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

Hyunwoo Ryu[1]*, Jiwoo Kim[1], Junwoo Chang[1], Hyun Seok Ahn[1], Joohwan Seo[2]
Taehan Kim[3], Yubin Kim[4], Chaewon Hwang[5,6], Jongeun Choi[1,2]†, Roberto Horowitz[2]
[1]Yonsei University, [2]University of California, Berkeley, [3]Samsung Research,
[4]Massachusetts Institute of Technology, [5]Ewha Womans University, [6]Work done at Yonsei University

Abstract

Diffusion generative modeling has become a promising approach for learning robotic manipulation tasks from stochastic human demonstrations. In this paper, we present Diffusion-EDFs*, a novel $SE(3)$-equivariant diffusion-based approach for visual robotic manipulation tasks. We show that our proposed method achieves remarkable data efficiency, requiring only 5 to 10 human demonstrations for effective end-to-end training in less than an hour. Furthermore, our benchmark experiments demonstrate that our approach has superior generalizability and robustness compared to state-of-the-art methods. Lastly, we validate our methods with real hardware experiments. Project Website:* https://sites.google.com/view/diffusion-edfs

1. Introduction

Diffusion models are increasingly being recognized as superior methods for modeling stochastic and multimodal policies [1, 3, 5, 6, 9, 26, 34, 36, 39, 48, 51]. However, these methods require numerous demonstrations and do not generalize well on novel task configurations that are not provided during training. On the contrary, equivariant methods are well known for their data efficiency and generalizability in learning robotic manipulation tasks [5, 6, 13, 23, 24, 27, 33, 42, 46, 47, 52, 56]. Specifically, $SE(3)$-equivariant models have been proven to be effective for 6-DoF manipulation tasks learning from vision [13, 24, 42, 46, 47]. *Equivariant Descriptor Fields* (EDFs) [42] achieve data-efficient end-to-end learning on 6-DoF visual robotic manipulation tasks by employing $SE(3)$ *bi-equivariant* [27, 42] energy-based models. However, training energy-based models like EDFs is typically very slow.

In this paper, we present Diffuion-EDFs, a diffusion-based alternative to EDFs with a significantly reduced training time ($\times$15 faster). Similarly to EDFs, we exploit the bi-equivariance (see Supp. A) and locality of robotic manipulation tasks in our method design. This enables our method to be trained end-to-end from only 5$\sim$10 human demonstrations without requiring any pre-training and object segmentation, yet are highly generalizable to out-of-distribution object configurations. We validate Diffusion-EDFs through simulation and real-robot experiments.

*First Author: tomato1mule@yonsei.ac.kr
†Corresponding Author: jongeunchoi@yonsei.ac.kr

2. Preliminaries

2.1. SO(3) Group Representation Theory

A *representation* $\mathbf{D}(g)$ is a map from a group $\mathcal{G}$ to a linear map on a vector space $\mathcal{W}$ that satisfies

$$\mathbf{D}(g)\mathbf{D}(h) = \mathbf{D}(gh) \quad \forall g, h \in \mathcal{G} \tag{1}$$

It is known that any representation of the special orthogonal group $SO(3)$ can be block-diagonalized into *irreducible representations*, which can be classified according to their angular frequency $l \in \{0, 1, 2, \dots\}$. The *real Wigner D-matrix* of degree l, denoted as $\mathbf{D}_l(R) : SO(3) \to \mathbb{R}^{(2l+1)\times(2l+1)}$ is a commonly used irreducible orthogonal representation of angular frequency l. The vectors transformed according to $\mathbf{D}_l(R)$ are called *type-l* vectors [20].

2.2. Equivariant Descriptor Fields

An Equivariant Descriptor Field (EDF) [42] $\boldsymbol{\varphi}(\boldsymbol{x}|O)$ is an $SO(3)$-equivariant and translation-invariant vector field on $\mathbb{R}^3$ generated by a point cloud $o \in \mathcal{O}$. EDFs are decomposed into the direct sum of irreducible subspaces

$$\boldsymbol{\varphi}(\boldsymbol{x}|O) = \bigoplus_{n=1}^{N} \boldsymbol{\varphi}^{(n)}(\boldsymbol{x}|O) \tag{2}$$

where $\boldsymbol{\varphi}^{(n)}(\boldsymbol{x}|O) : \mathbb{R}^3 \times \mathcal{O} \to \mathbb{R}^{2l_n+1}$ is a translation-invariant type-l_n vector field generated by O. Therefore, an EDF $\boldsymbol{\varphi}(\boldsymbol{x}|O)$ is transformed according to $\Delta g = (\Delta \boldsymbol{p}, \Delta R) \in SE(3)$, $\Delta \boldsymbol{p} \in \mathbb{R}^3$, $\Delta R \in SO(3)$ as

$$\boldsymbol{\varphi}(\Delta g\, \boldsymbol{x}|\Delta g \cdot O) = \mathbf{D}(\Delta R)\boldsymbol{\varphi}(\boldsymbol{x}|O) \tag{3}$$

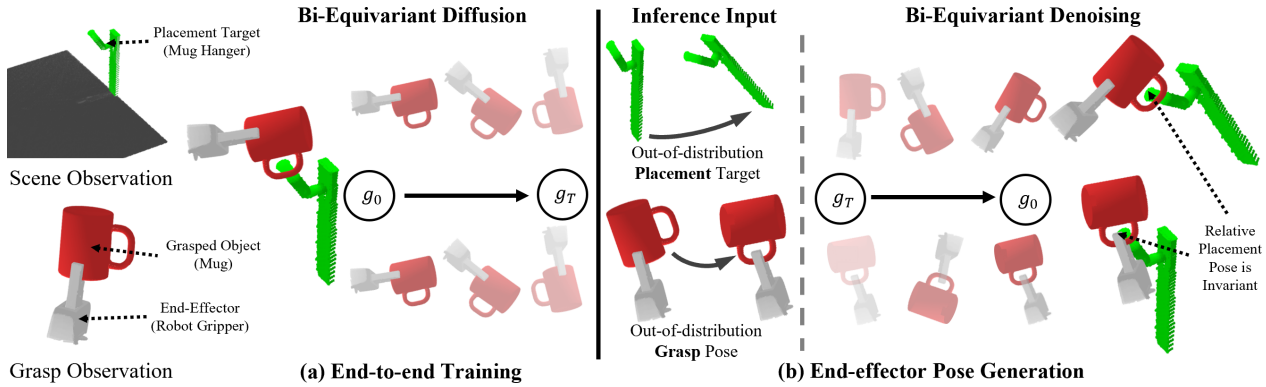

Figure 1. **Overview of Diffusion-EDFs. (a)** The target end-effector pose g_0 is bi-equivariantly diffused for the training of Diffusion-EDFs. **(b)** The end-effector pose is sampled from the policy by denoising with learned bi-equivariant score function. Due to the bi-equivariance, the trained policy can be effectively generalized to previously unseen configurations in the observation of the scene and the grasp.

where $\mathbf{D}(R)$ is the block-diagonal matrix whose sub-matrices are Wigner D-matrices $\{\mathbf{D}_{l_n}(R)\}_{n=1}^{n=N}$.

2.3. Brownian Diffusion on the SE(3) Manifold

Let $g_t \in SE(3)$ be generated by diffusing $g_0 \in SE(3)$ for time t. The Brownian diffusion process is defined by the following Lie group stochastic differential equation (SDE)

$$g_{t+dt} = g_t \exp\left[dW\right] \tag{4}$$

where dW is the standard Wiener process on $\mathfrak{se}(3)$ Lie algebra. The Brownian diffusion kernel $P_{t|0}(g_t|g_0) = \mathcal{B}_t(g_0^{-1}g_t)$ for the SDE in Eq. (4) can be decomposed into rotational and translational parts [15, 54] such that

$$\mathcal{B}_t(g) = \mathcal{N}(\boldsymbol{p}; \boldsymbol{\mu} = \mathbf{0}, \Sigma = tI)\, \mathcal{IG}_{SO(3)}(R; \epsilon = t/2) \tag{5}$$

$$\mathcal{IG}_{SO(3)}(R;\epsilon) = \sum_{l=0}^{\infty}(2l+1)e^{-\epsilon l(l+1)}\frac{\sin\left(l\theta+\frac{\theta}{2}\right)}{\sin\theta/2} \tag{6}$$

where $\mathcal{N}$ is the normal distribution on $\mathbb{R}^3$, $\mathcal{IG}_{SO(3)}$ is the isotropic Gaussian on $SO(3)$ [25, 31, 38, 43], $g = (\boldsymbol{p}, R) \in SE(3)$, $\boldsymbol{p} \in \mathbb{R}^3$, $R \in SO(3)$, and θ is the rotation angle of $SO(3)$ in the axis-angle parameterization.

2.4. Langevin Dynamics on the SE(3) Manifold

Let $\mathfrak{se}(3)$ be the Lie algebra that generates $SE(3)$. A *Lie derivative* $\mathcal{L}_{\mathcal{V}}$ along $\mathcal{V} \in \mathfrak{se}(3)$ of a differentiable function $f(g)$ on $SE(3)$ is defined as

$$\mathcal{L}_{\mathcal{V}} f(g) = \frac{d}{d\epsilon}\bigg|_{\epsilon=0} f(g\exp\left[\epsilon\mathcal{V}\right]) \tag{7}$$

The *Langevin dynamics* for a probability distribution $P(g)$ on $SE(3)$ is defined as follows [7, 11]:

$$g_{\tau+d\tau} = g_\tau \exp\left[\frac{1}{2}\nabla\log P(g)d\tau + dW\right] \tag{8}$$

$$\nabla\log P(g) = \sum_{i=1}^{6}\mathcal{L}_i \log P(g)\,\hat{\boldsymbol{e}}_i \tag{9}$$

where in the last line we denote the Lie derivative along the i-th basis $\hat{\boldsymbol{e}}_i \in \mathfrak{se}(3)$ as $\mathcal{L}_i$ instead of $\mathcal{L}_{\hat{\boldsymbol{e}}_i}$ for brevity. We denote the time for the Langevin dynamics as τ, as we reserve the notation t for the diffusion time. It is known that under mild assumptions, this process converges to $P(g)$ as $\tau \to \infty$ regardless of the initial distribution. Thus, one may sample from $P(g)$ with Langevin dynamics if the *score function* $\boldsymbol{s}(g) = \nabla\log P(g) : SE(3) \to \mathfrak{se}(3)$ is known.

3. Bi-equivariant Score Matching on the SE(3) Manifold

3.1. Problem Formulation

Let the target policy distribution be $P_0(g_0|o_s, o_e)$, where $g_0 \in SE(3)$ is the target end-effector pose, and o_s and o_e are the observed point clouds of the scene and the grasped object, respectively. Note that o_s is observed in the scene frame s, and o_e in the end-effector frame e. Following Ryu et al. [42], we model P_0 to be bi-equivariant (see Supp. A):

$$\begin{aligned} P_0(g|o_s, o_e) &= P_0(\Delta g\, g|\Delta g \cdot o_s, o_e) \\ &= P_0(g\Delta g^{-1}|o_s, \Delta g \cdot o_e) \end{aligned} \tag{10}$$

Now let $g_t \in SE(3)$ be the samples that are noised from g_0 by some diffusion process, where t denotes the diffusion time. Our goal is to train a model that denoises g_t, which is sampled from the diffused marginal distribution $P_t(g_t|o_s, o_e)$, into a denoised sample g, which follows the target distribution $P_0(g|o_s, o_e)$. This can be achieved with Annealed Langevin MCMC [4, 15, 22, 25, 49, 54] if the *score function* (see Sec. 2.4) of P_t is known. See Fig. 1 for the overview of Diffusion-EDFs.

3.2. Bi-equivariant Score Function

Let $\boldsymbol{s}(g|o_s, o_e) = \nabla\log P(g|o_s, o_e)$ be the score function of a probability distribution $P(g|o_s, o_e)$.

Proposition 1. *$\boldsymbol{s}(g|O_s, O_e)$ satisfies the following conditions for all $\Delta g \in SE(3)$ if $P(g|O_s, O_e)$ is bi-equivariant:*

$$\boldsymbol{s}(\Delta g\, g|\Delta g \cdot O_s, O_e) = \boldsymbol{s}(g|O_s, O_e) \tag{11}$$

$$\boldsymbol{s}(g\, \Delta g^{-1}|O_s, \Delta g \cdot O_e) = [\mathrm{Ad}_{\Delta g}]^{-T}\, \boldsymbol{s}(g|O_s, O_e) \tag{12}$$

Ad_g is the *adjoint representation* [11, 35, 37] of $SE(3)$ with $g = (\boldsymbol{p}, R)$, $\boldsymbol{p} \in \mathbb{R}^3$, and $R \in SO(3)$

$$Ad_g = \begin{bmatrix} R & [\boldsymbol{p}]^\wedge R \\ \emptyset & R \end{bmatrix} \tag{13}$$

where $[\boldsymbol{p}]^\wedge$ denotes the skew-symmetric 3×3 matrix of $\boldsymbol{p}$. See Supp. C.1 for the proof of Proposition 1.

3.3. Bi-equivariant Diffusion Process

Let the point cloud conditioned diffusion kernel under time t be $P_{t|0}(g|g_0, O_s, O_e)$ such that the diffused marginal $P_t(g|O_s, O_e)$ for $P_0(g|O_s, O_e)$ is defined as follows:

$$P_t(g|O_s, O_e) = \int_{SE(3)} dg_0\; P_{t|0}(g|g_0, O_s, O_e) P_0(g_0|O_s, O_e) \tag{14}$$

If the diffused marginal $P_t(g|O_s, O_e)$ is bi-equivariant, one may leverage Proposition 1 in the score model design.

Definition 1. *A bi-equivariant diffusion kernel $P_{t|0}$ is a square-integrable kernel that satisfies the following equations for all $\Delta g \in SE(3)$, except on a set of measure zero:*

$$\begin{aligned} P_{t|0}(g|g_0,\; O_s, O_e) &= P_{t|0}(\Delta g\, g|\Delta g\, g_0, \Delta g \cdot O_s, O_e) \\ &= P_{t|0}(g\, \Delta g^{-1}|g_0\, \Delta g^{-1}, O_s, \Delta g \cdot O_e) \end{aligned} \tag{15}$$

Proposition 2. *The diffused marginal P_t is guaranteed to be bi-equivariant for all bi-equivariant initial distribution P_0 if and only if the diffusion kernel $P_{t|0}$ is bi-equivariant.*

See Supp. C.2 for the proof of Proposition 2. Note that the Brownian diffusion kernel $P_{t|0}(g|g_0) = \mathcal{B}_t(g_0^{-1} g)$ in Eq. (5) is left invariant[1] but not right invariant[1], that is

$$\begin{aligned} &\forall\, \Delta g \in SE(3),\; P_{t|0}(\Delta g\, g|\Delta g\, g_0) = P_{t|0}(g|g_0) \\ &\exists\, \Delta g \in SE(3),\; P_{t|0}(g\, \Delta g^{-1}|g_0\, \Delta g^{-1}) \neq P_{t|0}(g|g_0) \end{aligned} \tag{16}$$

In fact, there exist no square-integrable kernel on $SE(3)$ that is bi-invariant[1] (see Supp. C.3). Therefore, a bi-equivariant diffusion kernel must be dependent on either O_s or O_e to absorb the left or right action of Δg.

To implement such bi-equivariant diffusion kernels, we propose using an equivariant *diffusion origin selection mechanism* $P(\boldsymbol{p}_{ed}|g_0^{-1} \cdot O_s, O_e)$ where $\boldsymbol{p}_{ed} \in \mathbb{R}^3$ is the origin where the diffusion process starts, such that

$$\begin{aligned} &P_{t|0}(g|g_0, O_s, O_e) \\ &= \int_{\mathbb{R}^3} d\boldsymbol{p}_{ed} P(\boldsymbol{p}_{ed}|g_0^{-1} \cdot O_s, O_e) \mathcal{B}_t(g_{ed}^{-1} g_0^{-1} g g_{ed}) \end{aligned} \tag{17}$$

where $g_{ed} = (\boldsymbol{p}_{ed}, I) \in SE(3)$ is a pure translation, and $\mathcal{B}_t$ is the Brownian kernel defined in Eq. (5). Note that $\boldsymbol{p}_{ed}$ is defined with respect to the end-effector frame e.

Proposition 3. *The diffusion kernel $P_{t|0}$ in Eq. (17) is bi-equivariant if the origin selection mechanism is equivariant*

$$\begin{aligned} &P\left(\boldsymbol{p}_{ed}|g_0^{-1} \cdot O_s, O_e\right) \\ &= P\left(\Delta g\, \boldsymbol{p}_{ed}|(\Delta g\, g_0^{-1}) \cdot O_s, \Delta g \cdot O_e\right) \end{aligned} \tag{18}$$

We provide the proof in Supp. C.4. A concrete realization of such equivariant diffusion origin selection mechanism $P(\boldsymbol{p}_{ed}|g_0^{-1} \cdot O_s, O_e)$ is discussed in Sec. D.1.

3.4. Score Matching Objectives

In contrast to Song and Ermon [49], Urain et al. [51], our diffusion kernel $P_{t|0}(g|g_0, O_s, O_e)$ in Eq. (17) is not the Brownian kernel. Still, the following mean squared error (MSE) loss can be used to train our score model $\boldsymbol{s}_t(g|O_s, O_e)$ without requiring the integration of Eq. (17):

$$\begin{aligned} \mathcal{J}_t &= \mathbb{E}_{g, g_0, g_{ed}, O_s, O_e}\left[J_t\right] \\ J_t &= \frac{1}{2}\left\|\boldsymbol{s}_t(g|O_s, O_e) - \nabla \log \mathcal{B}_t(g_{ed}^{-1} g_0^{-1} g g_{ed})\right\|^2 \end{aligned} \tag{19}$$

where $g_0 \sim P_0(g_0|O_s, O_e)$, $g_{ed} = (\boldsymbol{p}_{ed}, I) \sim P(\boldsymbol{p}_{ed}|g_0^{-1} \cdot O_s, O_e)$, and $g \sim P_{t|0}(g|g_0, O_s, O_e)$. We optimize $\mathcal{J}_t$ for sampled reference frame g_{ed} and diffusion time t. The minimizer of $\mathcal{J}_t$ is neither $\nabla \log K_t$ nor $\nabla \log P_{t|0}$ but the score function of the diffused marginal $\nabla \log P_t$, that is

$$\underset{\boldsymbol{s}_t(g|O_s, O_e)}{\arg\min}\; \mathcal{J}_t = \boldsymbol{s}_t^*(g|O_s, O_e) = \nabla \log P_t(g|O_s, O_e) \tag{20}$$

Although Eq. (20) is a straightforward adaptation of the MSE minimizer formula [49], we still provide the derivation in Supp. C.5 for completeness. While autograd packages can be used for the computation of $\nabla \log \mathcal{B}_t(g_{ed}^{-1} g_0^{-1} g g_{ed})$ [15, 25, 31, 42, 51, 54], we provide a more stable explicit form in Supp. B.

3.5. Bi-equivariant Score Model

We split our score model $\boldsymbol{s}_t(\cdot|O_s, O_e) : SE(3) \to \mathfrak{se}(3) \cong \mathbb{R}^6$ into the direct sum of translational and rotational parts

$$\boldsymbol{s}_t(g|O_s, O_e) = \left[\boldsymbol{s}_{\nu;t} \oplus \boldsymbol{s}_{\omega;t}\right](g|O_s, O_e) \tag{21}$$

where we denote the translational part with subscript ν and rotational part with subscript ω. Thus, $\boldsymbol{s}_{\nu;t}(\cdot|O_s, O_e) : SE(3) \to \mathbb{R}^3$ is the translational score and $\boldsymbol{s}_{\omega;t}(\cdot|O_s, O_e) : SE(3) \to \mathfrak{so}(3) \cong \mathbb{R}^3$ is the rotational score. To satisfy the

[1] We use the term *invariance* instead of equivariance since the kernel is neither conditioned by O_s nor O_e.

Scenario	Method	Without Pretraining	Without Obj. Seg.	Without Rot. Aug.	Mug			Bottle		
					Pick	Place	**Total**	Pick	Place	**Total**
Default (Trained Setup)	R-NDFs [47]	✗	✗	✓	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 [51]	✓	✗	✗	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, Poses, & Clutters[§]	R-NDFs [47]	✗	✗	✓	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 [51]	✓	✗	✗	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.

Table 1. Pick-and-place success rates for trained settings and out-of-distribution settings in simulated environment.

equivariance conditions in Eq. (11) and Eq. (12), we propose the following models:

$$\boldsymbol{s}_{\nu;t}(g|o_s,o_e) = \int_{\mathbb{R}^3} d^3\boldsymbol{x}\ \rho_{\nu;t}(\boldsymbol{x}|o_e)\ \widetilde{\boldsymbol{s}}_{\nu;t}(g,\boldsymbol{x}|o_s,o_e) \tag{22}$$

$$\begin{aligned}\boldsymbol{s}_{\omega;t}(g|o_s,o_e) = &\underbrace{\int_{\mathbb{R}^3} d^3\boldsymbol{x}\ \rho_{\omega;t}(\boldsymbol{x}|o_e)\ \widetilde{\boldsymbol{s}}_{\omega;t}(g,\boldsymbol{x}|o_s,o_e)}_{\text{Spin term}} \\ &+\underbrace{\int_{\mathbb{R}^3} d^3\boldsymbol{x}\ \rho_{\nu;t}(\boldsymbol{x}|o_e)\ \boldsymbol{x}\wedge\widetilde{\boldsymbol{s}}_{\nu;t}(g,\boldsymbol{x}|o_s,o_e)}_{\text{Orbital term}}\end{aligned} \tag{23}$$

where $\wedge$ denotes the cross product (wedge product). In these models, we compute the translational and rotational score using two different types of equivariant fields: 1) the equivariant density field $\rho_{\square;t}(\cdot|o_e):\mathbb{R}^3\to\mathbb{R}_{\geq 0}$, and 2) the time-conditioned score field $\widetilde{\boldsymbol{s}}_{\square;t}(\cdot|o_s,o_e):SE(3)\times\mathbb{R}^3\to\mathbb{R}^3$, where $\square$ is either ω or ν.

Proposition 4. *The score model in Eq. (21) satisfies Eq. (11) and Eq. (12) if for $\square=\omega,\nu$ the density and score fields satisfy the following conditions for all $\Delta g\in SE(3)$*

$$\rho_{\square;t}(\Delta g\,\boldsymbol{x}|\Delta g\cdot o_e) = \rho_{\square;t}(\boldsymbol{x}|o_e) \tag{24}$$

$$\widetilde{\boldsymbol{s}}_{\square;t}(\Delta g\,g,\boldsymbol{x}|\Delta g\cdot o_s,o_e) = \widetilde{\boldsymbol{s}}_{\square;t}(g,\boldsymbol{x}|o_s,o_e) \tag{25}$$

$$\widetilde{\boldsymbol{s}}_{\square;t}(g\,\Delta g^{-1},\Delta g\,\boldsymbol{x}|o_s,\Delta g\cdot o_e) = \Delta R\,\widetilde{\boldsymbol{s}}_{\square;t}(g,\boldsymbol{x}|o_s,o_e) \tag{26}$$

See Supp. C.6 for the proof. To achieve the left invariance (Eq. (25)) and right equivariance (Eq. (26)) of the score field, we propose using the following model with two EDFs:

$$\begin{aligned}&\widetilde{\boldsymbol{s}}_{\square;t}(g,\boldsymbol{x}|o_s,o_e) \\ &= \boldsymbol{\psi}_{\square;t}(\boldsymbol{x}|o_e)\ \otimes^{(\to 1)}_{\square;t}\ \mathbf{D}(R^{-1})\,\boldsymbol{\varphi}_{\square;t}(g\,\boldsymbol{x}|o_s)\end{aligned} \tag{27}$$

where $\boldsymbol{\varphi}_{\square;t}$ and $\boldsymbol{\psi}_{\square;t}$ are two different EDFs that respectively encode the point clouds o_s and o_e, and $\otimes^{(\to 1)}_{\square;t}$ is the time-conditioned equivariant tensor product [18, 50] with *Clebsch-Gordan coefficients* that maps the highly over-parametrized equivariant descriptors into a type-1 vector.

Proposition 5. *The score field model in Eq. (27) satisfies Eq. (25) and Eq. (26).*

The proof of Proposition 5 is provided in Supp. C.7. We provide the implementation details of EDFs in Supp. D.2 and the query density field model in Sec. D.3.

4. Experimental Results

We assess the data-efficiency and generalizability of diffusion-EDFs by benchmarking against state-of-the-art methods in simulation environment. The results are summarized in Tab.1. We also validate the effectiveness of diffusion-EDFs with real hardware experiments. Further details are provided in Supp. E.

5. Conclusion

In this paper, we present Diffusion-EDFs, a bi-equivariant diffusion-based generative model on the $SE(3)$ manifold for visual robotic manipulation with point cloud observations. Diffusion-EDFs significantly improve the slow training time and small receptive field of EDFs without losing their benefits. By thorough simulation and real hardware experiments, we validate Diffusion-EDFs' data efficiency and generalizability. One limitation of Diffusion-EDFs is the inability of control-level or trajectory-level inference. The application of geometric control framework [44, 45] or guided diffusion with motion planning cost [26, 51] can be considered in subsequent work. The other limitation is the necessity of the grasp observation procedure, which prevents its application to closed-loop inference. Future research may incorporate point cloud segmentation techniques to distinguish the grasp point cloud from the scene point cloud in a single observation.

Acknowledgments This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (No.RS-2023-00221762). This work was also supported by the Korea Institute of Science and Technology (KIST) intramural grants (2E31570), and a Berkeley Fellowship.

References

[1] Anurag Ajay, Yilun Du, Abhi Gupta, Joshua B. Tenenbaum, Tommi S. Jaakkola, and Pulkit Agrawal. Is conditional generative modeling all you need for decision-making? In *International Conference on Learning Representations (ICLR)*, 2023. 1

[2] Ondrej Biza, Skye Thompson, Kishore Reddy Pagidi, Abhinav Kumar, Elise van der Pol, Robin Walters, Thomas Kipf, Jan-Willem van de Meent, Lawson L. S. Wong, and Robert Platt. One-shot imitation learning via interaction warping. In *CoRL*, 2023. 14

[3] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023. 1

[4] Valentin De Bortoli, Emile Mathieu, Michael John Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. In *Advances in Neural Information Processing Systems*, 2022. 2

[5] Johann Brehmer, Joey Bose, Pim De Haan, and Taco Cohen. EDGI: Equivariant diffusion for planning with embodied agents. In *Workshop on Reincarnating Reinforcement Learning at ICLR 2023*, 2023. 1

[6] Johann Brehmer, Pim De Haan, Sönke Behrends, and Taco Cohen. Geometric algebra transformers. In *RSS 2023 Workshop on Symmetries in Robot Learning*, 2023. 1

[7] Roger Brockett. Notes on stochastic processes on manifolds. In *Systems and Control in the Twenty-first Century*, pages 75–100. Springer, 1997. 2

[8] Evangelos Chatzipantazis, Stefanos Pertigkiozoglou, Edgar Dobriban, and Kostas Daniilidis. SE(3)-equivariant attention networks for shape reconstruction in function space. In *The Eleventh International Conference on Learning Representations*, 2023. 9, 11

[9] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023. 1

[10] Gregory S Chirikjian. *Engineering applications of noncommutative harmonic analysis: with emphasis on rotation and motion groups*. CRC press, 2000. 3, 5

[11] Gregory S Chirikjian. *Stochastic models, information theory, and Lie groups, volume 2: Analytic methods and modern applications*. Springer Science & Business Media, 2011. 2, 3

[12] Gregory S Chirikjian. Partial bi-invariance of SE(3) metrics. *Journal of Computing and Information Science in Engineering*, 15(1), 2015. 3

[13] Ethan Chun, Yilun Du, Anthony Simeonov, Tomas Lozano-Perez, and Leslie Kaelbling. Local neural descriptor fields: Locally conditioned object representations for manipulation. *arXiv preprint arXiv:2302.03573*, 2023. 1, 9, 15

[14] David T Coleman, Ioan A Sucan, Sachin Chitta, and Nikolaus Correll. Reducing the barrier to entry of complex robotic software: a moveit! case study. *Journal of Software Engineering In Robotics*, 5(1):3–16, 2014. 18

[15] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. *International Conference on Learning Representations (ICLR)*, 2023. 2, 3, 10

[16] Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas J Guibas. Vector neurons: A general framework for SO(3)-equivariant networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12200–12209, 2021. 16

[17] Congyue Deng, Jiahui Lei, Bokui Shen, Kostas Daniilidis, and Leonidas Guibas. Banana: Banach fixed-point network for pointcloud segmentation with inter-part equivariance. *arXiv preprint arXiv:2305.16314*, 2023. 9

[18] Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. SE(3)-transformers: 3d roto-translation equivariant attention networks. *Advances in neural information processing systems*, 33:1970–1981, 2020. 4, 12

[19] Sergio Garrido-Jurado, Rafael Muñoz-Salinas, Francisco José Madrid-Cuevas, and Manuel Jesús Marín-Jiménez. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition*, 47(6):2280–2292, 2014. 17

[20] Mario Geiger and Tess Smidt. e3nn: Euclidean neural networks, 2022. 1

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 13

[22] Chin-Wei Huang, Milad Aghajohari, Joey Bose, Prakash Panangaden, and Aaron C Courville. Riemannian diffusion models. *Advances in Neural Information Processing Systems*, 35:2750–2761, 2022. 2

[23] Haojie Huang, Dian Wang, Robin Walters, and Robert Platt. Equivariant transporter network. In *Proceedings of Robotics: Science and Systems*, New York City, NY, USA, 2022. 1

[24] Haojie Huang, Dian Wang, Xupeng Zhu, Robin Walters, and Robert Platt. Edge grasp network: A graph-based SE(3)-invariant approach to grasp detection. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3882–3888. IEEE, 2023. 1

[25] Yesukhei Jagvaral, Francois Lanusse, and Rachel Mandelbaum. Diffusion generative models on SO(3). 2022. 2, 3

[26] Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning*, 2022. 1, 4

[27] Jiwoo Kim, Hyunwoo Ryu, Jongeun Choi, Joohwan Seo, Nikhil Potu Surya Prakash, Ruolin Li, and Roberto Horowitz. Robotic manipulation learning with equivariant descriptor fields: Generative modeling, bi-equivariance,

steerability, and locality. In *RSS 2023 Workshop on Symmetries in Robot Learning*, 2023. 1, 9, 10, 15

[28] Frederic Koehler, Alexander Heckett, and Andrej Risteski. Statistical efficiency of score matching: The view from isoperimetry. In *The Eleventh International Conference on Learning Representations*, 2023. 13

[29] Alexander B Kyatkin and Gregory S Chirikjian. Regularized solutions of a nonlinear convolution equation on the euclidean group. *Acta Applicandae Mathematica*, 53:89–123, 1998. 5

[30] Mathieu Labbé and François Michaud. Rtab-map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation. *Journal of field robotics*, 36(2):416–446, 2019. 17

[31] Adam Leach, Sebastian M Schmon, Matteo T Degiacomi, and Chris G Willcocks. Denoising diffusion probabilistic models on SO(3) for rotational alignment. In *ICLR 2022 Workshop on Geometrical and Topological Representation Learning*, 2022. 2, 3

[32] Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. In *The Eleventh International Conference on Learning Representations*, 2023. 11, 12

[33] Yen-Chen Lin, Pete Florence, Andy Zeng, Jonathan T Barron, Yilun Du, Wei-Chiu Ma, Anthony Simeonov, Alberto Rodriguez Garcia, and Phillip Isola. Mira: Mental imagery for robotic affordances. In *Conference on Robot Learning*, pages 1916–1927. PMLR, 2023. 1

[34] Weiyu Liu, Tucker Hermans, Sonia Chernova, and Chris Paxton. Structdiffusion: Object-centric diffusion for semantic rearrangement of novel objects. In *Workshop on Language and Robotics at CoRL 2022*, 2022. 1

[35] Kevin M Lynch and Frank C Park. *Modern robotics*. Cambridge University Press, 2017. 3, 7

[36] Utkarsh Aashu Mishra and Yongxin Chen. Reorientdiff: Diffusion model based reorientation for object manipulation. In *RSS 2023 Workshop on Learning for Task and Motion Planning*, 2023. 1

[37] Richard M Murray, Zexiang Li, and S Shankar Sastry. *A mathematical introduction to robotic manipulation*. CRC press, 2017. 3, 7

[38] Dmitry I Nikolayev and Tatjana I Savyolov. Normal distribution on the rotation group SO(3). *Textures and Microstructures*, 29, 1970. 2, 3

[39] Tim Pearce, Tabish Rashid, Anssi Kanervisto, Dave Bignell, Mingfei Sun, Raluca Georgescu, Sergio Valcarcel Macua, Shan Zheng Tan, Ida Momennejad, Katja Hofmann, and Sam Devlin. Imitating human behaviour with diffusion models. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[40] Hung Pham and Quang-Cuong Pham. A new approach to time-optimal path parameterization based on reachability analysis. *IEEE Transactions on Robotics*, 34(3):645–659, 2018. 18

[41] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 11

[42] Hyunwoo Ryu, Hong in Lee, Jeong-Hoon Lee, and Jongeun Choi. Equivariant descriptor fields: SE(3)-equivariant energy-based models for end-to-end visual robotic manipulation learning. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 3, 9, 10, 11, 14, 15, 16, 19

[43] TM Ivanova TI Savyolova. Normal distributions on SO(3). In *Programming And Mathematical Techniques In Physics-Proceedings Of The Conference On Programming And Mathematical Methods For Solving Physical Problems*, page 220. World Scientific, 1994. 2

[44] Joohwan Seo, Nikhil Potu Surya Prakash, Alexander Rose, and Roberto Horowitz. Geometric impedance control on SE(3) for robotic manipulators. *IFAC World Congress*, 2023. 4

[45] Joohwan Seo, Nikhil P. S. Prakash, Xiang Zhang, Changhao Wang, Jongeun Choi, Masayoshi Tomizuka, and Roberto Horowitz. Contact-rich se(3)-equivariant robot manipulation task learning via geometric impedance control. *IEEE Robotics and Automation Letters*, 9(2):1508–1515, 2024. 4

[46] Anthony Simeonov, Yilun Du, Andrea Tagliasacchi, Joshua B Tenenbaum, Alberto Rodriguez, Pulkit Agrawal, and Vincent Sitzmann. Neural descriptor fields: SE(3)-equivariant object representations for manipulation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6394–6400. IEEE, 2022. 1, 14, 16

[47] Anthony Simeonov, Yilun Du, Yen-Chen Lin, Alberto Rodriguez Garcia, Leslie Pack Kaelbling, Tomás Lozano-Pérez, and Pulkit Agrawal. SE(3)-equivariant relational rearrangement with neural descriptor fields. In *Conference on Robot Learning*, pages 835–846. PMLR, 2023. 1, 4, 14, 15, 16

[48] Anthony Simeonov, Ankit Goyal, Lucas Manuelli, Lin Yen-Chen, Alina Sarmiento, Alberto Rodriguez, Pulkit Agrawal, and Dieter Fox. Shelving, stacking, hanging: Relational pose diffusion for multi-modal rearrangement. *Conference on Robot Learning*, 2023. 1, 16

[49] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 2, 3, 13

[50] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018. 4, 12

[51] Julen Urain, Niklas Funk, Jan Peters, and Georgia Chalvatzaki. SE(3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion. *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 1, 3, 4, 2, 13, 14, 15, 16, 19

[52] Dian Wang, Robin Walters, Xupeng Zhu, and Robert Platt. Equivariant q learning in spatial action spaces. In *Conference on Robot Learning*, pages 1713–1723. PMLR, 2022. 1

[53] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. SAPIEN: A simulated part-based interactive environment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 15

[54] Jason Yim, Brian L Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. SE(3) diffusion model with application to protein backbone generation. *International Conference on Machine Learning*, 2023. 2, 3, 10

[55] Anthony Zee. *Group theory in a nutshell for physicists*. Princeton University Press, 2016. 2

[56] Andy Zeng, Pete Florence, Jonathan Tompson, Stefan Welker, Jonathan Chien, Maria Attarian, Travis Armstrong, Ivan Krasin, Dan Duong, Vikas Sindhwani, and Johnny Lee. Transporter networks: Rearranging the visual world for robotic manipulation. *Conference on Robot Learning (CoRL)*, 2020. 1

[57] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. 17

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

Supplementary Material

A. Bi-equivariance

For robust pick-and-place manipulation, the trained policy needs to be generalizable to previously unseen configurations of the target objects to pick/place. This can be achieved by inferring end-effector poses that keep the relative pose between the grasped object and the placement target invariant. Note that in our formulation, picking is essentially a special case of placing tasks, in which the gripper is *placed* at appropriate grasp points of the target object to pick with an appropriate orientation.

Consider the scenario in which the policy is trained with a demonstration (g_{we}, o_s, o_e) in which g_{we} is the end-effector pose, and o_s and o_e are respectively the point cloud observations of the scene and grasp. We denote the world frame using subscript w and the end-effector frame using subscript e. Note that o_s is observed in frame w and o_e in frame e. Now, let the placement target be moved by $\Delta g = g_{w'w}$, inducing the transformation of the observation $o_s \rightarrow \Delta g\, o_s$. This is equivalent to changing the world reference frame from w to w' with respect to the observation. Therefore, the end-effector pose should also be transformed equivariantly as $g_{we} \rightarrow g_{w'e} = \Delta g\, g_{we}$ (see Fig. 2-(a)). This *scene equivariance* is also referred to as *left equivariance* [27, 42], as the transformation Δg comes to the left side of g_{we}.

On the other hand, consider the transformation of the grasped object $\Delta g = g_{e'e}$, which induces the transformation of the observation $o_e \rightarrow \Delta g\, o_e$. This is equivalent to changing the end-effector reference frame from e to e' with respect to the observation. In the world frame, this corresponds to the transformation of the end-effector pose by $g_{we} \rightarrow g_{we'} = g_{we}\Delta g^{-1}$ (see Fig. 2-(b)). This *grasp equivariance* is also referred to as *right equivariance* [27, 42], as the transformation Δg^{-1} comes to the right side of g_{we}. Combining these left and right equivariance conditions, we obtain the bi-equivariance condition, which can be formally expressed in a probabilistic form as Eq. (10).

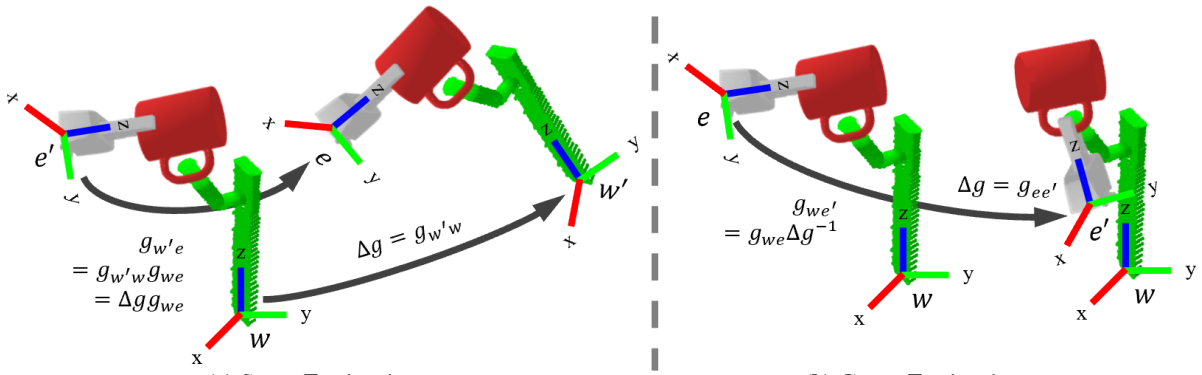

Figure 2. **Scene Equivariance and Grasp Equivariance. (a)** The end-effector pose must follow the transformation of the placement target within the scene. This *scene equivariance* can be achieved by multiplying the transformation Δg on the left side of the end-effector pose. Therefore, we also refer to this property as the *left equivariance*. **(b)** The end-effector pose must move contravariantly to the transformation of the grasped object to compensate for the changes. This *grasp equivariance* involves the inverse transformation Δg^{-1} coming to the right side of the end-effector pose. Therefore, we also refer to this property as the *right equivariance*.

B. Analytic Form of the Target Score in Eq. (19)

In this section, we provide the analytic form of the target score function in Eq. (19)

$$\nabla \log \mathcal{B}_t(g_{ed}^{-1} g_0^{-1} g g_{ed}) \tag{28}$$

By definition, the i-th component of the target score function is calculated as follows:

$$\begin{aligned}
\mathcal{L}_i \log \mathcal{B}_t(g_{ed}^{-1} g_0^{-1} g g_{ed}) &= \left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \log \mathcal{B}_t \left(g_{ed}^{-1} g_0^{-1} g \exp[\epsilon \hat{\boldsymbol{e}}_i] g_{ed}\right) \\
&= \left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \log \mathcal{B}_t \left(g_{ed}^{-1} g_0^{-1} g\, g_{ed} \exp[\epsilon Ad_{g_{ed}^{-1}} \hat{\boldsymbol{e}}_i]\right) \\
&= \left[\mathcal{L}_{Ad_{g_{ed}^{-1}} \hat{\boldsymbol{e}}_i} \log \mathcal{B}_t\right] \left(g_{ed}^{-1} g_0^{-1} g g_{ed}\right) \\
&= \left[\sum_{j=1}^{6} \left[Ad_{g_{ed}^{-1}}\right]_{ji} \mathcal{L}_j \log \mathcal{B}_t\right] \left(g_{ed}^{-1} g_0^{-1} g g_{ed}\right) \\
&= \sum_{j=1}^{6} \left[Ad_{g_{ed}^{-1}}\right]_{ji} \left[\mathcal{L}_j \log \mathcal{B}_t\right] \left(g_{ed}^{-1} g_0^{-1} g g_{ed}\right)
\end{aligned} \tag{29}$$

$$\Rightarrow \nabla \log \mathcal{B}_t(g_{ed}^{-1} g_0^{-1} g g_{ed}) = [Ad_{g_{ed}}]^{-T} \left[\nabla \log \mathcal{B}_t\right] \left(g_{ed}^{-1} g_0^{-1} g g_{ed}\right) \tag{30}$$

Therefore, all we need is the score of the Brownian diffusion kernel $\nabla \log \mathcal{B}_t(g)$ which can be decomposed into its translation and rotation parts using Eq. (5)

$$\nabla \log \mathcal{B}_t(g) = \nabla \log \mathcal{N}(\boldsymbol{p}; \boldsymbol{\mu} = \mathbf{0}, \Sigma = tI) + \nabla \log \mathcal{IG}_{SO(3)}(R; \epsilon = t/2) \tag{31}$$

where $\nabla \log \mathcal{N}(\boldsymbol{p}; \boldsymbol{\mu} = \mathbf{0}, \Sigma = tI) = -\boldsymbol{p}/t$ can be easily computed. A common practice for the calculation of the rotational part $\nabla \log \mathcal{IG}_{SO(3)}(R; \epsilon = t/2)$ is to use automatic differentiation packages [15, 25, 31, 42, 51, 54]. However, the explicit form can be easily calculated without automatic differentiation packages.

$$\mathcal{L}_i \ \log \mathcal{IG}_{SO(3)}(R; \epsilon) = \frac{\mathcal{L}_i \, \mathcal{IG}_{SO(3)}(R; \epsilon)}{\mathcal{IG}_{SO(3)}(R; \epsilon)} \tag{32}$$

$$\mathcal{L}_i \, \mathcal{IG}_{SO(3)}(R; \epsilon) = \sum_{l=0}^{\infty} (2l+1) \exp\left[-l(l+1)\epsilon\right] \left[\frac{(l+1)\sin(l\theta) - l\sin((l+1)\theta)}{\cos(\theta) - 1}\right] \left[\frac{-\mathrm{tr}\left[R[\hat{\boldsymbol{e}}_i]^\wedge\right]}{2\sin\theta}\right] \tag{33}$$

We denote the skew-symmetric matrix of the i-th $\mathfrak{so}(3)$ basis $\hat{\boldsymbol{e}}_i$ as $[\hat{\boldsymbol{e}}_i]^\wedge$, whose matrix element is $[\hat{\boldsymbol{e}}_i]^\wedge_{jk} = -\epsilon_{ijk}$ where ϵ_{ijk} is the Levi-Civita symbol.

The derivation is as follows. First, we rewrite Eq. (6) with the *character* $\mathcal{X}(R)$ of $SO(3)$ [55].

$$\mathcal{IG}_{SO(3)}(R; \epsilon) = \sum_{l=0}^{\infty} (2l+1) \exp\left[-l(l+1)\epsilon\right] \mathcal{X}_l(R) \tag{34}$$

$$\mathcal{X}_l(R) = \mathrm{tr}\left[D_l(R)\right] = \sin\left((2l+1)\frac{\theta}{2}\right) / \sin(\frac{\theta}{2}) \tag{35}$$

$\theta \in (0, \pi)$ is the rotation angle of R. Now we calculate the Lie derivative of $\mathcal{IG}_{SO(3)}$ as follows:

$$\mathcal{L}_i \, \mathcal{IG}_{SO(3)}(R; \epsilon) = \sum_{l=0}^{\infty} (2l+1) \exp\left[-l(l+1)\epsilon\right] \mathcal{L}_i \, \mathcal{X}_l(R) \tag{36}$$

$$\mathcal{L}_i \, \mathcal{X}_l(R) = \left[\frac{(l+1)\sin(l\theta) - l\sin((l+1)\theta)}{\cos(\theta) - 1}\right] \mathcal{L}_i \, \theta \tag{37}$$

$$\mathcal{L}_i \, \theta = \left[\frac{-1}{\sin\theta}\right] \mathcal{L}_i \left[\cos\theta\right] \tag{38}$$

The last line can be easily calculated using $\cos\theta = \frac{1}{2}\left(\mathrm{tr}\left[R\right] - 1\right)$ and $\mathcal{L}_{\mathcal{V}} \, \mathrm{tr}\left[R\right] = \mathrm{tr}\left[R[\mathcal{V}]^\wedge\right]$.

$$\mathcal{L}_i \left[\cos\theta\right] = \frac{1}{2}\left(\mathrm{tr}\left[R[\hat{\boldsymbol{e}}_i]^\wedge\right]\right) \tag{39}$$

Combining these results, one can derive Eq. (33). In practice, the infinite sum in Eq. (33) is approximated with $\sum_{l=0}^{l_{max}}$ where $l_{max} = 1000 \sim 10000$, which can be computed within a millisecond when appropriately parallelized. Although we have derived Eq. (33) for $\theta = (0, \pi)$, the result can be asymptotically extended to $\theta = 0$ and π as $\mathcal{IG}_{SO(3)}$ is an infinitely differentiable on $SO(3)$ [38].

C. Proofs and Derivations

C.1. Proof of Proposition 1

Proof of the left invariance of the score function.

$$\begin{aligned}
\mathcal{L}_i \log P(\Delta g\, g|\Delta g \cdot o_s, o_e) &= \left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \log P(\Delta g\, g \exp\left[\epsilon \hat{\boldsymbol{e}}_i\right]|\Delta g \cdot o_s, o_e) \\
&= \left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \log P(g \exp\left[\epsilon \hat{\boldsymbol{e}}_i\right]|o_s, o_e) \\
&= \mathcal{L}_i \log P(g|o_s, o_e)
\end{aligned}$$

where we used $P(\Delta g\, g|\Delta g \cdot o_s, o_e) = P(g|o_s, o_e)$ in the second line. □

Proof of the right equivariance of the score function.

$$\begin{aligned}
\mathcal{L}_i \log P(g\, \Delta g^{-1}|o_s, \Delta g \cdot o_e) &= \left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \log P(g\, \Delta g^{-1} \exp\left[\epsilon \hat{\boldsymbol{e}}_i\right]|o_s, \Delta g \cdot o_e) \\
&= \left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \log P(g\, \Delta g^{-1} \exp\left[\epsilon \hat{\boldsymbol{e}}_i\right] \Delta g|o_s, o_e) \\
&= \left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \log P(g \exp\left[\epsilon Ad_{\Delta g^{-1}} \hat{\boldsymbol{e}}_i\right]|o_s, o_e) \\
&= \mathcal{L}_{Ad_{\Delta g^{-1}} \hat{\boldsymbol{e}}_i} \log P(g|o_s, o_e) \\
&= \mathcal{L}_{\sum_j \left[Ad_{\Delta g^{-1}}\right]_{ji} \hat{\boldsymbol{e}}_j} \log P(g|o_s, o_e) \\
&= \sum_{j=1}^{6} \left[Ad_{\Delta g^{-1}}\right]_{ji} \mathcal{L}_j \log P(g|o_s, o_e) \\
&\qquad (\because \text{Linearity of Lie-derivatives [11]}\ \ \mathcal{L}_{\sum_i v_i \hat{\boldsymbol{e}}_i} = \textstyle\sum_i v_i \mathcal{L}_i) \\
\Rightarrow \nabla \log P(g\, \Delta g^{-1}|o_s, \Delta g \cdot o_e) &= \left[Ad_{\Delta g^{-1}}\right]^T \nabla \log P(g|o_s, o_e) = \left[Ad_{\Delta g}\right]^{-T} \nabla \log P(g|o_s, o_e)
\end{aligned}$$

where we denote the (j, i)-th matrix element of $Ad_{\Delta g^{-1}}$ with $\left[Ad_{\Delta g^{-1}}\right]_{ji}$. We used $P(g\, \Delta g^{-1}|o_s, \Delta g \cdot o_e) = P(g|o_s, o_e)$ in the second line. □

C.2. Proof of Proposition 2

It is straightforward to prove the bi-equivariance of the diffused marginal using the bi-invariance of the integral measure (Haar measure) dg

$$\int_{SE(3)} d(\Delta g\, g) = \int_{SE(3)} dg = \int_{SE(3)} d(g\, \Delta g) \qquad \forall \Delta g \in SE(3) \tag{40}$$

where $dg = dR\, d\boldsymbol{p} = \frac{1}{8\pi^2} (\sin\beta)\, d\alpha\, d\beta\, d\gamma\, dx\, dy\, dz$ in the rotation-translation coordinate with the Euler angles α, β, γ and the frame origin x, y, z. See Chirikjian [10, 11, 12], Murray et al. [37] and Appendix A of Ryu et al. [42] for more details on the bi-invariant integral measure of $SE(3)$.

We first prove that the marginal is bi-equivariant if the kernel is bi-equivariant.

Proof of left equivariance.

$$
\begin{aligned}
P_t(g|o_s, o_e) &= \int_{SE(3)} dg_0 \; P_{t|0}(g|g_0, o_s, o_e) P_0(g_0|o_s, o_e) \\
&= \int_{SE(3)} dg_0 \; P_{t|0}(g|g_0, o_s, o_e) P_0(\Delta g \, g_0|\Delta g \cdot o_s, o_e) && (\because \text{Eq. (10)}) \\
&= \int_{SE(3)} dg_0 \; P_{t|0}(\Delta g \, g|\Delta g \, g_0, \Delta g \cdot o_s, o_e) P_0(\Delta g \, g_0|\Delta g \cdot o_s, o_e) && (\because \text{Eq. (15)}) \\
&= \int_{SE(3)} dg_0 \; P_{t|0}(\Delta g \, g|g_0, \Delta g \cdot o_s, o_e) P_0(g_0|\Delta g \cdot o_s, o_e) && (\because \text{Eq. (40)}, \; \Delta g \, g_0 \to g_0) \\
&= P_t(\Delta g \, g|\Delta g \cdot o_s, o_e)
\end{aligned}
$$

□

Proof of right equivariance.

$$
\begin{aligned}
P_t(g|o_s, o_e) &= \int_{SE(3)} dg_0 \; P_{t|0}(g|g_0, o_s, o_e) P_0(g_0|o_s, o_e) \\
&= \int_{SE(3)} dg_0 \; P_{t|0}(g|g_0, o_s, o_e) P_0(g_0 \, \Delta g^{-1}|o_s, \Delta g \cdot o_e) && (\because \text{Eq. (10)}) \\
&= \int_{SE(3)} dg_0 \; P_{t|0}(g \, \Delta g^{-1}|g_0 \, \Delta g^{-1}, o_s, \Delta g \cdot o_e) P_0(g_0 \, \Delta g^{-1}|o_s, \Delta g \cdot o_e) && (\because \text{Eq. (15)}) \\
&= \int_{SE(3)} dg_0 \; P_{t|0}(g \, \Delta g^{-1}|g_0, o_s, \Delta g \cdot o_e) P_0(g_0|o_s, \Delta g \cdot o_e) && (\because \text{Eq. (40)}, \; g_0 \, \Delta g^{-1} \to g_0) \\
&= P_t(g \, \Delta g^{-1}|o_s, \Delta g \cdot o_e)
\end{aligned}
$$

□

Similarly, it can be proven that the kernel must be bi-equivariant (up to measure zero) to guarantee the bi-equivariance of the diffused marginal for any arbitrary initial distribution $dP_0 = P_0 \, dg_0$.

Proof.

$$
\begin{aligned}
P_t(g|o_s, o_e) &= \int_{SE(3)} dg_0 \; P_{t|0}(g|g_0, o_s, o_e) P_0(g_0|o_s, o_e) \\
P_t(\Delta g \, g|\Delta g \cdot o_s, o_e) &= \int_{SE(3)} dg_0 \; P_{t|0}(\Delta g \, g|g_0, \Delta g \cdot o_s, o_e) P_0(g_0|\Delta g \cdot o_s, o_e) \\
&= \int_{SE(3)} dg_0 \; P_{t|0}(\Delta g \, g|g_0, \Delta g \cdot o_s, o_e) P_0(\Delta g^{-1} \, g_0|o_s, o_e) && (\because \text{Eq. (10)}) \\
&= \int_{SE(3)} dg_0 \; P_{t|0}(\Delta g \, g|\Delta g \, g_0, \Delta g \cdot o_s, o_e) P_0(g_0|o_s, o_e) && (\because \text{Eq. (40)}, \; g_0 \to \Delta g \, g_0)
\end{aligned}
$$

$$
\begin{aligned}
P_t(g \, \Delta g^{-1}|o_s, \Delta g \cdot o_e) &= \int_{SE(3)} dg_0 \; P_{t|0}(g \, \Delta g^{-1}|g_0, o_s, \Delta g \cdot o_e) P_0(g_0|o_s, \Delta g \cdot o_e) \\
&= \int_{SE(3)} dg_0 \; P_{t|0}(g \, \Delta g^{-1}|g_0, o_s, \Delta g \cdot o_e) P_0(g_0 \, \Delta g|o_s, o_e) && (\because \text{Eq. (10)}) \\
&= \int_{SE(3)} dg_0 \; P_{t|0}(g \, \Delta g^{-1}|g_0 \, \Delta g^{-1}, o_s, \Delta g \cdot o_e) P_0(g_0|o_s, o_e) && (\because \text{Eq. (40)}, \; g_0 \to g_0 \, \Delta g^{-1})
\end{aligned}
$$

$$\Rightarrow \int_{SE(3)} dg_0\ P_0(g_0|o_s,o_e) \times \left[P_{t|0}(g|g_0,o_s,o_e) - P_{t|0}(\Delta g\, g|\Delta g\, g_0, \Delta g\cdot o_s, o_e)\right] = 0$$
$$\int_{SE(3)} dg_0\ P_0(g_0|o_s,o_e) \times \left[P_{t|0}(g|g_0,o_s,o_e) - P_{t|0}(g\,\Delta g^{-1}|g_0\,\Delta g^{-1}, o_s, \Delta g\cdot o_e)\right] = 0$$

Therefore, for this equation to hold for any arbitrary bi-equivariant initial distribution $dP_0 = P_0\, dg_0$, the diffusion kernel must be bi-equivariant $\forall g, \Delta g \in SE(3)$

$$P_{t|0}(g|g_0,o_s,o_e) - P_{t|0}(\Delta g\, g|\Delta g\, g_0, \Delta g\cdot o_s, o_e) = 0 \tag{41}$$
$$P_{t|0}(g|g_0,o_s,o_e) - P_{t|0}(g\,\Delta g^{-1}|g_0\,\Delta g^{-1}, o_s, \Delta g\cdot o_e) = 0 \tag{42}$$
$$\Rightarrow\ P_{t|0}(g|g_0,o_s,o_e) = P_{t|0}(\Delta g\, g|\Delta g\, g_0, \Delta g\cdot o_s, o_e) = P_{t|0}(g\,\Delta g^{-1}|g_0\,\Delta g^{-1}, o_s, \Delta g\cdot o_e) \tag{43}$$

□

C.3. Non-existence of Bi-Invariant Diffusion Kernels on SE(3)

Note that any left invariant kernel $P_{t|0}(g|g_0)$ can be written in a univariate form $K_t(g_0^{-1}g)$.

$$P_{t|0}(\Delta g\, g|\Delta g\, g_0) = P_{t|0}(g|g_0) \quad \forall\, \Delta g, g \quad \Rightarrow \quad P_{t|0}(g|g_0) = P_{t|0}(g_0^{-1}g|I) \quad \forall\, g \tag{44}$$

The right invariance requires this kernel to satisfy $K_t(\Delta g\, g\, \Delta g^{-1}) = K_t(g)$, meaning that it is a *class function*, which does not exist for $L^2(SE(3))$ [10, 29].

C.4. Proof of Proposition 3

Note that the Brownian diffusion kernel $\mathcal{B}_t(g)$ is right-invariant to rotation, that is,

$$\begin{aligned} &\mathcal{B}_t\left((g_0\,\Delta R)^{-1}\,(g\,\Delta R)\right) = \mathcal{B}_t(g_0^{-1}\,g) \\ &\quad \Rightarrow \mathcal{B}_t(\Delta R^{-1}\,g\,\Delta R) = \mathcal{B}_t(g) \qquad \forall \Delta R \in SO(3) \end{aligned} \tag{45}$$

where we abuse the notation to denote the action of a pure rotation ΔR on $g = (\boldsymbol{p}, R)$ as $\Delta R\, g = (\Delta R\, \boldsymbol{p}, \Delta R\, R)$ and $g\,\Delta R = (\boldsymbol{p}, R\,\Delta R)$. Eq. (45) holds because the Gaussian distribution in Eq. (5) is rotation-invariant and $\mathcal{IG}_{SO(3)}$ in Eq. (6) is a linear combination of *characters* of $SO(3)$, which are *class functions* due to the permutation invariance of trace operations (see Supp. B and C.3). Consider the following diffusion kernel with the equivariant origin selection mechanism in Eq. (18):

$$P_{t|0}(g|g_0,o_s,o_e) = \int_{\mathbb{R}^3} d\boldsymbol{p}_{ed} P\left(\boldsymbol{p}_{ed}|g_0^{-1}\cdot o_s, o_e\right) \mathcal{B}_t\left((g_0 \triangleleft \boldsymbol{p}_{ed})^{-1}\,(g \triangleleft \boldsymbol{p}_{ed})\right) \tag{46}$$

where $\triangleleft \boldsymbol{p}_{ed} : SE(3) \to SE(3)$ denotes the right action of a pure translation $\boldsymbol{p}_{ed} \in \mathbb{R}^3$ onto $g = (\boldsymbol{p}, R) \in SE(3)$ such that

$$g \triangleleft \boldsymbol{p}_{ed} = \begin{bmatrix} R & \boldsymbol{p} \\ \emptyset & 1 \end{bmatrix} \begin{bmatrix} I & \boldsymbol{p}_{ed} \\ \emptyset & 1 \end{bmatrix} = \begin{bmatrix} R & R\,\boldsymbol{p}_{ed} + \boldsymbol{p} \\ \emptyset & 1 \end{bmatrix} \tag{47}$$

Note that the following equation holds for all $g_1, g_2 \in SE(3)$ and $\boldsymbol{p}_{ed} \in \mathbb{R}^3$:

$$\begin{aligned} (g_1\, g_2) \triangleleft \boldsymbol{p}_{ed} &= \begin{bmatrix} R_1 & \boldsymbol{p}_1 \\ \emptyset & 1 \end{bmatrix} \begin{bmatrix} R_2 & \boldsymbol{p}_2 \\ \emptyset & 1 \end{bmatrix} \begin{bmatrix} I & \boldsymbol{p}_{ed} \\ \emptyset & 1 \end{bmatrix} \\ &= \begin{bmatrix} R_1\, R_2 & R_1\,(R_2 \boldsymbol{p}_{ed} + \boldsymbol{p}_2) + \boldsymbol{p}_1 \\ \emptyset & 1 \end{bmatrix} \\ &= \begin{bmatrix} R_1 & \boldsymbol{p}_1 \\ \emptyset & 1 \end{bmatrix} \begin{bmatrix} I & g_2\,\boldsymbol{p}_{ed} \\ \emptyset & 1 \end{bmatrix} \begin{bmatrix} R_2 & \mathbf{0} \\ \emptyset & 1 \end{bmatrix} \\ &= (g_1 \triangleleft (g_2\,\boldsymbol{p}_{ed}))\ \Delta R_2 \end{aligned} \tag{48}$$

The bi-equivariance of $P_{t|0}$ can be proved using Eq. (45) and Eq. (48).

Proof. The proof of left equivariance is straightforward as $g_0^{-1} \cdot \mathcal{O}_s = (\Delta g\, g_0)^{-1} \cdot (\Delta g \cdot \mathcal{O}_s)$. The proof of right equivariance is as follows:

$$
\begin{aligned}
& P_{t|0}(g\,\Delta g^{-1} | g_0\,\Delta g^{-1}, \mathcal{O}_s, \Delta g \cdot \mathcal{O}_e) \\
&= \int_{\mathbb{R}^3} d\boldsymbol{p}_{ed} P\left(\boldsymbol{p}_{ed} | (\Delta g\, g_0^{-1}) \cdot \mathcal{O}_s, \Delta g \cdot \mathcal{O}_e\right) \mathcal{B}_t \left(\left((g_0\,\Delta g^{-1}) \triangleleft \boldsymbol{p}_{ed}\right)^{-1} \left((g\,\Delta g^{-1}) \triangleleft \boldsymbol{p}_{ed}\right) \right) \\
&= \int_{\mathbb{R}^3} d\boldsymbol{p}_{ed} P\left(\Delta g^{-1}\,\boldsymbol{p}_{ed} | g_0^{-1} \cdot \mathcal{O}_s, \mathcal{O}_e\right) \mathcal{B}_t \left(\left((g_0\,\Delta g^{-1}) \triangleleft \boldsymbol{p}_{ed}\right)^{-1} \left((g\,\Delta g^{-1}) \triangleleft \boldsymbol{p}_{ed}\right) \right) && (\because \text{Eq. (18)}) \\
&= \int_{\mathbb{R}^3} d\boldsymbol{p}_{ed} P\left(\Delta g^{-1}\,\boldsymbol{p}_{ed} | g_0^{-1} \cdot \mathcal{O}_s, \mathcal{O}_e\right) \mathcal{B}_t \left(\Delta R\, \left(g_0 \triangleleft \left(\Delta g^{-1}\,\boldsymbol{p}_{ed}\right)\right)^{-1} \left(g \triangleleft \left(\Delta g^{-1}\,\boldsymbol{p}_{ed}\right)\right)\, \Delta R^{-1} \right) && (\because \text{Eq. (48)}) \\
&= \int_{\mathbb{R}^3} d\boldsymbol{p}_{ed} P\left(\Delta g^{-1}\,\boldsymbol{p}_{ed} | g_0^{-1} \cdot \mathcal{O}_s, \mathcal{O}_e\right) \mathcal{B}_t \left(\left(g_0 \triangleleft \left(\Delta g^{-1}\,\boldsymbol{p}_{ed}\right)\right)^{-1} \left(g \triangleleft \left(\Delta g^{-1}\,\boldsymbol{p}_{ed}\right)\right) \right) && (\because \text{Eq. (45)}) \\
&= \int_{\mathbb{R}^3} d\boldsymbol{p}_{ed} P\left(\boldsymbol{p}_{ed} | g_0^{-1} \cdot \mathcal{O}_s, \mathcal{O}_e\right) \mathcal{B}_t \left(\left(g_0 \triangleleft (\boldsymbol{p}_{ed})\right)^{-1} \left(g \triangleleft (\boldsymbol{p}_{ed})\right) \right) \\
& && (\because \text{invariance of Euclidean integral under roto-translation, } \boldsymbol{p}_{ed} \to \Delta g\,\boldsymbol{p}_{ed}) \\
&= P_{t|0}(g | g_0, \mathcal{O}_s, \mathcal{O}_e)
\end{aligned}
$$

□

In fact, any left-invariant kernel that is also right-invariant to rotation as in Eq. (45) can be used.

C.5. Derivation of Eq. (20)

We first show that $\boldsymbol{s}_t^*(g|\mathcal{O}_s, \mathcal{O}_e) = \mathbb{E}_{g_0, g_{ed}|g, \mathcal{O}_s, \mathcal{O}_e}\left[\nabla \log K_t(g_{ed}^{-1} g_0^{-1} g g_{ed})\right]$ using a simple variational calculus.

Proof. Let $\delta \boldsymbol{s}_t(g|\mathcal{O}_s, \mathcal{O}_e)$ be a perturbation of the score model $\boldsymbol{s}_t(g|\mathcal{O}_s, \mathcal{O}_e)$. For the optimal score model $\boldsymbol{s}_t^*(g|\mathcal{O}_s, \mathcal{O}_e)$, any small perturbation would result in zero perturbation of the objective.

$$
\begin{aligned}
\boldsymbol{s}_t^*(g|\mathcal{O}_s, \mathcal{O}_e) &= \underset{\boldsymbol{s}_t(g|\mathcal{O}_s, \mathcal{O}_e)}{\arg\min}\ \mathcal{J}_t\left[\boldsymbol{s}_t(g|\mathcal{O}_s, \mathcal{O}_e)\right] \\
\Rightarrow\ \delta \mathcal{J}_t\left[\boldsymbol{s}_t^*(g|\mathcal{O}_s, \mathcal{O}_e)\right] &= 0 \qquad \forall\ \delta \boldsymbol{s}_t^*(g|\mathcal{O}_s, \mathcal{O}_e)
\end{aligned}
\tag{49}
$$

The explicit form of the perturbation of the objective with regard to $\delta \boldsymbol{s}_t(g|\mathcal{O}_s, \mathcal{O}_e)$ is written as follows:

$$
\begin{aligned}
\delta \mathcal{J}_t\left[\boldsymbol{s}_t(g|\mathcal{O}_s, \mathcal{O}_e)\right] &= \delta \left(\mathbb{E}_{g, g_0, g_{ed}, \mathcal{O}_s, \mathcal{O}_e}\left[\frac{1}{2} \left\| \boldsymbol{s}_t(g|\mathcal{O}_s, \mathcal{O}_e) - \nabla \log K_t(g_{ed}^{-1} g_0^{-1} g g_{ed}) \right\|^2 \right] \right) \\
&= \mathbb{E}_{g, \mathcal{O}_s, \mathcal{O}_e}\left[\delta \boldsymbol{s}_t(g|\mathcal{O}_s, \mathcal{O}_e) \cdot \left[\boldsymbol{s}_t(g|\mathcal{O}_s, \mathcal{O}_e) - \mathbb{E}_{g_0, g_{ed}|g, \mathcal{O}_s, \mathcal{O}_e}\left[\nabla \log K_t(g_{ed}^{-1} g_0^{-1} g g_{ed})\right] \right] \right]
\end{aligned}
\tag{50}
$$

Therefore, assuming $P_t(g|\mathcal{O}_s, \mathcal{O}_e) > 0\ \ \forall g, \mathcal{O}_s, \mathcal{O}_e$, the optimal score model must be

$$
\boldsymbol{s}_t^*(g|\mathcal{O}_s, \mathcal{O}_e) = \mathbb{E}_{g_0, g_{ed}|g, \mathcal{O}_s, \mathcal{O}_e}\left[\nabla \log K_t(g_{ed}^{-1} g_0^{-1} g g_{ed})\right] \tag{51}
$$

□

We now show that $\mathbb{E}_{g_0, g_{ed}|g, \mathcal{O}_s, \mathcal{O}_e}\left[\nabla \log K_t(g_{ed}^{-1} g_0^{-1} g g_{ed})\right] = \nabla \log P_t(g|\mathcal{O}_s, \mathcal{O}_e)$.

Proof.

$$
\begin{aligned}
&\mathbb{E}_{g_0,g_{ed}|g,O_s,O_e}\left[\nabla \log K_t(g_{ed}^{-1}g_0^{-1}gg_{ed})\right] \\
&= \int dg_0 \int dg_{ed}\; P(g_0,g_{ed}|g,o_s,o_e;t)\frac{\nabla K_t(g_{ed}^{-1}g_0^{-1}gg_{ed})}{K_t(g_{ed}^{-1}g_0^{-1}gg_{ed})} \\
&= \int dg_0 \int dg_{ed}\; \left[P(g|g_0,g_{ed},o_s,o_e;t)\frac{P(g_0,g_{ed}|o_s,o_e)}{P_t(g|o_s,o_e)}\right]\frac{\nabla K_t(g_{ed}^{-1}g_0^{-1}gg_{ed})}{K_t(g_{ed}^{-1}g_0^{-1}gg_{ed})} \\
&= \int dg_0 \int dg_{ed}\; \cancel{P(g|g_0,g_{ed},o_s,o_e;t)}\frac{P(g_0,g_{ed}|o_s,o_e)}{P_t(g|o_s,o_e)}\frac{\nabla K_t(g_{ed}^{-1}g_0^{-1}gg_{ed})}{\cancel{K_t(g_{ed}^{-1}g_0^{-1}gg_{ed})}} \\
&\qquad\qquad (\because\; P(g|g_0,g_{ed},o_s,o_e;t) = P(g|g_0,g_{ed};t) = K_t(g_{ed}^{-1}g_0^{-1}gg_{ed})\;) \\
&= \frac{1}{P_t(g|o_s,o_e)}\int dg_0 \int dg_{ed}\; P(g_0,g_{ed}|o_s,o_e)\nabla K_t(g_{ed}^{-1}g_0^{-1}gg_{ed}) \\
&= \frac{1}{P_t(g|o_s,o_e)}\nabla \int dg_0 \int dg_{ed}\; P(g_0,g_{ed}|o_s,o_e)K_t(g_{ed}^{-1}g_0^{-1}gg_{ed}) \\
&= \frac{1}{P_t(g|o_s,o_e)}\nabla \int dg_0\, P_0(g_0|o_s,o_e) \int dg_{ed}\; P(g_{ed}|g_0^{-1}\cdot o_s,o_e)K_t(g_{ed}^{-1}g_0^{-1}gg_{ed}) \\
&= \frac{1}{P_t(g|o_s,o_e)}\nabla P_t(g|o_s,o_e) \qquad\qquad (\because \text{Eq. (17) and Eq. (14)}) \\
&= \frac{\nabla P_t(g|o_s,o_e)}{P_t(g|o_s,o_e)} = \nabla \log P_t(g|o_s,o_e)
\end{aligned}
$$

□

Therefore, we prove that $\boldsymbol{s}_t^*(g|o_s,o_e) = \nabla \log P_t(g|o_s,o_e)$.

C.6. Proof of Proposition 4

For readers' convenience, we reproduce the bi-equivariance conditions for the score functions in Proposition 1 with explicit components.

$$
\boldsymbol{s}(\Delta g\, g|\Delta g\cdot o_s,o_e) = \boldsymbol{s}(g|o_s,o_e) \tag{52}
$$

$$
\begin{aligned}
\boldsymbol{s}(g\,\Delta g^{-1}|o_s,\Delta g\cdot o_e) &= [\mathrm{Ad}_{\Delta g}]^{-T}\,\boldsymbol{s}(g|o_s,o_e) \\
&= \begin{bmatrix} \Delta R & \emptyset \\ [\Delta \boldsymbol{p}]^{\wedge}\Delta R & \Delta R \end{bmatrix}\begin{bmatrix} \boldsymbol{s}_\nu(g|o_s,o_e) \\ \boldsymbol{s}_\omega(g|o_s,o_e)\end{bmatrix} \\
&= \Delta R\,\boldsymbol{s}_\nu(g|o_s,o_e) \oplus [\Delta R\,\boldsymbol{s}_\omega(g|o_s,o_e) + \Delta \boldsymbol{p}\wedge \Delta R\,\boldsymbol{s}_\nu(g|o_s,o_e)]
\end{aligned} \tag{53}
$$

where we used the fact that the inverse transpose of the adjoint matrix is as follows [35, 37]:

$$
[\mathrm{Ad}_{\Delta g}]^{-T} = \begin{bmatrix} \Delta R & \emptyset \\ [\Delta \boldsymbol{p}]^{\wedge}\Delta R & \Delta R \end{bmatrix} \tag{54}
$$

We begin by proving the bi-equivariance of the linear (translational) score term

Proof. The left invariance of the linear score model is proved as

$$
\begin{aligned}
\boldsymbol{s}_{\nu;t}(\Delta g\, g|\Delta g\cdot o_s,o_e) &= \int_{\mathbb{R}^3} d^3\boldsymbol{x}\; \rho_{\nu;t}(\boldsymbol{x}|o_e)\;\widetilde{\boldsymbol{s}}_{\nu;t}(\Delta g\, g,\boldsymbol{x}|\Delta g\cdot o_s,o_e) \\
&= \int_{\mathbb{R}^3} d^3\boldsymbol{x}\; \rho_{\nu;t}(\boldsymbol{x}|o_e)\;\widetilde{\boldsymbol{s}}_{\nu;t}(g,\boldsymbol{x}|o_s,o_e) \qquad (\because \text{Eq. (25)}) \\
&= \boldsymbol{s}_{\nu;t}(g|o_s,o_e)
\end{aligned}
$$

The right equivariance of the linear score model is proved as

$$\begin{aligned}
\boldsymbol{s}_{\nu;t}(g\,\Delta g^{-1}|o_s,\Delta g\cdot o_e) &= \int_{\mathbb{R}^3} d^3\boldsymbol{x}\;\rho_{\nu;t}(\boldsymbol{x}|\Delta g\cdot o_e)\;\widetilde{\boldsymbol{s}}_{\nu;t}(g\,\Delta g^{-1},\boldsymbol{x}|o_s,\Delta g\cdot o_e)\\
&= \int_{\mathbb{R}^3} d^3\boldsymbol{x}\;\rho_{\nu;t}(\Delta g\,\boldsymbol{x}|\Delta g\cdot o_e)\;\widetilde{\boldsymbol{s}}_{\nu;t}(g\,\Delta g^{-1},\Delta g\,\boldsymbol{x}|o_s,\Delta g\cdot o_e)\\
&\qquad\qquad (\because \text{invariance of Euclidean integral under roto-translation } \boldsymbol{x}\to\Delta g\,\boldsymbol{x})\\
&= \int_{\mathbb{R}^3} d^3\boldsymbol{x}\;\rho_{\nu;t}(\boldsymbol{x}|o_e)\;\Delta R\,\widetilde{\boldsymbol{s}}_{\nu;t}(g,\boldsymbol{x}|o_s,o_e) \qquad (\because \text{Eq. (24) and Eq. (26)})\\
&= \Delta R\int_{\mathbb{R}^3} d^3\boldsymbol{x}\;\rho_{\nu;t}(\boldsymbol{x}|o_e)\;\widetilde{\boldsymbol{s}}_{\nu;t}(g,\boldsymbol{x}|o_s,o_e)\\
&= \Delta R\,\boldsymbol{s}_{\nu;t}(g|o_s,o_e)
\end{aligned}$$

□

Let the angular (rotational) score model be decomposed into the spin term $\boldsymbol{s}_{\text{spin};t}$ and the orbital term $\boldsymbol{s}_{\text{orbital};t}$ as in Eq. (23). The bi-equivariance of spin term in the angular (rotational) score model

$$\boldsymbol{s}_{\text{spin};t}(g|o_s,o_e) = \int_{\mathbb{R}^3} d^3\boldsymbol{x}\;\rho_{\omega;t}(\boldsymbol{x}|o_e)\;\widetilde{\boldsymbol{s}}_{\omega;t}(g,\boldsymbol{x}|o_s,o_e) \tag{55}$$

$$\boldsymbol{s}_{\text{spin};t}(\Delta g\,g|\Delta g\cdot o_s,o_e) = \boldsymbol{s}_{\text{spin};t}(g|o_s,o_e) \tag{56}$$

$$\boldsymbol{s}_{\text{spin};t}(g\,\Delta g^{-1}|o_s,\Delta g\cdot o_e) = \Delta R\,\boldsymbol{s}_{\text{spin};t}(g|o_s,o_e) \tag{57}$$

can be proven in a similar fashion to the linear score model. It can be shown that the orbital term satisfies the following bi-equivariance condition

$$\boldsymbol{s}_{\text{orbital};t}(g|o_s,o_e) = \int_{\mathbb{R}^3} d^3\boldsymbol{x}\;\rho_{\nu;t}(\boldsymbol{x}|o_e)\;\boldsymbol{x}\wedge\widetilde{\boldsymbol{s}}_{\nu;t}(g,\boldsymbol{x}|o_s,o_e) \tag{58}$$

$$\boldsymbol{s}_{\text{orbital};t}(\Delta g\,g|\Delta g\cdot o_s,o_e) = \boldsymbol{s}_{\text{orbital};t}(g|o_s,o_e) \tag{59}$$

$$\boldsymbol{s}_{\text{orbital};t}(g\,\Delta g^{-1}|o_s,\Delta g\cdot o_e) = \Delta\boldsymbol{p}\wedge\Delta R\,\boldsymbol{s}_{\text{orbital};t}(g|o_s,o_e) \tag{60}$$

Proof. The left invariance is straightforward, as the linear score field $\widetilde{\boldsymbol{s}}_{\nu;t}$ is left-invariant as Eq. (25). The right equivariance can be proved as follows

$$\begin{aligned}
&\boldsymbol{s}_{\text{orbital};t}(g\,\Delta g^{-1}|o_s,\Delta g\cdot o_e)\\
&= \int_{\mathbb{R}^3} d^3\boldsymbol{x}\;\rho_{\nu;t}(\boldsymbol{x}|\Delta g\cdot o_e)\;\boldsymbol{x}\wedge\widetilde{\boldsymbol{s}}_{\nu;t}(g\,\Delta g^{-1},\boldsymbol{x}|o_s,\Delta g\cdot o_e)\\
&= \int_{\mathbb{R}^3} d^3\boldsymbol{x}\;\rho_{\nu;t}(\Delta g^{-1}\,\boldsymbol{x}|o_e)\;\boldsymbol{x}\wedge\Delta R\,\widetilde{\boldsymbol{s}}_{\nu;t}(g,\Delta g^{-1}\,\boldsymbol{x}|o_s,o_e) \qquad (\because \text{Eq. (24) and Eq. (26)})\\
&= \int_{\mathbb{R}^3} d^3\boldsymbol{x}\;\rho_{\nu;t}(\boldsymbol{x}|o_e)\;(\Delta R\,\boldsymbol{x}+\Delta\boldsymbol{p})\wedge\Delta R\,\widetilde{\boldsymbol{s}}_{\nu;t}(g,\boldsymbol{x}|o_s,o_e)\\
&\qquad\qquad (\because \text{invariance of Euclidean integral under roto-translation } \boldsymbol{x}\to\Delta g\,\boldsymbol{x}=\Delta R\,\boldsymbol{x}+\Delta\boldsymbol{p})\\
&= \Delta R\left[\int_{\mathbb{R}^3} d^3\boldsymbol{x}\;\rho_{\nu;t}(\boldsymbol{x}|o_e)\;\boldsymbol{x}\wedge\widetilde{\boldsymbol{s}}_{\nu;t}(g,\boldsymbol{x}|o_s,o_e)\right]+\Delta\boldsymbol{p}\wedge\Delta R\int_{\mathbb{R}^3} d^3\boldsymbol{x}\;\rho_{\nu;t}(\boldsymbol{x}|o_e)\;\widetilde{\boldsymbol{s}}_{\nu;t}(g,\boldsymbol{x}|o_s,o_e)\\
&\qquad\qquad (\because\; R\,\boldsymbol{x}\wedge R\,\boldsymbol{y}=R\,(\boldsymbol{x}\wedge\boldsymbol{y})\;\;\forall\,\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^3)\\
&= \Delta R\,\boldsymbol{s}_{\text{orbital};t}(g|o_s,o_e)+\Delta\boldsymbol{p}\wedge\Delta R\,\boldsymbol{s}_{\nu;t}(g|o_s,o_e)
\end{aligned}$$

□

As a result, the angular (rotational) score model

$$\boldsymbol{s}_{\omega;t}(g|o_s,o_e) = \boldsymbol{s}_{\text{orbital};t}(g|o_s,o_e)+\boldsymbol{s}_{\text{spin};t}(g|o_s,o_e) \tag{61}$$

satisfies the following bi-equivariance

$$\boldsymbol{s}_{\omega;t}(\Delta g\, g|\Delta g \cdot o_s, o_e) = \boldsymbol{s}_{\omega;t}(g|o_s, o_e) \tag{62}$$

$$\begin{aligned}\boldsymbol{s}_{\omega;t}(g\,\Delta g^{-1}|o_s, \Delta g \cdot o_e) =& \Delta R\,[\boldsymbol{s}_{\text{orbital};t}(g|o_s, o_e) + \boldsymbol{s}_{\text{spin};t}(g|o_s, o_e)] + \Delta \boldsymbol{p} \wedge \Delta R\, \boldsymbol{s}_{\nu;t}(g|o_s, o_e) \\ =& \Delta R\, \boldsymbol{s}_{\omega;t}(g|o_s, o_e) + \Delta \boldsymbol{p} \wedge \Delta R\, \boldsymbol{s}_{\nu;t}(g|o_s, o_e)\end{aligned} \tag{63}$$

Hence, we have proven Proposition 4 that the score model in Eq. (21) is bi-equivariant, satisfying Eq. (52) and Eq. (53).

C.7. Proof of Proposition 5

Proof.

$$\begin{aligned}\widetilde{\boldsymbol{s}}_{\square;t}(\Delta g\, g, \boldsymbol{x}|\Delta g \cdot o_s, o_e) &= \boldsymbol{\psi}_{\square;t}(\boldsymbol{x}|o_e) \otimes^{(\to 1)}_{\square;t} \mathbf{D}(R^{-1}\Delta R^{-1})\, \boldsymbol{\varphi}_{\square;t}(\Delta g\, g\, \boldsymbol{x}|\Delta g \cdot o_s) \\ &= \boldsymbol{\psi}_{\square;t}(\boldsymbol{x}|o_e) \otimes^{(\to 1)}_{\square;t} \mathbf{D}(R^{-1}\Delta R^{-1})\mathbf{D}(\Delta R)\, \boldsymbol{\varphi}_{\square;t}(g\, \boldsymbol{x}|o_s) && (\because \text{Eq. (3)}) \\ &= \boldsymbol{\psi}_{\square;t}(\boldsymbol{x}|o_e) \otimes^{(\to 1)}_{\square;t} \mathbf{D}(R^{-1}\cancel{\Delta R^{-1}\Delta R})\, \boldsymbol{\varphi}_{\square;t}(g\, \boldsymbol{x}|o_s) && (\because \text{Eq. (1)}) \\ &= \widetilde{\boldsymbol{s}}_{\square;t}(g, \boldsymbol{x}|o_s, o_e)\end{aligned}$$

$$\begin{aligned}&\widetilde{\boldsymbol{s}}_{\square;t}(g\,\Delta g^{-1}, \Delta g\, \boldsymbol{x}|o_s, \Delta g \cdot o_e) \\ &= \boldsymbol{\psi}_{\square;t}(\Delta g\, \boldsymbol{x}|\Delta g \cdot o_e) \otimes^{(\to 1)}_{\square;t} \mathbf{D}(\Delta R\, R^{-1})\, \boldsymbol{\varphi}_{\square;t}(g\,\cancel{\Delta g^{-1}}\cancel{\Delta g}\, \boldsymbol{x}|o_s) \\ &= \mathbf{D}(\Delta R)\, \boldsymbol{\psi}_{\square;t}(\boldsymbol{x}|o_e) \otimes^{(\to 1)}_{\square;t} \mathbf{D}(\Delta R\, R^{-1})\, \boldsymbol{\varphi}_{\square;t}(g\, \boldsymbol{x}|o_s) && (\because \text{Eq. (3)}) \\ &= \mathbf{D}(\Delta R)\, \boldsymbol{\psi}_{\square;t}(\boldsymbol{x}|o_e) \otimes^{(\to 1)}_{\square;t} \mathbf{D}(\Delta R)\, \mathbf{D}(R^{-1})\, \boldsymbol{\varphi}_{\square;t}(g\, \boldsymbol{x}|o_s) && (\because \text{Eq. (1)}) \\ &= \mathbf{D}_1(\Delta R) \left[\boldsymbol{\psi}_{\square;t}(\boldsymbol{x}|o_e) \otimes^{(\to 1)}_{\square;t} \mathbf{D}(R^{-1})\, \boldsymbol{\varphi}_{\square;t}(g\, \boldsymbol{x}|o_s)\right] && (\because\ [\mathbf{D}(R)\boldsymbol{v}] \otimes^{(\to l)} [\mathbf{D}(R)\boldsymbol{w}] = \mathbf{D}_l(R)\left[\boldsymbol{v} \otimes^{(\to l)} \boldsymbol{w}\right]) \\ &= \Delta R\, \widetilde{\boldsymbol{s}}_{\square;t}(g, \boldsymbol{x}|o_s, o_e)\end{aligned}$$

where in the last line we assume that the degree-1 Wigner D-matrix $\mathbf{D}_1(\cdot)$ is in the real basis with $x - y - z$ axis ordering. Note that the last line only holds in this specific choice of basis. Therefore, the type-1 or higher descriptors of the two EDFs must be defined in this basis. □

D. Implementation

In this section, we first provide the specific implementation of the bi-equivariant diffusion origin selection mechanism, which was postponed in Sec. 3.3. We then provide a novel multiscale EDF architecture, and the query points model. Further details such as non-dimensionalization and denoising schedule are then provided.

D.1. Diffusion Origin Selection Mechanism

For most manipulation tasks, specific local sub-geometries are more significant than the global geometry of the target object in determining its pose. Several works have addressed the importance of incorporating such locality in equivariant methods [8, 13, 17, 27, 42]. In manipulation tasks, contact-rich sub-geometries are more likely to be important than the others. We exploit this property by selecting the origin of diffusion near contact-rich sub-geometries.

Let $n_r(\boldsymbol{x}, o)$ be the number of points in a point cloud o that is within a contact radius r from a point $\boldsymbol{x} \in \mathbb{R}^3$. We use the following diffusion origin selection mechanism with r as a hyperparameter.

$$P\left(\boldsymbol{p}_{ed}|g_0^{-1} \cdot o_s, o_e\right) \propto \sum_{\boldsymbol{p} \in O_e} n_r\left(\boldsymbol{p},\ g_0^{-1} \cdot o_s\right) \delta^{(3)}(\boldsymbol{p}_{ed} - \boldsymbol{p}) \tag{64}$$

where $\delta^{(3)}(\boldsymbol{p})$ is the Dirac delta function on $\mathbb{R}^3$. We find that this strategy enables our models to pay more attention to such contact-rich and relevant sub-geometries without explicit supervision.

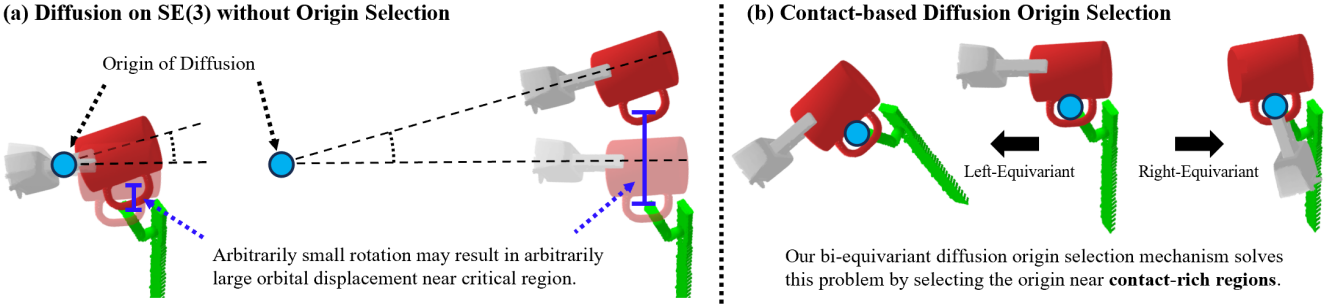

Figure 3. **Necessity of Diffusion Origin Selection Mechanism.** **(a)** A small rotational diffusion may result in arbitrarily large orbital displacement near the critical region depending on the diffusion origin. **(b)** We employ a contact-based diffusion origin selection mechanism. This not only allows bi-equivariant diffusion process but also stabilizes learning by minimizing the orbital impact of the rigid body rotation near the critical regions.

Necessity of Diffusion Frame/Origin Selection Mechanism. We first discuss why a diffusion frame/origin selection mechanism is necessary for our diffusion model on the $SE(3)$ manifold. For simplicity, we confine our argument only to the diffusion origin selection mechanism as Proposition 3 suggests.

In Sec. 3.3, we introduced the concept of diffusion frame/origin selection mechanism to achieve bi-equivariance in the diffusion process. However, the diffusion frame/origin selection has further implication, even for non-equivariant diffusion models on the $SE(3)$ manifold. As illustrated in Fig. 3, an arbitrarily small rotational perturbation may result in an arbitrarily large orbital displacement near the critical region depending on the choice of the origin, leading to an unstable diffusion and denoising process. This is in contrast to typical Euclidean diffusion models because vector addition is a commutative operation, and hence origin fixing has no effect. Therefore, a proper diffusion process for our problem must include a diffusion origin selection procedure to minimize the orbital effect of rotation near critical regions.

A natural selection of the diffusion origin for manipulation tasks is the origin of the end-effector frame itself. However, this origin selection is not equivariant to the grasped object, making our diffusion kernel only left-equivariant and not right-equivariant. Another natural diffusion origin is the centroid of the point cloud, which was utilized by Yim et al. [54] and Corso et al. [15] for protein docking problems. Indeed, this is a special case of an equivariant origin selection mechanism that satisfies Proposition 3. However, as pointed out by Ryu et al. [42] and Kim et al. [27], centroids are often dominated by the global geometry rather than the critical sub-geometry of the target objects. Please recall that this is why R-NDFs suffer without object segmentation. While the protein-ligand interaction problem in Yim et al. [54] and Corso et al. [15] has additional torsional degrees of freedom to debias this centroid artifact, it won't translate to our problem since the points in o_e are only actuated by the end-effector pose g.

Equivariant Diffusion Origin Selection Mechanism with Contact Heuristics. An important quality of a good diffusion origin selection mechanism is that the selected origin should not be too far away from the critical contact-rich region. As illustrated in Fig. 3, even a small rotational diffusion may take the critical region of the grasped object (the handle of the mug) far away from the placement target (the tip of the hanger), making training unstable. Although this problem can be resolved by reducing the rotational noise scale of the diffusion process, it requires meticulous task-specific hyperparameter tuning. Furthermore, as can be seen in Eq. (53), the rotational score consists of the pure rotational term and the orbital term. By studying the orbital term, one may notice that this term is non-dimensionalized by the product of the displacement term $\Delta \boldsymbol{p}$, which is proportional to the length unit, and the translational score $\boldsymbol{s}_\nu$, which is reciprocal to the length unit. Although these two dimensionful quantities neatly cancel out each other's unit, this structure inevitably increases the variance of the score estimation when the displacement term $\Delta \boldsymbol{p}$ is too large. For instance, a small translational score term in the reference frame of the critical region may induce a large rotational score term in the end-effector frame if the displacement $\Delta \boldsymbol{p}$ between these two frames is large. This is natural because a small rotation in the end-effector frame can dramatically change the probability of the pose if $\Delta \boldsymbol{p}$ is large. Therefore, it is always optimal to work in a diffusion origin near the critical region, such that $\Delta \boldsymbol{p}$ is kept minimal. This is the reason why we propose a contact-based diffusion origin selection mechanism in Eq. (64), which selects the origin near the important contact-rich sub-geometries.

We find that this origin selection mechanism stabilizes training by enabling Diffusion-EDFs to correctly identify important contact rich sub-geometries from the grasp observation o_e. This can be verified by visualizing the strength of the query

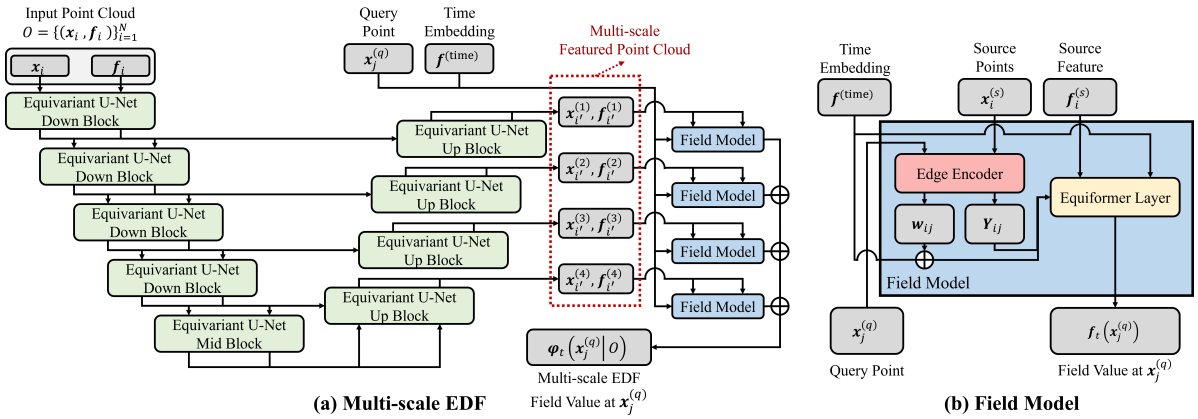

Figure 4. **Architecture of multiscale EDF.** Our multiscale EDF model is composed of a feature extracting part and a field model part. See Fig. 5 for details of each module in the architecture. **(a)** The feature extractor encodes the input point cloud into multiscale featured point clouds. We use an U-Net-like GNN architecture for the feature extractor part. **(b)** The encoded multiscale point clouds are passed into the field model part along with the query point and time embedding. The field model outputs the time-conditioned EDF field value at the query point. We simply sum up the output from each scale to obtain the EDF field value at the query point.

weight field. Fig. 6 illustrates the query points in colors according to their query weights. Query points with high weights are represented in cyan and those with near-zero weights are in black. As can be seen in the figure, the query weight field of the trained Diffusion-EDFs successfully assign high weight to the mug's handle, which is the most significant sub-geometry when placing it on a hanger.

D.2. Architecture of Equivariant Descriptor Fields

For faster sampling, we separate our implementation of EDFs into the feature extractor and the field model (see Fig. 4) as Ryu et al. [42] and Chatzipantazis et al. [8]. The feature extractor is a deep $SE(3)$-equivariant GNN encoder that is run only once at the beginning of the denoising process. On the other hand, the field model is much shallower and faster GNN that is utilized for each denoising step. It takes the encoded feature points from the feature extractor as input and computes the field value at a given query point.

For denoising, the receptive field of our model should cover the whole scene. However, the original EDFs [42] have small receptive fields due to memory constraints. We address this issue with our U-Net-like multiscale architecture, which maintains a wide receptive field without losing local high-frequency details. This increased receptive field enables Diffusion-EDFs to understand scene-level context.

In our multiscale EDF architecture, we use smaller message passing radius for small-scale points and larger radius for large-scale points. To keep the number of graph edges constant, we apply point pooling to larger-scale points with *Farthest Point Sampling* (FPS) algorithm [41]. For the field model, we find that a single layer is sufficient, although it is possible to stack multiple layers as Chatzipantazis et al. [8]. We use Equiformer [32] as the $SE(3)$-equivariant backbone GNN, with the addition of skip connections through point pooling layers. We illustrate our multiscale UNet-like architecture in Fig. 4. See Fig. 5 for the illustration of each module used in Fig. 4.

D.3. Score Model

We use the weighted query points model similar to Ryu et al. [42] for $\rho(\boldsymbol{x}|O)$

$$\rho(\boldsymbol{x}|o_e) = \sum_{\boldsymbol{q}\in Q(O_e)} w(\boldsymbol{x}|o_e)\delta^{(3)}(\boldsymbol{x}-\boldsymbol{q}) \tag{65}$$

where $Q(\cdot) : o_e \mapsto \{\boldsymbol{q}_n\}_{n=1}^{N_q}$ is the *query points function* which outputs the set of N_q query points, and $w(\cdot|o_e) : \mathbb{R}^3 \to \mathbb{R}_{\geq 0}$ is the *query weight field* that assigns weights to each query point. The query points function and query weight field are $SE(3)$-equivariant such that

$$\begin{aligned} Q(\Delta g \cdot o_e) &= \{\Delta g\, \boldsymbol{q}_n | \boldsymbol{q}_n \in Q(o_e)\} && \forall \Delta g \in SE(3) \\ w(\boldsymbol{x}|o_e) &= w(\Delta g\, \boldsymbol{x}|\Delta g \cdot o_e) && \forall \Delta g \in SE(3) \end{aligned}$$

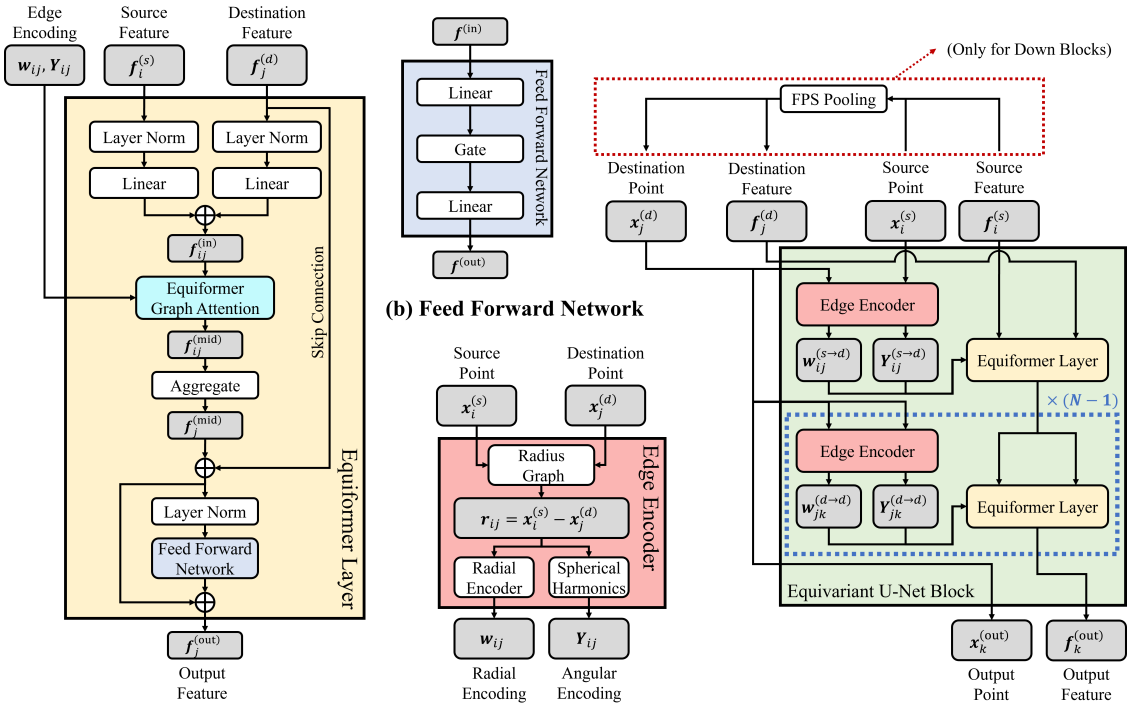

Figure 5. **Overview of Modules Used in Multiscale EDF. (a)** We employ Equiformer [32] to achieve $SE(3)$-equivariance in our model. **(b)** We use an equivariant feed forward network with gate activation from Equiformer. **(c)** We use radius graph to construct graph from points. Graph edge length and orientation are respectively encoded by a radial encoder and spherical harmonics [18, 32, 50]. **(d)** Multiple equiformer layers are stacked and form the equivariant U-Net Block. FPS pooling is used in downward blocks to obtain coarse-grained destination points from source points in lower scale-space.

We use FPS algorithm for $Q(o_e)$. Although it is not strictly deterministic, we observe negligible impact from this stochasticity. For the implementation of the query weight field $w(\boldsymbol{x}|o)$, we use an EDF with a single scalar (type-0) output. With this query points model, Eq. (22) and Eq. (23) become tractable summation forms

$$\boldsymbol{s}_{\nu;t}(g|o_s,o_e) = \sum_{\boldsymbol{q}\in Q(O_e)} w(\boldsymbol{q}|o_e)\,\widetilde{\boldsymbol{s}}_{\nu;t}(g,\boldsymbol{q}|o_s,o_e) \tag{66}$$

$$\boldsymbol{s}_{\omega;t}(g|o_s,o_e) = \sum_{\boldsymbol{q}\in Q(O_e)} w(\boldsymbol{q}|o_e)\,\left[\widetilde{\boldsymbol{s}}_{\omega;t}(g,\boldsymbol{q}|o_s,o_e) + \boldsymbol{q}\wedge\widetilde{\boldsymbol{s}}_{\nu;t}(g,\boldsymbol{q}|o_s,o_e)\right] \tag{67}$$

Nondimensionalization. We assume that the output of the score field model in Eq. (27) is a dimensionless quantity. Therefore, we obtain the dimensionful score by taking

$$\widetilde{\boldsymbol{s}}_{\nu;t} \to \frac{1}{L\sqrt{t}}\widetilde{\boldsymbol{s}}_{\nu;t}, \qquad \widetilde{\boldsymbol{s}}_{\omega;t} \to \frac{1}{\sqrt{t}}\widetilde{\boldsymbol{s}}_{\omega;t}$$

where L is the characteristic length scale unit. The reason for dividing $1/\sqrt{t}$ is because the norm of the target score tend to scale with $O(1/\sqrt{t})$. Likewise, we divide the linear score field by L because score field is a gradient and thus scales reciprocally to the characteristic length scale.

For computational efficiency, we use identical EDFs for $\square = \omega, \nu$ in Eq. (27). In addition, we remove the time dependence of the grasp EDF $\psi_t(\boldsymbol{x}|o_e)$ so that its field value is computed only once at the beginning of the denoising process. In

(a) Query points of a real-world mug observation.

(b) Query points of a real-world bottle observation.

Figure 6. **Learned Query Points**. The figure depicts the point clouds of a real mug and bottle with their query points visualized in colors according to their weights. The query points with the highest weight values are illustrated in cyan. The query weight field of the trained Diffusion-EDFs assigns high weight to **(a)** the mug's handle, which is the most significant sub-geometry when placing it on a hanger, and **(b)** the bottom of the bottle, which is the most significant sub-geometry when placing it on a shelf.

conclusion, our actual implementations of Eq. (66) and Eq. (67) are as follows:

$$\boldsymbol{s}_{\nu;t}(g|o_s,o_e) = \frac{1}{L\sqrt{t}} \sum_{\boldsymbol{q}\in Q(O_e)} w(\boldsymbol{q}|o_e) \left[\boldsymbol{\psi}(\boldsymbol{q}|o_e) \otimes_{\nu;t}^{(\to 1)} \mathbf{D}(R^{-1})\, \boldsymbol{\varphi}_t(g\,\boldsymbol{q}|o_s)\right] \tag{68}$$

$$\begin{aligned}\boldsymbol{s}_{\omega;t}(g|o_s,o_e) = &\frac{1}{\sqrt{t}} \sum_{\boldsymbol{q}\in Q(O_e)} w(\boldsymbol{q}|o_e)\, \frac{\boldsymbol{q}}{L} \wedge \left[\boldsymbol{\psi}(\boldsymbol{q}|o_e) \otimes_{\nu;t}^{(\to 1)} \mathbf{D}(R^{-1})\, \boldsymbol{\varphi}_t(g\,\boldsymbol{q}|o_s)\right] \\ &+ \frac{1}{\sqrt{t}} \sum_{\boldsymbol{q}\in Q(O_e)} w(\boldsymbol{q}|o_e) \left[\boldsymbol{\psi}(\boldsymbol{q}|o_e) \otimes_{\omega;t}^{(\to 1)} \mathbf{D}(R^{-1})\, \boldsymbol{\varphi}_t(g\,\boldsymbol{q}|o_s)\right]\end{aligned} \tag{69}$$

D.4. Sampling with Annealed Langevin Dynamics

It is known to be difficult and unstable to train and sample with the score function for a sparse distribution [28, 49]. To address this issue, *Annealed Langevin Markov Chain Monte Carlo* [49] leverages the score of the diffused marginal P_t instead of P_0. A diffused marginal $P_t(g)$ for a diffusion kernel $P_{t|0}(g|g_0)$ is defined on the $SE(3)$ manifold as

$$P_t(g) = \int_{SE(3)} dg_0 P_{t|0}(g|g_0) P_0(g_0). \tag{70}$$

We utilize the trained score function $s_t(g) = \nabla \log P_t(g)$ for the annealed Langevin MCMC on $SE(3)$ [51] as

$$g_{\tau+d\tau} = g_\tau \exp\left[\frac{1}{2}\boldsymbol{s}_{t(\tau)}(g_\tau|o_s,o_e)d\tau + dW\right]. \tag{71}$$

where $t(\tau)$ is the diffusion time scheduling, which is gradually annealed to zero as $\tau \to \infty$, such that $t(\tau=\infty)=0$. This process will converge to $P_0(g)$ regardless of the initial distribution if it is annealed sufficiently slowly and $\lim_{t\to 0} P_t = P_0$. This SDE can be discretized using the forward Euler-Maruyama method such that

$$g_{n+1} = g_n \exp\left[\frac{1}{2}\boldsymbol{s}_{t[n]}(g_n|o_s,o_e)\alpha[n] + \sqrt{\alpha[n]}\boldsymbol{z}_n\right], \quad \boldsymbol{z}_n \sim \mathcal{N}(\mathbf{0}, I) \tag{72}$$

where $t[n]$ and $\alpha[n]$ are respectively the diffusion time and Langevin step size, both of which are scheduled according to the step count n. A commonly used scheduling scheme is taking $\alpha[n] \propto t[n]$ with either a linear or log-linear $t[n]$ schedule [21, 49, 51]. However, the convergence is very slow with this scheduling. Therefore, we use $\alpha[n] \propto t[n]^{k_1}$ schedule with

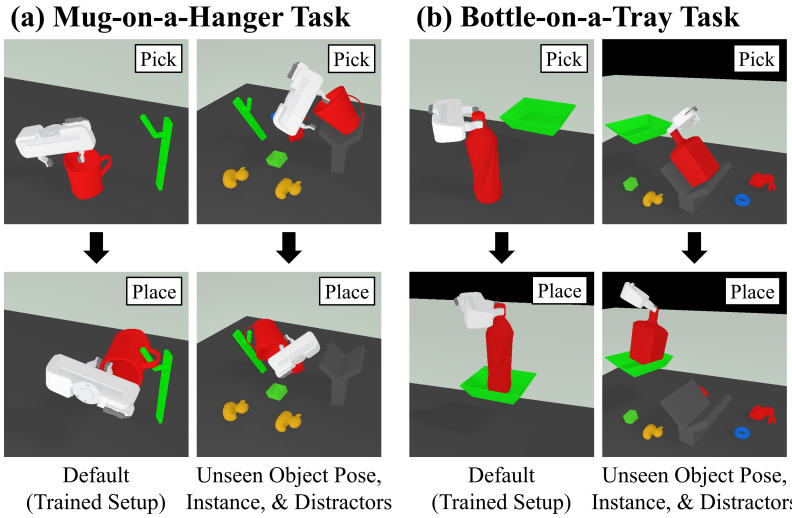

Figure 7. **Simulation Experiments.** (a) In the *Mug-on-a-Hanger* task, a red mug should be picked up by its rim and placed on a green hanger by its handle. (b) In the *Bottle-on-a-Tray* task, a red bottle should be picked up by its cap and placed on a green tray.

a hyperparameter $k_1 < 1$. To suppress the instability caused by large step sizes when t is small, we also gradually lower the *temperature*[2] of the process. This can be done by using $\sqrt{\alpha[n]T[n]}\boldsymbol{z}_n$ instead of $\sqrt{\alpha[n]}\boldsymbol{z}_n$ for the noise term with the temperature schedule $T[n] = t[n]^{k_2}$, where $k_2 \geq 0$ is another hyperparameter. Intuitively, this makes the sampling process to smoothly transition into a simple gradient descent optimization as $t[n] \to 0$, and hence $T[n] \to 0$. We empirically found that this strategy significantly improves the convergence time without compromising the accuracy and diversity of the sampled poses. The resulting sampling algorithm with a small number ϵ is

$$g_{n+1} = g_n \exp\left[\frac{\epsilon}{2}\boldsymbol{s}_{t[n]}(g_n|o_s, o_e)t[n]^{k_1} + \sqrt{\epsilon}\, t[n]^{\frac{k_1+k_2}{2}}\boldsymbol{z}_n\right], \quad \boldsymbol{z}_n \sim \mathcal{N}(\mathbf{0}, I) \tag{73}$$

We use $k_1 = 0.5$ and $k_2 = 1.0$ for the step size and temperature scheduling. For the diffusion time $t[n]$, we use piecewise linear scheduling. For example, we linearly schedule the diffusion time for $t = 1$ to $t = 0.1$ and then with $t = 0.1$ to $t = 0.01$. Similar to diffusion-based image generation models, we separate a low-resolution model and high-resolution model instead of using a single model. We use the low-resolution model for higher t and the high-resolution model for lower t. Similar to Ryu et al. [42], we solve Eq. (73) in the quaternion-translation parameterization of $SE(3)$ instead of performing the actual exponential mapping in Eq. (73).

E. Experiments and Results

E.1. Simulation Benchmarks.

We compare diffusion-EDFs with a state-of-the-art $SE(3)$-equivariant method (R-NDFs [47]) and a state-of-the-art denoising diffusion-based method ($SE(3)$-Diffusion Fields [51]) under an evaluation protocol similar to Simeonov et al. [46, 47], Ryu et al. [42], and Biza et al. [2]. In particular, we measure the pick-and-place success rate for two different object categories: mugs and bottles (see Fig. 7). We assess the generalizability of each method under four previously unseen scenarios: 1) novel object instances, 2) novel object poses, 3) novel clutters of distracting objects, and 4) all three combined.

All the models are trained with ten task demonstrations performed by humans. We train Diffusion-EDFs in a fully end-to-end manner without using any pre-training or object segmentation. In contrast, we evaluate R-NDFs and $SE(3)$-Diffusion Fields for both with and without object segmentation pipelines. For $SE(3)$-Diffusion Fields, we use rotational augmentation as they lack $SE(3)$-equivariance. For R-NDFs, we additionally use category-specific pre-trained weights from the original implementation [47]. It took 20∼45 minutes to train Diffusion-EDFs for single pick or place task with RTX 3090 GPU and i9-12900k CPU.

As shown in Tab. 2, Diffusion-EDFs consistently outperform both the $SE(3)$-equivariant baseline (R-NDFs [47]) and diffusion model baseline ($SE(3)$-DiffusionFields [51]) in almost all scenarios, despite not being provided with pre-training

[2]This temperature annealing should not be confused with that of the *'annealed'* Langevin MCMC in which the diffusion time t is decreased.

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 [47]	✗	✗	✓	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 [51]	✓	✗	✗	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 [47]	✗	✗	✓	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 [51]	✓	✗	✗	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 [47]	✗	✗	✓	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 [51]	✓	✗	✗	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 [47]	✗	✓	✓	0.00	0.00	0.00	0.00	0.00	0.00
	SE(3)-DiffusionFields [51]	✓	✓	✗	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 [47]	✗	✗	✓	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 [51]	✓	✗	✗	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.

Table 2. Pick-and-place success rates in various out-of-distribution settings in simulated environment.

or segmented inputs. In particular, the baseline models completely fail with unsegmented observations. Without object segmentation, R-NDFs achieve zero success rates due to the lack of locality in their method design [13, 27, 42]. While slightly better than R-NDFs, $SE(3)$-DiffusionFields also record low success rates, presumably due to the lack of $SE(3)$-equivariance. On the other hand, Diffusion-EDFs maintain total success rates around 80% even in the most adversarial scenarios due to the local equivariance [27, 42] inherited from EDFs and our local contact-based diffusion frame selection mechanism.

E.1.1 Simulation Environment

Evaluations are performed in a simulated enviroment using SAPIEN [53] with nine ceiling-mounted depth cameras. We assume a perfect observation to remove the influence of point cloud processing pipelines, which is orthogonal to our research. We also remove the impact of robot's kinematic constraints by using a floating gripper-only robot instead of simulating the full robot. In addition, we turn off the collision between the environment and allow the robot to teleport to the pre-pick/place pose in order to get rid of failures related to motion planning. We evaluate the success of pick or place by turning off the collision between the environment (including the table) and the target object to manipulate, and measuring the object's z-axis position. If the object is not firmly grasped by the gripper or is not placed on the intended placement target, the object will fall after removing the environmental collision. Therefore, we measure the z-axis position to automatically assess whether the object has not fallen, meaning that the manipulation has succeeded.

E.1.2 Method Details

For each task, we train the models using ten human-generated demonstrations, in which five object instances in only upright poses are used. In other words, each of the five object instances is demonstrated for two different pick/place poses. In the

training data set, we do not use distracting objects. We used a custom-built web-based GUI to collect human demonstrations.

Diffusion-EDFs. We only use ten human demonstrations to train Diffusion-EDFs in a fully end-to-end manner. No additional prior knowledge such as pre-training, object segmentation, pose estimation or data augmentation is used for Diffusion-EDFs. For preprocessing, we use simple voxel downsampling to reduce the number of points.

R-NDFs. For R-NDFs, we use the pre-trained weights from the original implementation of Simeonov et al. [47]. These weights were trained with a self-supervised learning method that relies on massive amount (150 gigabytes) object geometry that are specific to the target object categories (mug, bowl, bottle; 50 gigabytes for each). Although we do not use bowls in our experiment, we still use the weights trained from all three object categories, which achieve better performance than weights trained from only a single object category [46, 47]. Still, we observe that R-NDFs fail to place the mug on our mug hanger. We presume that this is due to the discrepancy of the hanger's shape in our experiment and the ones used for pre-training, which were procedurally generated [47]. Therefore, we do R-NDFs an additional favor of using the pre-trained hanger instances instead of our hanger for the evaluation. Lastly, we also tried to naively pre-train the NDFs using the reconstructed meshes from the point clouds in our ten task demonstrations, but resulted in suboptimal performance (less than 5% success rate). These attempts show the importance of the end-to-end trainability of EDFs [42] and Diffusion-EDFs. R-NDFs cannot be used for uncommon object categories, as they require immense amount of category-specific data for pre-training. Procedural generation has also turned out to be unable to resolve this problem because it cannot cover all variations in the category, which was evident in the case of the mug hanger mentioned above.

We also evaluate R-NDFs both with and without object segmentation. It should be noted that the ability to infer without object segmentation is important not only because of its convenience. This allows the model to understand *scene-level contexts* beyond a single target object, which is further evidenced by our real hardware experiments in Sec. E.2. The experimental results in Tab. 2 clearly show that R-NDFs are unable to make inference without object segmentation. As pointed out by Ryu et al. [42], we presume this is because of the violation of locality in R-NDFs, such as centroid subtraction.

SE(3)-Diffusion Fields. In contrast to R-NDFs, we train $SE(3)$-Diffusion Fields [51] using only the ten demonstrations as Diffusion-EDFs. Following Urain et al. [51], we jointly train the model to match both the signed distance function and the score function. We specifically use the *PoiNt-SE(3)-DiF* variant in the original paper [51]. Although this model utilizes $SO(3)$-equivariant point cloud encoder based on VN-PointNet [16], the overall architecture is not equivariant. Therefore, we use $SO(3)$ rotational data augmentation to complement the lack of equivariance.

Similar to R-NDFs, we evaluate $SE(3)$-Diffusion Fields both with and without object segmentation. With object segmentation, $SE(3)$-Diffusion Fields could learn to pick up the target object, although the success rates are much lower than Diffusion-EDFs. Without object segmentation, they achieve success rates lower than 15% for all scenarios.

E.2. Real Hardware Experiment Details

We further evaluate our Diffusion-EDFs on three real-world tasks: the *mug-on-a-hanger* task, *bowls-on-dishes* task, and *bottles-on-a-shelf* task. We illustrate these tasks in Fig. 9, and the experiment pipeline in Fig. 8.

The mug-on-a-hanger task is similar to the one in the simulation benchmark. In this task, even a minor error of a centimeter can result in complete failure due to noisy observation and the small size of mug handles. In addition, the placement pose heavily depends on the posture of the grip, requiring full 6-DoF inference capability. We also experiment with novel objects in oblique poses that were not presented during training. Diffusion-EDFs successfully learned to solve this task from only ten human demonstrations, demonstrating their ability to perform 1) accurate 6-DoF manipulation tasks with 2) previously unseen object instances and 3) out-of-distribution poses.

In the bowls-on-dishes task, the robot should pick up the bowls and place them on the dishes of matching colors in red-green-blue order. Note that this sequential task requires scene-level comprehension, which is impossible for methods that rely on object segmentation. For example, the robot should not pick up the blue bowl unless the red and green bowls are already on the dishes. Diffusion EDFs successfully learned to solve this sequential task (in correct order) from only ten human demonstrations, which consists of red, green, and blue subtasks. This validates Diffusion-EDFs' ability to 1) solve sequential problems; 2) understand scene-level contexts; and 3) process color-critical information.

Lastly, in the bottles-on-a-shelf task, the robot should pick up multiple bottles one by one and place them on a shelf. In this task, we provide three identical bottle instances for both training and evaluation. Non-probabilistic methods such as R-NDFs are known to suffer from such multimodalities in the task [48]. Methods that depend on object segmentation are also unable

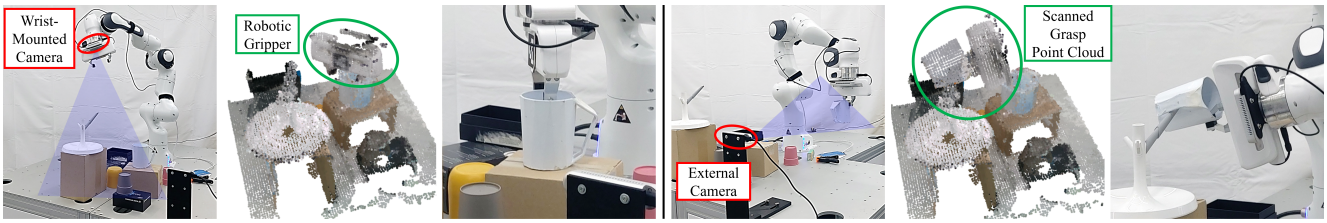

Figure 8. **Real Hardware Experiment Pipeline** 1) The scene point cloud is observed via 3D SLAM algorithm with the wrist-mounted RGB-D Camera. 2) Diffusion-EDFs infer the gripper pose to pick up the target object. 3) The robot executes picking if the pose is reachable. 4) The grasp point cloud is scanned with an external RGB-D camera. 5) Diffusion-EDFs infer the gripper pose to place the grasped object on the placement target. 6) The robot executes placement if the pose is reachable. See Supp. E.2 for more details.

to solve this task, as they cannot differentiate between bottles that are already placed on the shelf and those that are not. To evaluate generalization, we also experiment with object instances and quantities that were not presented during training. Diffusion-EDFs successfully learned the task from four human demonstrations (consisting of three sequential pick-and-place subtasks for each bottle), showcasing their robustness to stochastic and multimodal tasks.

In conclusion, our experiments demonstrate that Diffusion-EDFs are capable of: 1) accurately generating 6-DoF poses; 2) understanding scene-level contexts; 3) learning from stochastic demonstrations; and 4) generalizing to novel object instances and poses in real-world robotic manipulation, despite being trained with a limited number of demonstrations. We summarize the key challenges of each task in Tab. 3. For the experimental results, please refer to the supplementary video at the project website: https://sites.google.com/view/diffusion-edfs.

E.2.1 Experimental Setup

We use a Franka Emika Panda robot arm with two Intel RealSense D415 RGB-D cameras. The first camera is attached to the wrist of the robot. The robot moves around the workspace to observe RGB-D images of the scene from multiple viewpoints. We employ RTAB-Map [30], a 3D SLAM technique, to convert these observations into a point cloud of the scene. Rather than relying on visual odometry, we take advantage of the forward kinematics solution from the robot's joint encoders, which is more precise. Although we use 3D SLAM-based approach in our experiments, this procedure can be skipped if multiple well-calibrated external cameras are available. The second camera is installed on the table to observe the point cloud of the robot's gripper. This external camera is calibrated to the ArUco marker [19] frame attached to the robot's end-effector. All the point clouds are post-processed using Open3D [57], in which we remove statistical outliers and apply voxel filtering. We also apply hue and lightness augmentation for the training data to obtain robustness under light condition changes.

In our experimental procedure, the robot first moves along a predefined trajectory to observe the scene. RTAB-Map is used to convert these observations into the point cloud of the scene in real time. Diffusion-EDFs take this point cloud to generate the end-effector poses to pick the target object. After picking the object, the robot moves to the predefined grasp observation pose. The robot then rotates its grasped object by 360°, and the external camera observes it. These observations are then registered into the grasp point cloud. For the scene point cloud, we use the same one that we used to infer the pick pose. With these two point clouds, Diffusion-EDFs infer the end-effector poses to place the grasped object onto the placement target. For the collection of human demonstrations, we follow a procedure similar to that in the aforementioned inference pipeline. The only difference is that the target pose demonstration is manually provided by a human instead of Diffusion-EDFs.

E.2.2 System Engineering

Motion Primitives. While it is theoretically possible to generate a collision-free motion plan for any reachable goal pose, it is challenging in reality due to the imprecise nature of point cloud observations. Therefore, determining how to approach the target pose is also an important problem. As we focus only on the problem of inferring the target pose itself in this work, we simply assume that we already have task-specific motion primitives to approach the generated goal pose. In all three real-world tasks, we use a simple motion primitive of picking along the end-effector's z-axis direction (the direction in which the gripper is pointing), and placing the target object in the top-down direction. The robot first moves to the pre-pick/place pose by following the collision-free trajectory found by an off-the-shelf motion planner. The motion primitives are then used to approach the generated target pick/place pose from the previous pre-pick/place pose. After successful picking or placing, we

Figure 9. **Real Hardware Experiments.** (a) In the *mug-on-a-hanger* task, the white mug must be picked and placed on the white hanger. (b) In the *bowls-on-dishes* task, the bowls must be picked and placed on the dishes of matching color in red-green-blue order. (c) In the *bottles-on-a-shelf* task, multiple bottles must be picked and placed on the shelf one by one. The experimental results can be found in the supplementary video at the project website: https://sites.google.com/view/diffusion-edfs.

initiate post-pick/place primitives. We simply lift up the end-effector for the post-pick primitive. For the post-place primitive, we retract the end-effector towards the opposite direction that was taken in the pre-pick manuever. We use MoveIt [14] for motion planning and use the TOPP-RA [40] algorithm to time-parameterize our waypoint-based motion primitives.

Although we use predefined motion primitives, not every problem can be solved in this way. Therefore, more general approach should also encompass learning not only the target pose but also the approach direction. We expect that our score model in Eq. (21) can be used for this purpose with slight modifications. The approach direction can be represented as the

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

Table 3. Key challenges of each task

displacement between the pre-pick/place pose and the target pose. This displacement can be effectively expressed as an $\mathfrak{se}(3)$ Lie algebra vector. Therefore, our score model can be modified to equivariantly infer this Lie algebra vector that represents the approach direction. We leave this research for future studies.

Energy-based Critic. Due to the collision and kinematic constraint of the robot, not every pose generated by Diffusion-EDFs are feasible. Although we ignored this problem in our simulation experiment, this problem must be considered in real robot applications. Therefore, similar to Urain et al. [51] and Ryu et al. [42], we generate multiple samples in parallel and reject infeasible poses one by one until a reachable pose is found.

However, it is difficult to ensure convergence for every generated sample as we use a limited number of Langevin steps to achieve reasonable inference time (5$\sim$17 seconds). The number of unconverged samples tend to be larger in our real-world experiment with noisy observations than in the simulated ones with perfect observations. Furthermore, rejecting infeasible poses often leads to the elimination of correct poses and the selection of unconverged wrong poses. Urain et al. [51] and Ryu et al. [42] circumvented this problem by sorting the generated samples according to the learned energy function, which evaluates the quality of the generated poses. In contrast to these works, however, our method does not have an explicit scalar function that can be utilized.

Therefore, we train an auxiliary energy function to sort the generated poses according to their quality. We first modify the bi-equivariant energy function of Ryu et al. [42] to allow diffusion time conditioning. We then take the Lie derivatives to obtain the energy-based score model similar to Urain et al. [51]. This energy-based score model is trained using the loss function in Eq. (19) with proper non-dimensionalization. Although this score-matching model is far less accurate than our original model in Eq. (21) due to the inflexible nature of energy-based diffusion models, the trained energy function is sufficient to distinguish between unconverged samples and converged samples.

With the learned energy function, we first sort the generated samples according to their energy value. If the energy function is well trained, lower-energy samples should be better than higher-energy samples. However, in contrast to the MCMC-based training of Ryu et al. [42], our diffusion-based energy function training does not have a contrastive mechanism to penalize the model for assigning low energy to outlier poses. Therefore, our energy function often assigns too low energy values to outlier poses, although the training is much faster. Nevertheless, we find that simply rejecting too-low-energy outliers effectively solves this problem. Therefore, we remove the first few samples from the sorted list and start from samples with moderately low energy. We then try motion planning for each sample until a feasible pose is found. This strategy drastically improves the success rate of pick-and-place tasks in our real-world tasks.

E.2.3 Experimental Results Details

Note that it is difficult to precisely measure the performance of Diffusion-EDFs for real-world tasks as the success rate is determined not only by the inference quality but also the quality of observation, localization, and motion planning. For instance, noisy observation and localization cause success rates to drop for subtasks that require high precision, such as mug placement and bottle picking, even though Diffusion-EDFs accurately generated correct target poses. Challenges associated with motion planning can also reduce the success rate, particularly for subtasks that require difficult 6-DoF manipulation, such as mug placement. We achieve over 90% success rate for all subtasks except the mug placement and bottle picking. For these two tasks, the success rates are roughly around 80%. The majority of the errors in these tasks were caused by a slight lack of accuracy in the position that was less than a centimeter. Note that these real hardware success rates may largely differ across systems, depending on the quality of observation, calibration, motion planning and control pipelines, which are orthogonal to our research. The experimental results with our real robot manipulator system can be found in the supplementary video[3]. In the video, our robot performs 5 to 6 pick-and-place operations in one take without failure, showcasing that Diffusion-EDFs can solve all three real-world tasks with high success rates.

For more reproducible results, we also provide example input data and codes that we used to generate end-effector poses for the three real-world tasks with Diffusion-EDFs. These supplementary materials can provide an idea of Diffusion-EDFs'

[3]Supplementary video can be found at the project website: https://sites.google.com/view/diffusion-edfs.

pure inference performance for noisy real-world observations without the complications related to motion planning and localization. The samples generated by Diffusion-EDF for the mug-on-hanger and bowls-on-dishes tasks are illustrated in Figs. 10 and 11, respectively. The samples generated by Diffusion-EDF for the bottles-on-shelf task are illustrated in Figs. 12 and 13. Diffusion-EDFs combined with the energy-based critic in Sec. E.2.2 can successfully infer appropriate poses for all these tasks in more than 90% of the cases, although it is important to note that this success rate is subject to human evaluation and may vary based on the individual's criteria.

For mugs and bottles, it takes 5$\sim$6 seconds to generate 20 poses for picking, and 9$\sim$10 seconds to generate 10 poses for placing. For bowls, it takes 7 seconds to generate 20 poses for picking and 17 seconds to generate 10 poses for placing. The sampling is slower for the bowls-on-dishes task because the point clouds in this task have more points than in the other tasks. As mentioned in Sec. D.4, we use two different models for low-resolution and high-resolution denoising. In addition, we use the energy-based critic to sort the sampled poses according to their quality. Therefore, three different models must be trained for each pick and place tasks. It takes less than 24 minutes to train each model for mug-picking and less than 36 minutes for mug-placing with an RTX3090 GPU. The bottles-on-a-shelf task requires slightly longer training time, amounting to 27 minutes for picking and 43 minutes for placing with an RTX3090 GPU. The bowls-on-dishes task requires a much longer training time because it consists of three different subtasks. It takes less than 47 minutes of training for picking and less than 1.3 hours for the placing. Note that the three models can be trained in parallel. Therefore, it takes less than an hour with three RTX3090 GPU to train our method for all tasks except for the bowl-placing task.

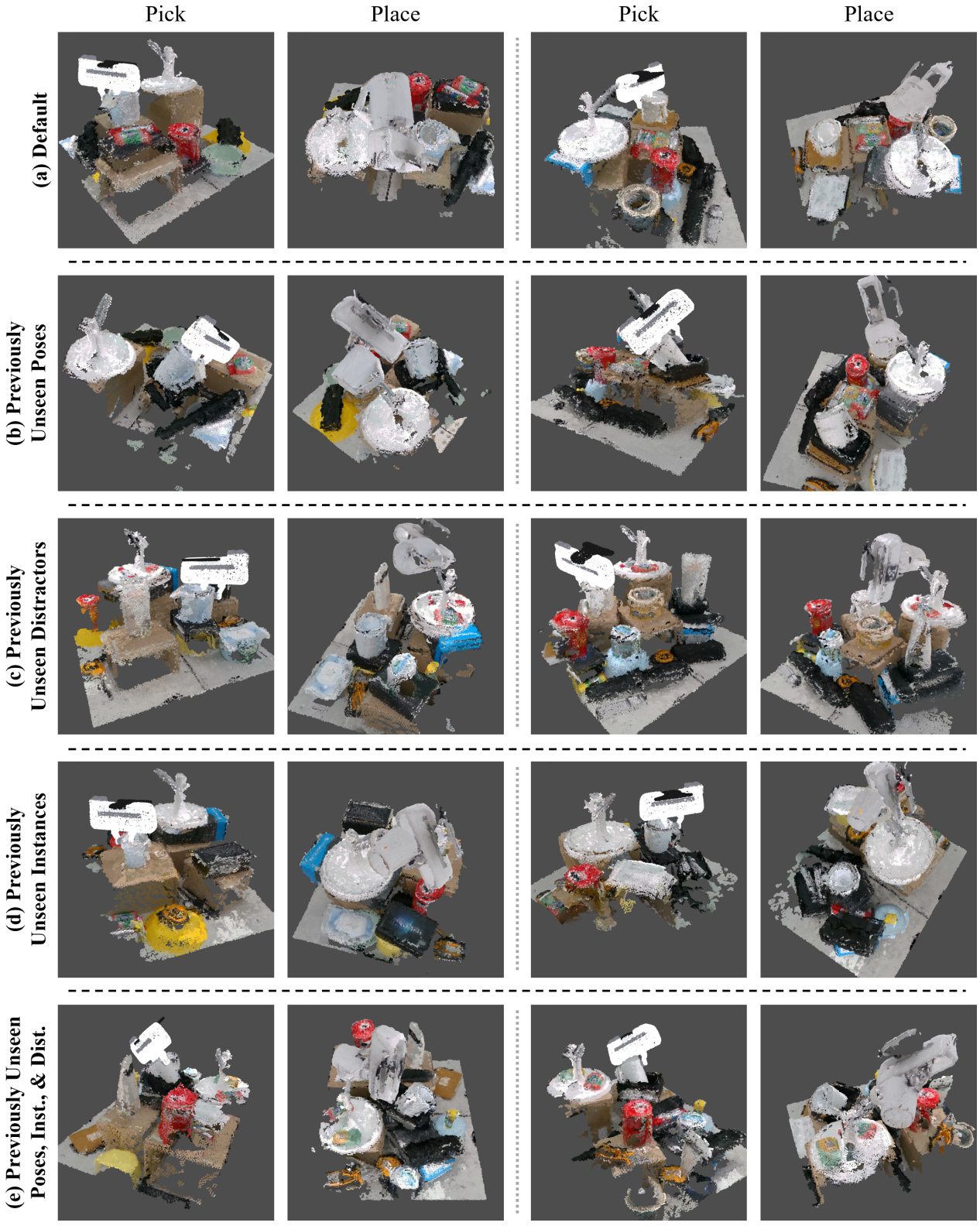

Figure 10. **Samples Generated by Diffusion-EDFs for Real-world Mug-on-a-hanger Task.** The figure depicts the end-effector pose samples for picking and placing a white mug on a white mug hanger. Diffusion-EDFs trained with only ten human demonstrations generated these samples from the real-world point cloud observations of the scene and grasp. Similar to our simulation experiments, we experiment for the **(a)** default scenario, **(b)** previously unseen target object poses (oblique; note that we only trained Diffusion-EDFs for upright poses) scenario, **(c)** previously unseen adversarial distractors (in white color) scenario, **(d)** previously unseen target object instances scenario, and **(e)** the all scenarios combined. The denoising diffusion process can be found in the supplementary video at https://sites.google.com/view/diffusion-edfs.

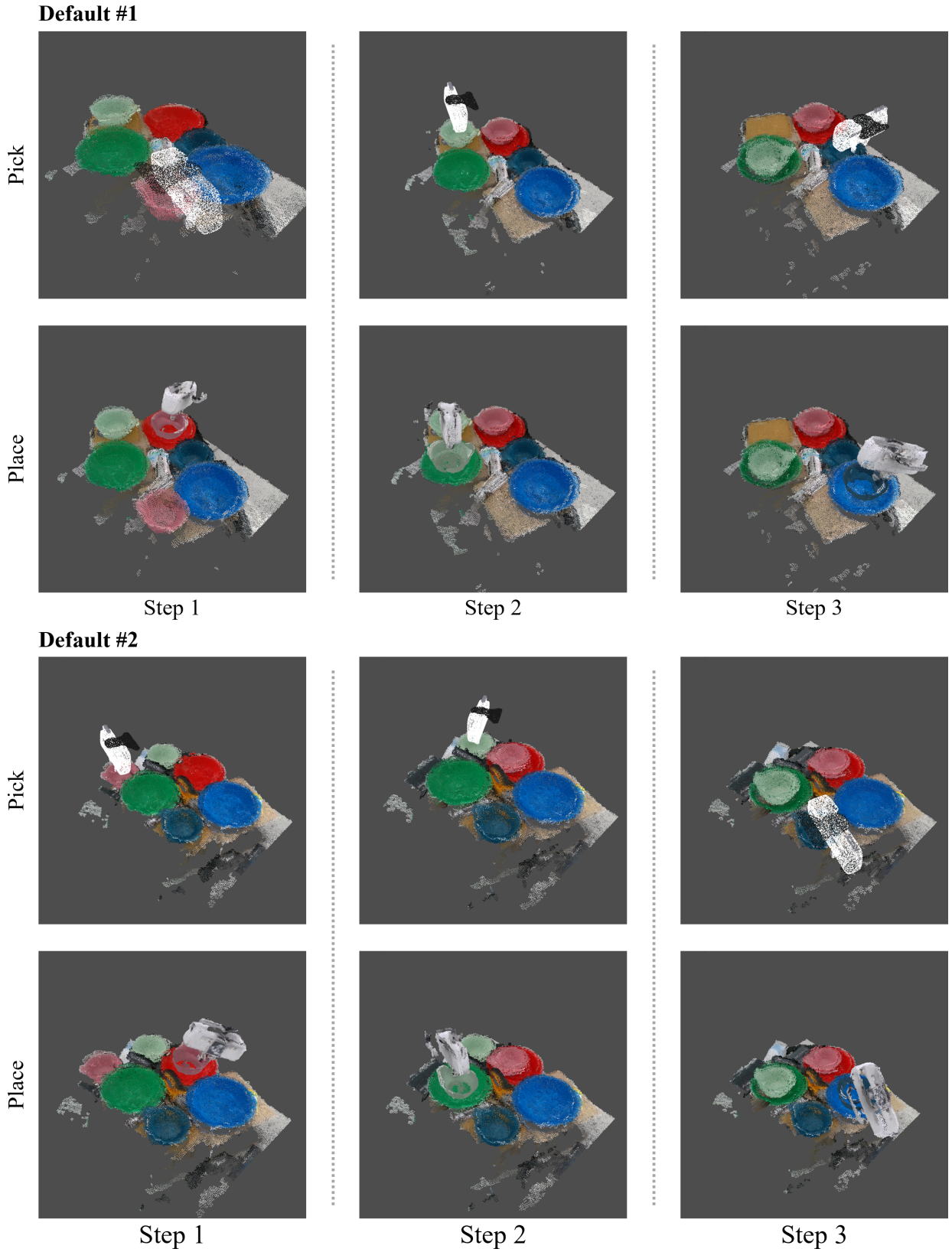

Figure 11. **Samples Generated by Diffusion-EDFs for Real-world Bowls-on-dishes Task.** The figure depicts the end-effector pose samples for picking and placing bowls on the dishes of matching colors in red-green-blue order. Diffusion-EDFs trained with only ten human demonstrations (three colored subtasks for each) generated these samples from the real-world point cloud observations of the scene and grasp. The denoising diffusion process can be found in the supplementary video at https://sites.google.com/view/diffusion-edfs.

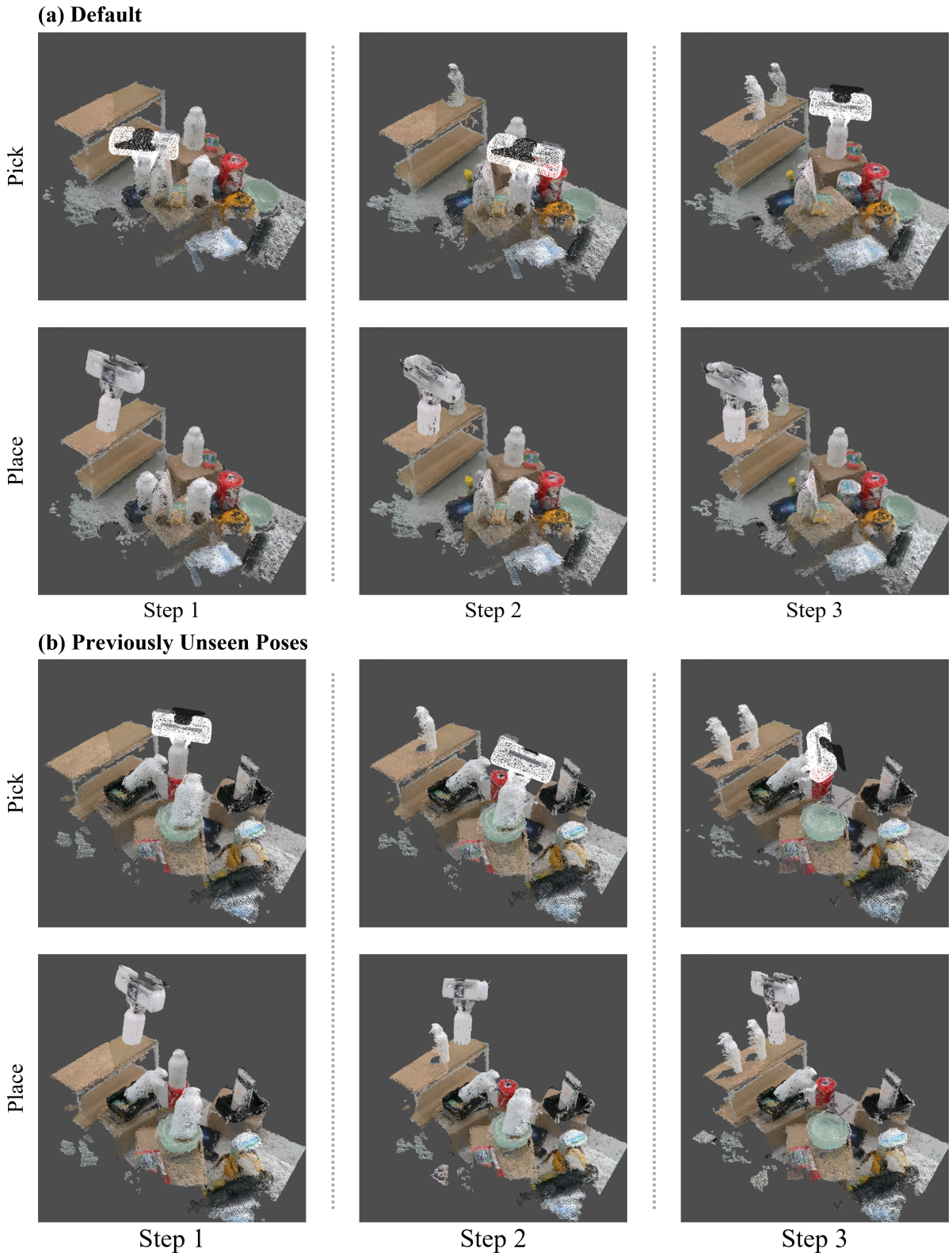

Figure 12. **Samples Generated by Diffusion-EDFs for Real-world Bottles-on-a-shelf Task.** The figure depicts the end-effector pose samples for picking and placing multiples bottles on a shelf. Diffusion-EDFs trained with only four human demonstrations (three sequential subtasks for each) generated these samples from the real-world point cloud observations of the scene and grasp. Similar to our simulation experiments, we experiment for the **(a)** default scenario and **(b)** previously unseen target object poses (oblique; note that we only trained Diffusion-EDFs for upright poses) sceneraio. The denoising diffusion process can be found in the supplementary video at https://sites.google.com/view/diffusion-edfs.

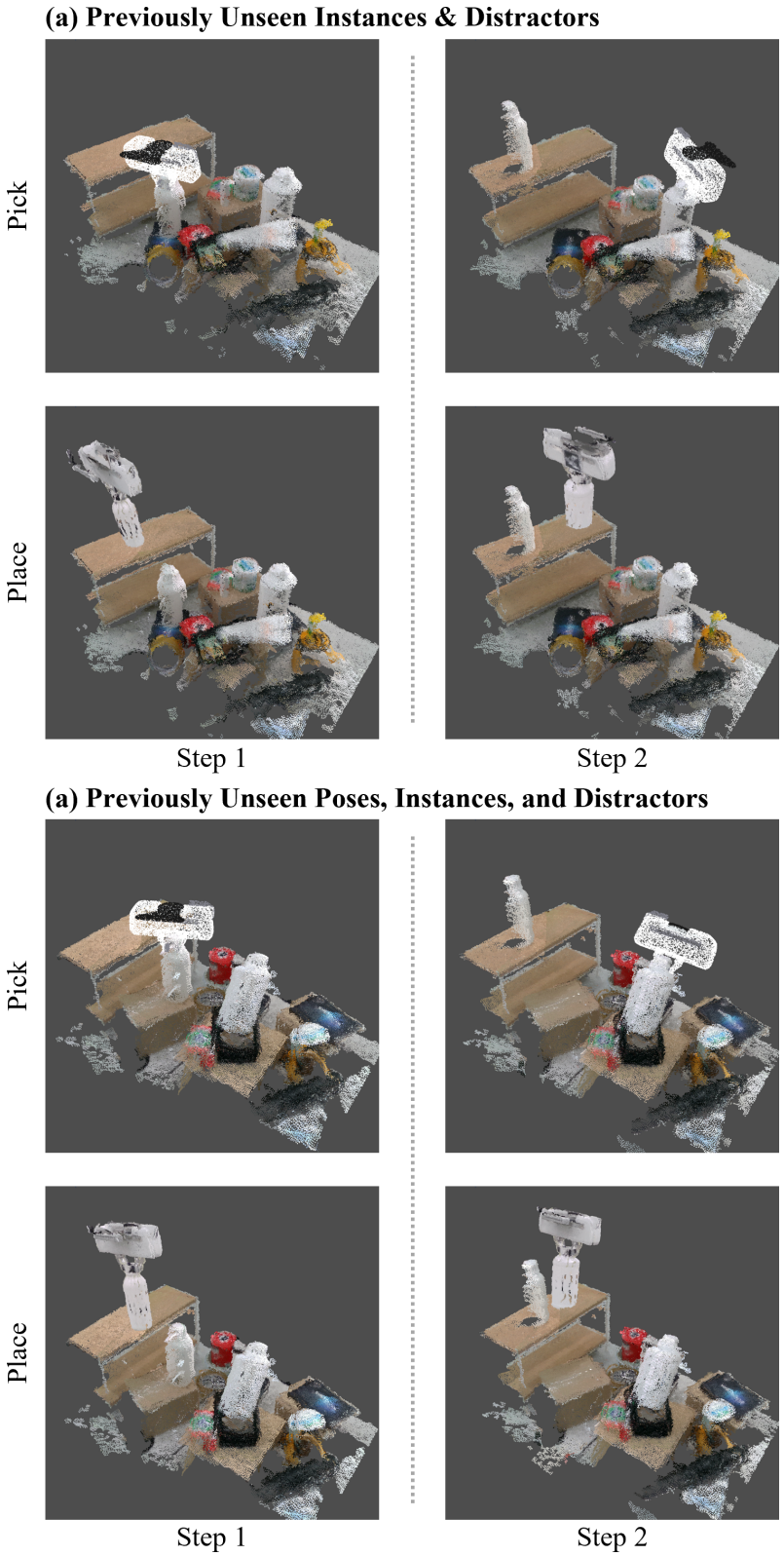

Figure 13. **Samples Generated by Diffusion-EDFs for Real-world Bottles-on-a-shelf Task (Previously Unseen Instances).** The figure depicts the end-effector pose samples for picking and placing multiples bottles on a shelf. In contrast to Fig. 12, we experiment with previously unseen bottle instances. Diffusion-EDFs trained with only four human demonstrations (three sequential subtasks for each) generated these samples from the real-world point cloud observations of the scene and grasp. Similar to Fig. 12, we experiment with both the **(a)** trained poses and **(b)** previously unseen poses (oblique; note that we only trained Diffusion-EDFs for upright poses). The denoising diffusion process can be found in the supplementary video at https://sites.google.com/view/diffusion-edfs.