We use $SE(3)$-steerable neural fields to implement bi-equivariant score model for $\square = \omega, \nu$

$$\tilde{\boldsymbol{s}}_{\square;t}(g,\boldsymbol{x}|O_s,O_e) = \boldsymbol{\psi}_{\square;t}(\boldsymbol{x}|O_e)\otimes_{\square;t}^{(\to 1)}\mathbf{D}(R^{-1})\boldsymbol{\varphi}_{\square;t}(g\boldsymbol{x}|O_s)$$

- A multiscale architecture based on $SE(3)$-equivariant GNNs are used to encode point cloud observations.



(a) Multi-scale EDF

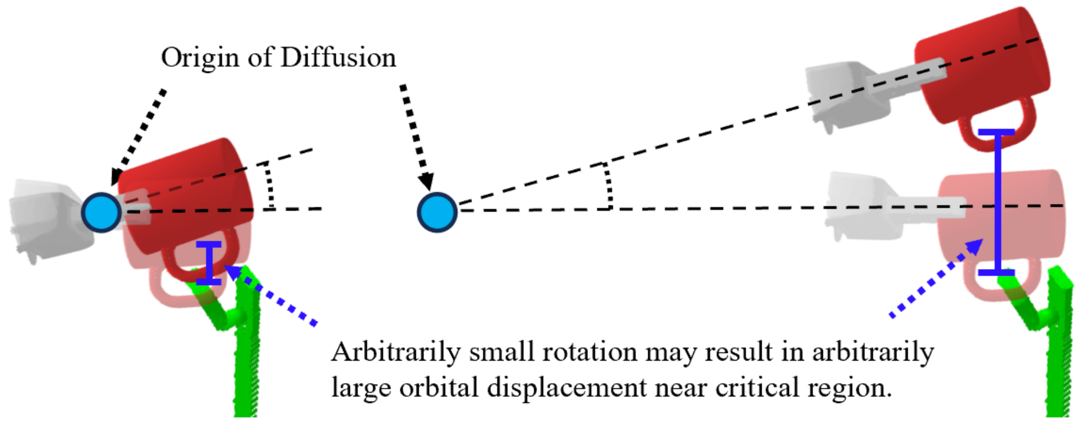
Steerability: $\varphi(\Delta g\ \boldsymbol{x}|\Delta g\cdot O) = \mathbf{D}(\Delta R)\varphi(\boldsymbol{x}|O) \quad \forall\Delta g = (\Delta\boldsymbol{p},\Delta R)\in SE(3)$

$\mathbf{D}(R)$: Block-diagonal of Real Wigner D-matrices; $\otimes^{(\to 1)}$: Row-rank Clebsch-Gordon Tensor Product to a Spin-1 Tensor

Bi-equivariant Diffusion via Contact-based Origin Selection

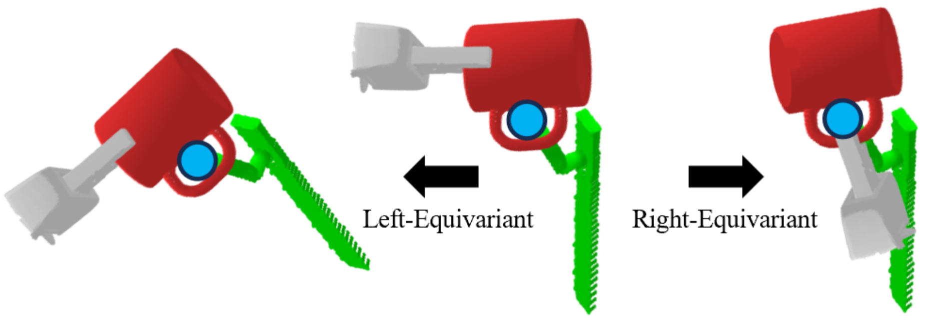
- There is no bi-invariant kernel on $SE(3)$ such that $P_{t|0}(g_t|g_o) = P_{t|0}(\Delta g g_t|\Delta g g_o) = P_{t|0}(g_t\Delta g|g_o\Delta g)\ \forall\ \Delta g$
- Therefore, $P_{t|0}$ must have input dependency; it should depend on either O_s or O_g
- We achieve bi-equivariance by diffusing with Brownian noise but in an equivariantly sampled diffusion origin.

(a) Diffusion on SE(3) without Origin Selection



Arbitrarily small rotation may result in arbitrarily large orbital displacement near critical region.

(b) Contact-based Diffusion Origin Selection



Our bi-equivariant diffusion origin selection mechanism solves this problem by selecting the origin near **contact-rich regions**.

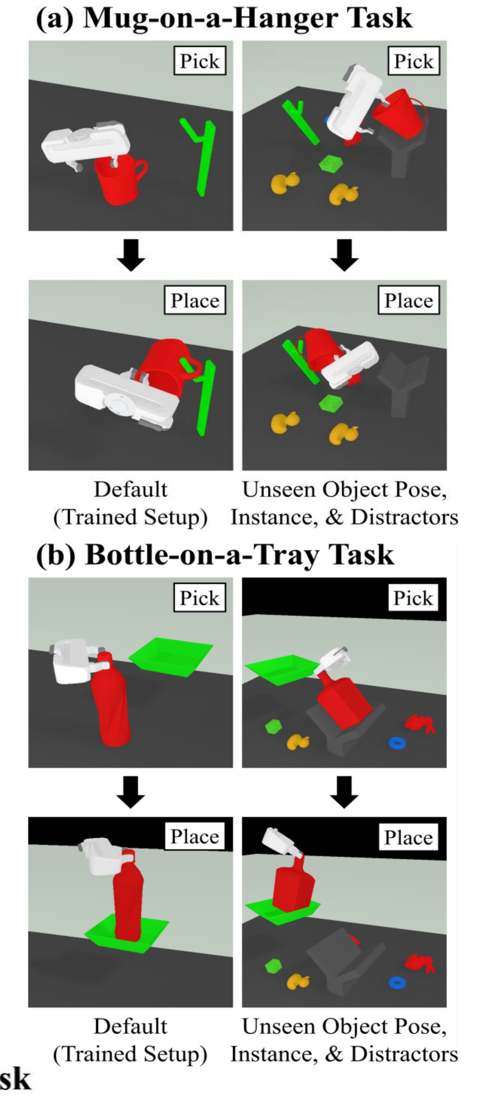
Simulation Benchmark Results

Scenario	Method	Without Pretraining	Without Obj. Seg.	Without Rot. Aug.	Mug Pick	Mug Place	Mug Total	Bottle Pick	Bottle Place	Bottle Total
Default (Trained Setup)	R-NDFs [68]	✗	✗	✓	0.83	**0.97**	0.81	0.91	0.73	0.67
		✗	✓	✓	0.00	0.00	0.00	0.00	0.00	0.00
	SE(3)-DiffusionFields [75]	✓	✗	✗	0.75	(n/a)	(n/a)	0.47	(n/a)	(n/a)
		✓	✓	✗	0.11	(n/a)	(n/a)	0.01	(n/a)	(n/a)
	Diffusion-EDFs (Ours)	✓	✓	✓	**0.99**	0.96	**0.95**	**0.97**	**0.85**	**0.83**
Previously Unseen Instances	R-NDFs [68]	✗	✗	✓	0.73	0.70	0.51	0.90	0.87	0.79
		✗	✓	✓	0.00	0.00	0.00	0.00	0.00	0.00
	SE(3)-DiffusionFields [75]	✓	✗	✗	0.55	(n/a)	(n/a)	0.57	(n/a)	(n/a)
		✓	✓	✗	0.14	(n/a)	(n/a)	0.00	(n/a)	(n/a)
	Diffusion-EDFs (Ours)	✓	✓	✓	**0.96**	**0.96**	**0.92**	**0.99**	**0.91**	**0.90**
Previously Unseen Poses	R-NDFs [68]	✗	✗	✓	0.84	0.93	0.78	0.65	0.72	0.47
		✗	✓	✓	0.00	0.00	0.00	0.00	0.00	0.00
	SE(3)-DiffusionFields [75]	✓	✗	✗	0.75	(n/a)	(n/a)	0.47	(n/a)	(n/a)
		✓	✓	✗	0.00	(n/a)	(n/a)	0.04	(n/a)	(n/a)
	Diffusion-EDFs (Ours)	✓	✓	✓	**0.98**	**0.98**	**0.96**	**0.98**	**0.81**	**0.79**
Previously Unseen Clutters§	R-NDFs [68]	✗	✓	✓	0.00	0.00	0.00	0.00	0.00	0.00
	SE(3)-DiffusionFields [75]	✓	✓	✗	0.06	(n/a)	(n/a)	0.03	(n/a)	(n/a)
	Diffusion-EDFs (Ours)	✓	✓	✓	**0.91**	**1.00**	**0.91**	**0.96**	**0.91**	**0.87**
Previously Unseen Instances, Poses, & Clutters§	R-NDFs [68]	✗	✗	✓	0.71§	0.75§	0.53§	0.85§	0.84§	0.72§
		✗	✓	✓	0.00	0.00	0.00	0.00	0.00	0.00
	SE(3)-DiffusionFields [75]	✓	✗	✗	0.58§	(n/a)	(n/a)	0.59§	(n/a)	(n/a)
		✓	✓	✗	0.03	(n/a)	(n/a)	0.00	(n/a)	(n/a)
	Diffusion-EDFs (Ours)	✓	✓	✓	**0.89**	**0.89**	**0.79**	**0.98**	**0.89**	**0.87**

§Models with segmented inputs are tested without cluttered objects to guarantee perfect object segmentation.

(a) Mug-on-a-Hanger Task

(b) Bottle-on-a-Tray Task



Real Robot Experiments

We evaluate Diffusion-EDFs on three real-world scenarios:

1. Picking a mug and placing on a hanger.
2. Picking up bowls and placing on the dish with the matching color, in red-green-blue order.
3. Picking one of the bottles and placing it on a shelf, until there is no bottle left.

End-to-end trained from scratch with **only 10 human demo**.

***Table: Core challenges of each task**

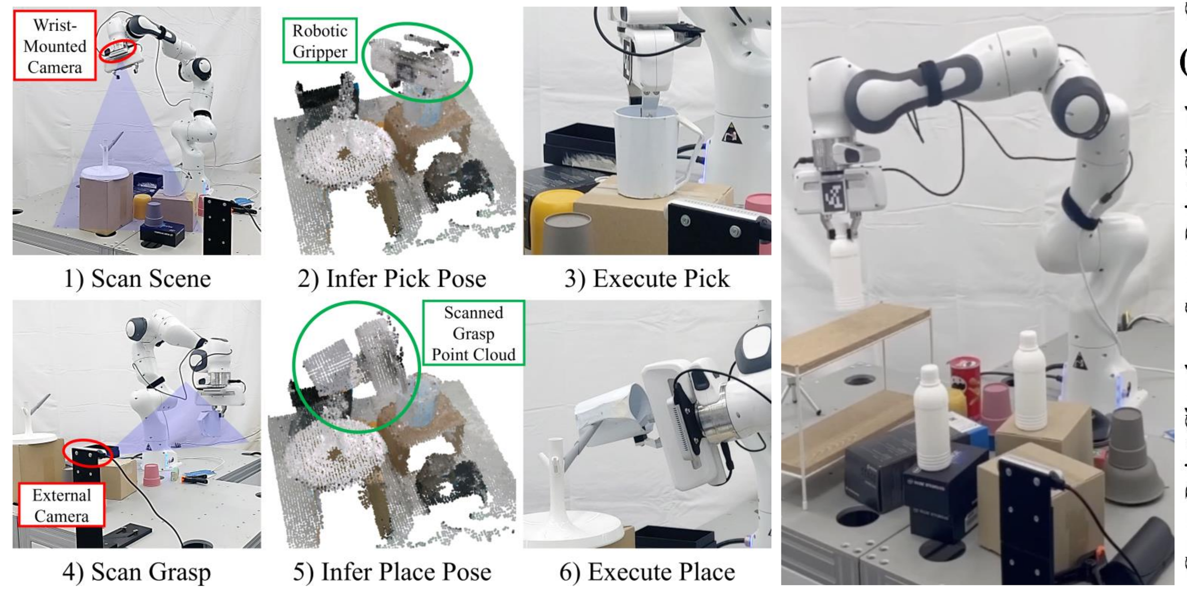
Mug-on-a-hanger	Bowls-on-dishes	Bottles-on-a-shelf
Accurate 6-DoF inference	Sequential problem	Multimodal distribution
Unseen object pose	Scene-level understanding	Variable object number
Unseen object instance	Color-critical	Unseen object instance

(a) Mug-on-a-Hanger Task

(b) Bowls-on-Dishes Task (in Red-Green-Blue Order)

(c) Bottles-on-a-Shelf Task

Real Hardware Experiment Outline



1) Scan Scene 2) Infer Pick Pose 3) Execute Pick 4) Scan Grasp 5) Infer Place Pose 6) Execute Place