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

Hyunwoo Ryu[1], Jiwoo Kim[1], Hyunseok An[1], Junwoo Chang[1], Joohwan Seo[2], Taehan Kim[3], Yubin Kim[4], Chaewon Hwang[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
{tomato1mule,nfsshift9801,junwoochang,hs991210,jongeunchoi}@yonsei.ac.kr
{joohwan_seo, horowitz}@berkeley.edu
taehan11.kim@samsung.com, ybkim95@media.mit.edu, hcw0221@ewhain.net

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, 4, 6, 7, 11, 35, 50, 52, 56, 69, 75]. In particular, $SE(3)$-*Diffusion Fields* [75] apply diffusion-based learning on the $SE(3)$ manifold to generate grasp poses of the end-effector. However, these methods require numerous demonstrations and do not generalize well on novel task configurations that are not provided during training.

In contrast, equivariant methods are well known for their data efficiency and generalizability in learning robotic manipulation tasks [6, 7, 15, 32, 33, 37, 49, 61, 67, 68, 78, 86]. In particular, several recent works explore the use of $SE(3)$-equivariant models for learning 6-DoF manipulation tasks with point cloud observations [15, 33, 61, 67, 68].

*Corresponding author: Jongeun Choi (jongeunchoi@yonsei.ac.kr)

Equivariant Descriptor Fields (EDFs) [61] achieve data-efficient end-to-end learning on 6-DoF visual robotic manipulation tasks by employing $SE(3)$ *bi-equivariant* [37, 61] energy-based models. However, EDFs require more than 10 hours to learn from only a few demonstrations due to the inefficient training of energy-based models.

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.

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

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

The vector space $\mathcal{W}$ where $\mathbf{D}(g)$ acts on is called the *representation space* of $\mathbf{D}(g)$. It is known that any representation of the special orthogonal group $SO(3)$ can be block-diagonalized into smaller representations by a change of basis. *Irreducible representations* are representations that cannot be reduced anymore, and hence constitute the building blocks of any larger representation.

According to the representation theory of $SO(3)$, all irreducible representations are classified according to their angular frequency $l \in \{0, 1, 2, \dots\}$, a non-negative integer number called *type*, or *spin*. Any type-l, or spin-l

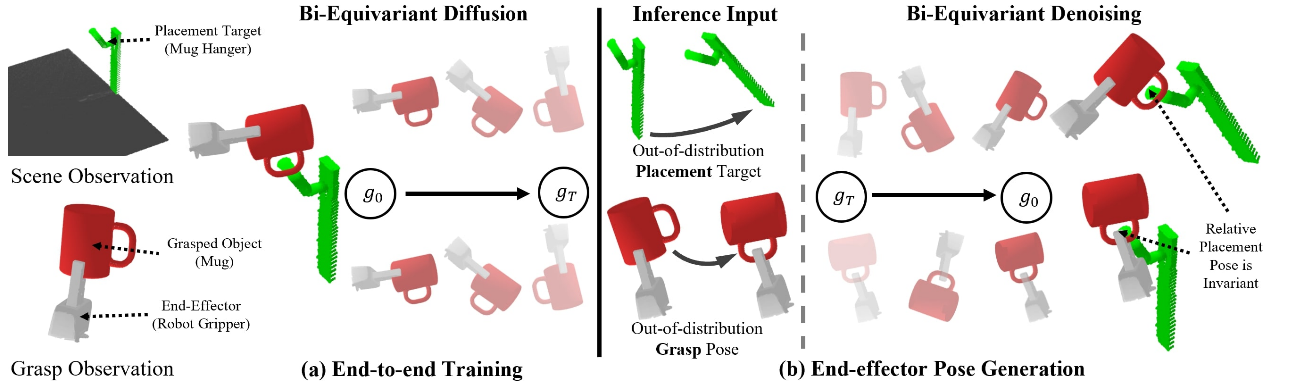


Figure 1. **Overview of Diffusion-EDFs. (a)** The target end-effector pose g_0 is bi-equivariantly diffused for the training of Diffusion-EDFs. **(b)** The end-effector pose is sampled from the policy by denoising with learned bi-equivariant score function. Due to the bi-equivariance, the trained policy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

representations are equivalent representations of the *real Wigner D-matrix* of degree l, denoted as $\mathbf{D}_l(R) : SO(3) \to \mathbb{R}^{(2l+1)\times(2l+1)}$. We refer to the vectors in the representation space of $\mathbf{D}_l(R)$ as *type-l*, or *spin-l vectors*. Type-0 representations have zero angular frequency, i.e. $\mathbf{D}_0(R) = 1$, meaning that type-0 vectors are *scalars* that are invariant under rotations. On the other hand, type-1 representations are identical when rotated by 360°, as their angular frequency is 1. Following the convention of E3NN [29], we use the x-y-z basis in which $\mathbf{D}_1(R) = R$. Therefore, type-1 vectors are typical spatial vectors in $\mathbb{R}^3$. In general, $\mathbf{D}_l(R)$ is identical when rotated by $\theta = 2\pi/l$, making higher-type vectors more suitable for encoding high-frequency details.

2.2. Equivariant Descriptor Fields

An Equivariant Descriptor Field (EDF) [61] $\boldsymbol{\varphi}(\boldsymbol{x}|O)$ 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

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

where $\mathbf{D}(R)$ is the block-diagonal matrix whose submatrices 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)

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

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 [17, 84] such that

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

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

where $\mathcal{N}$ is the normal distribution on $\mathbb{R}^3$, $\mathcal{IG}_{SO(3)}$ is the isotropic Gaussian on $SO(3)$ [34, 42, 55, 63], $g = (\boldsymbol{p}, R) \in SE(3)$, $\boldsymbol{p} \in \mathbb{R}^3$, $R \in SO(3)$, and θ is the rotation angle of $SO(3)$ in the axis-angle parameterization. CDF sampling is used for the sampling of $\mathcal{IG}_{SO(3)}$ [42].

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

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

Let $dP(g) = P(g)dg$ be a distribution on $SE(3)$ with the invariant probability distribution function $P(g)$. The *Langevin dynamics* for $dP(g)$ is defined as follows [8, 13]:

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

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

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 $dP(g)$ as

$\tau \to \infty$ regardless of the initial distribution. Thus, one may sample from $dP(g)$ 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[1] 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. [61], 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. A detailed explanation of this diffusion process will be deferred to a subsequent section. 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 [5, 17, 31, 34, 71, 84] 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* [13, 51, 53] 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[2] but not right invariant[2], 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[2] (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 use an equivariant *diffusion frame selection mechanism* $P(g_{ed}|g_0^{-1} \cdot O_s, O_e)$ where $g_{ed} \in SE(3)$ is the pose of the diffusion frame d with respect to the end-effector frame e

$$\begin{aligned} &P_{t|0}(g|g_0, O_s, O_e) \\ &= \int_{SE(3)} dg_{ed} P(g_{ed}|g_0^{-1} \cdot O_s, O_e) K_t(g_{ed}^{-1} g_0^{-1} g g_{ed}) \end{aligned} \tag{17}$$

where $K_t(g_0^{-1}g)$ is any left invariant kernel (see Supp. C.3). The diffusion procedure is as follows:

D1. A target pose $g_0 \sim P_0(g_0|O_s, O_e)$ is sampled.

D2. A diffusion frame $g_{ed} \sim P(g_{ed}|g_0^{-1} \cdot O_s, O_e)$ is sampled.

D3. A diffusion displacement $\Delta g_{t|0} \sim K_t(\Delta g_{t|0})$ is sampled.

D4. $\Delta g_{t|0}$ is applied to the demonstrated end-effector pose g_0 in the diffusion frame d, that is, $g_t = g_0\, g_{ed}\, \Delta g_{t|0}\, g_{ed}^{-1}$ where $g_t \sim P_t$ is the diffused end-effector pose.

Proposition 3. *The diffusion kernel $P_{t|0}$ in Eq. (17) is bi-equivariant if the diffusion frame selection mechanism $P(g_{ed}|g_0^{-1} \cdot O_s, O_e)$ satisfies the following property:*

$$P(g_{ed}|g_0^{-1} \cdot O_s, O_e) = P(\Delta g\, g_{ed}|(\Delta g\, g_0^{-1}) \cdot O_s, \Delta g \cdot O_e) \tag{18}$$

[1] For notational simplicity, we do not distinguish the probability distribution $dP = Pdg$ from the probability distribution function (PDF) P where dg denotes the bi-invariant volume form [13, 53, 85] on $SE(3)$.

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

See Supp. C.4 for the proof. In practice, however, the orientational part of the frame selection mechanism may be difficult to implement. Remarkably, for the specific case in which K_t is the Brownian diffusion kernel $\mathcal{B}_t$, only the translation part of the frame selection is required for Eq. (17) to be bi-equivariant. Therefore, we modify our diffusion frame selection mechanism as follows:

$$P(g_{ed}|g_0^{-1} \cdot O_s, O_e) = P(\boldsymbol{p}_{ed}|g_0^{-1} \cdot O_s, O_e)\, \delta(R_{ed}) \tag{19}$$

where $\delta(R)$ is the Dirac delta on $SO(3)$ and $P(\boldsymbol{p}_{ed}|g_0^{-1} \cdot O_s, O_e)$ is the *diffusion origin selection mechanism*.

Proposition 4. *The diffusion kernel $P_{t|0}$ in Eq. (17) with the frame selection mechanism in Eq. (19) is bi-equivariant if K_t in Eq. (17) is the Brownian diffusion kernel and the origin selection mechanism in Eq. (19) is equivariant that*

$$\begin{aligned} &P\left(\boldsymbol{p}_{ed}|g_0^{-1} \cdot O_s, O_e\right) \\ &= P\left(\Delta g\, \boldsymbol{p}_{ed}|(\Delta g\, g_0^{-1}) \cdot O_s, \Delta g \cdot O_e\right) \end{aligned} \tag{20}$$

We provide the proof in Supp. C.5. 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. 4.1.

3.4. Score Matching Objectives

In contrast to Song and Ermon [71], Urain et al. [75], 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):

$$\begin{aligned} \mathcal{J}_t &= \mathbb{E}_{g, g_0, g_{ed}, O_s, O_e}\left[J_t\right] \\ J_t &= \frac{1}{2}\left\|\boldsymbol{s}_t(g|O_s, O_e) - \nabla \log K_t(g_{ed}^{-1} g_0^{-1} g g_{ed})\right\|^2 \end{aligned} \tag{21}$$

where $g_0 \sim P_0(g_0|O_s, O_e)$, $g_{ed} \sim P(g_{ed}|g_0^{-1} \cdot O_s, O_e)$, 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

$$\underset{\boldsymbol{s}_t(g|O_s, O_e)}{\arg\min}\ \mathcal{J}_t = \boldsymbol{s}_t^*(g|O_s, O_e) = \nabla \log P_t(g|O_s, O_e) \tag{22}$$

Although Eq. (22) is a straightforward adaptation of the MSE minimizer formula [71], we still provide the derivation in Supp. C.6 for completeness. In practice, we use the Brownian diffusion kernel $\mathcal{B}_t$ for K_t to exploit Proposition 4. Therefore, training with Eq. (21) requires the computation of $\nabla \log \mathcal{B}_t(g_{ed}^{-1} g_0^{-1} g g_{ed})$. While autograd packages can be used for this computation [17, 34, 42, 61, 75, 84], we use 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

$$\boldsymbol{s}_t(g|O_s, O_e) = \left[\boldsymbol{s}_{\nu;t} \oplus \boldsymbol{s}_{\omega;t}\right](g|O_s, O_e) \tag{23}$$

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

$$\boldsymbol{s}_{\nu;t}(g|O_s, O_e) = \int_{\mathbb{R}^3} d^3\boldsymbol{x}\ \rho_{\nu;t}(\boldsymbol{x}|O_e)\ \widetilde{\boldsymbol{s}}_{\nu;t}(g, \boldsymbol{x}|O_s, O_e) \tag{24}$$

$$\begin{aligned} \boldsymbol{s}_{\omega;t}(g|O_s, O_e) = &\underbrace{\int_{\mathbb{R}^3} d^3\boldsymbol{x}\ \rho_{\omega;t}(\boldsymbol{x}|O_e)\ \widetilde{\boldsymbol{s}}_{\omega;t}(g, \boldsymbol{x}|O_s, O_e)}_{\text{Spin term}} \\ &+ \underbrace{\int_{\mathbb{R}^3} d^3\boldsymbol{x}\ \rho_{\nu;t}(\boldsymbol{x}|O_e)\ \boldsymbol{x} \wedge \widetilde{\boldsymbol{s}}_{\nu;t}(g, \boldsymbol{x}|O_s, O_e)}_{\text{Orbital term}} \end{aligned} \tag{25}$$

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 5. *The score model in Eq. (23) satisfies Eq. (11) and Eq. (12) if for $\square = \omega, \nu$ the density and score fields satisfy the following conditions for all $\Delta g \in SE(3)$*

$$\rho_{\square;t}(\Delta g\, \boldsymbol{x}|\Delta g \cdot O_e) = \rho_{\square;t}(\boldsymbol{x}|O_e) \tag{26}$$

$$\widetilde{\boldsymbol{s}}_{\square;t}(\Delta g\, g, \boldsymbol{x}|\Delta g \cdot O_s, O_e) = \widetilde{\boldsymbol{s}}_{\square;t}(g, \boldsymbol{x}|O_s, O_e) \tag{27}$$

$$\widetilde{\boldsymbol{s}}_{\square;t}(g\, \Delta g^{-1}, \Delta g\, \boldsymbol{x}|O_s, \Delta g \cdot O_e) = \Delta R\, \widetilde{\boldsymbol{s}}_{\square;t}(g, \boldsymbol{x}|O_s, O_e) \tag{28}$$

See Supp. C.7 for the proof. To achieve the left invariance (Eq. (27)) and right equivariance (Eq. (28)) of the score field, we propose using the following model with two EDFs:

$$\begin{aligned} &\widetilde{\boldsymbol{s}}_{\square;t}(g, \boldsymbol{x}|O_s, O_e) \\ &= \boldsymbol{\psi}_{\square;t}(\boldsymbol{x}|O_e)\ \otimes_{\square;t}^{(\to 1)}\ \mathbf{D}(R^{-1})\, \boldsymbol{\varphi}_{\square;t}(g\, \boldsymbol{x}|O_s) \end{aligned} \tag{29}$$

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 $\otimes_{\square;t}^{(\to 1)}$ is the time-conditioned equivariant tensor product [26, 74] with *Clebsch-Gordan coefficients* that maps the highly overparametrized equivariant descriptors into a type-1 vector.

Proposition 6. *The score field model in Eq. (29) satisfies Eq. (27) and Eq. (28).*

We provide the proof of Proposition 6 in Supp. C.8.

4. Implementation

In this section, we first provide the specific implementation of the bi-equivariant diffusion frame 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 provided in Supp. D

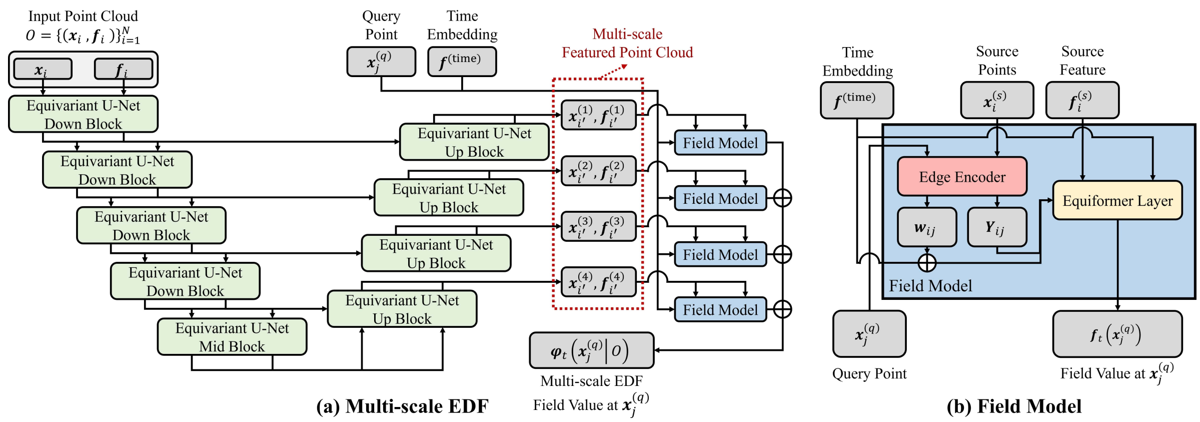


Figure 2. **Architecture of multiscale EDF.** Our multiscale EDF model is composed of a feature extracting part and a field model part. See Fig. 7 in Supp. D.3 for details on each module in the architecture. **(a)** The feature extractor encodes the input point cloud into multiscale featured point clouds. 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.

4.1. 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 [9, 15, 20, 37, 61]. 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.

$$\begin{aligned} &P\left(\boldsymbol{p}_{ed}|g_0^{-1}\cdot O_s, O_e\right) \\ &\propto \sum_{\boldsymbol{p}\in O_e} n_r\left(\boldsymbol{p},\ g_0^{-1}\cdot O_s\right)\delta^{(3)}(\boldsymbol{p}_{ed}-\boldsymbol{p}) \end{aligned} \tag{30}$$

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. See Supp. D.4 for more details.

4.2. Architecture of Equivariant Descriptor Fields

For faster sampling, we separate our implementation of EDFs into the feature extractor and the field model (see Fig. 2) as Ryu et al. [61] and Chatzipantazis et al. [9]. 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 [61] 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 [58]. For the field model, we find that a single layer is sufficient, although it is possible to stack multiple layers as Chatzipantazis et al. [9]. We use Equiformer [45] as the $SE(3)$-equivariant backbone GNN, with the addition of skip connections through point pooling layers. See Fig. 2 for an illustration of our architecture. More details can be found in Supp. D.3.

4.3. Score Model

We use the weighted query points model similar to Ryu et al. [61] for $\rho(\boldsymbol{x}|O)$

$$\rho(\boldsymbol{x}|O_e) = \sum_{\boldsymbol{q}\in Q(O_e)} w(\boldsymbol{x}|O_e)\delta^{(3)}(\boldsymbol{x}-\boldsymbol{q}) \tag{31}$$

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

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

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. (24) and Eq. (25) become tractable summation forms

$$\boldsymbol{s}_{\nu;t}(g|O_s,O_e) = \sum_{\boldsymbol{q}\in Q(O_e)} w(\boldsymbol{q}|O_e)\,\widetilde{\boldsymbol{s}}_{\nu;t}(g,\boldsymbol{q}|O_s,O_e) \tag{32}$$

$$\begin{aligned}\boldsymbol{s}_{\omega;t}(g|O_s,O_e) = &\sum_{\boldsymbol{q}\in Q(O_e)} w(\boldsymbol{q}|O_e)\,\widetilde{\boldsymbol{s}}_{\omega;t}(g,\boldsymbol{q}|O_s,O_e) \\ &+ \sum_{\boldsymbol{q}\in Q(O_e)} w(\boldsymbol{q}|O_e)\,\boldsymbol{q}\wedge\widetilde{\boldsymbol{s}}_{\nu;t}(g,\boldsymbol{q}|O_s,O_e)\end{aligned} \tag{33}$$

5. Experiments and Results

Simulation Benchmarks. We compare diffusion-EDFs with a state-of-the-art $SE(3)$-equivariant method (R-NDFs [68]) and a state-of-the-art denoising diffusion-based method ($SE(3)$-Diffusion Fields [75]) under an evaluation protocol similar to Simeonov et al. [67, 68], Ryu et al. [61], and Biza et al. [3]. In particular, we measure the pick-and-place success rate for two different object categories: mugs and bottles (see Fig. 3). 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. See Supp. E.1 for more details on the experimental setup.

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 [68]. 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. 1, Diffusion-EDFs consistently outperform both the $SE(3)$-equivariant baseline (R-NDFs [68]) and diffusion model baseline ($SE(3)$-DiffusionFields [75]) in almost all scenarios, despite not being provided with pre-training 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 [15, 37, 61]. 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 [37, 61] inherited from EDFs and our local contact-based diffusion frame selection mechanism.

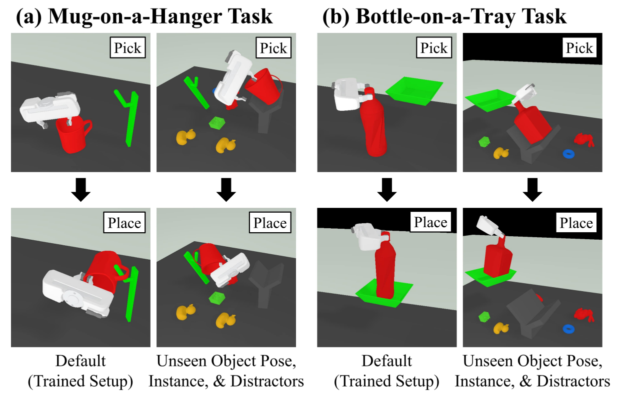


Figure 3. **Simulation Experiments.** (a) In the *Mug-on-a-Hanger* task, a red mug should be picked up by its rim and placed on a green hanger by its handle. (b) In the *Bottle-on-a-Tray* task, a red bottle should be picked up by its cap and placed on a green tray.

Real Hardware Experiments. 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. 5, and the experiment pipeline in Fig. 4. More details on the training and evaluation setups can be found in Supp. E.2.

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.

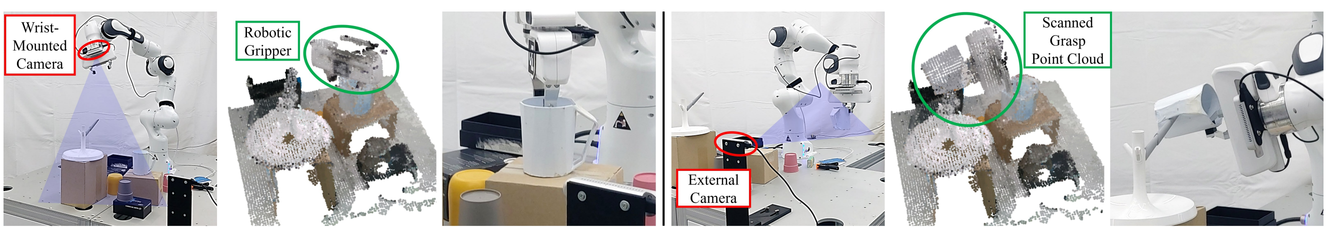
In the bowls-on-dishes task, the robot should pick up the bowls and place them on the dishes of matching colors in red-green-blue order. Note that this sequential task requires scene-level comprehension, which is impossible for methods that rely on object segmentation. For example, the robot should not 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 [69]. Methods that depend

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

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

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



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

Figure 4. **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.

on object segmentation are also unable 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.

We summarize the key challenges of each task in Tab. 2. For the experimental results, please refer to the supplementary materials and our project website: `https://sites.google.com/view/diffusion-edfs`

6. Related Works

Equivariant Robot Learning. Several works in robot learning utilize $SE(2)$-equivariance for efficient behavior cloning [32, 36, 47, 64, 73, 82, 86] and reinforcement

Figure 5. **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.

learning [76–78, 90]. Although these methods can be extended to problems that are not strictly $SE(2)$-symmetric [79, 80], they still suffer from highly spatial out-of-plane tasks [49, 61]. To address this issue, $SE(3)$-equivariance has been explored in robotic manipulation learning [6, 7, 15, 33, 37, 61, 67, 68]. Equivariant modeling has also been shown to be effective in robot control [37, 39, 66, 87].

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 2. Key challenges of each task

SE(3)-Equivariant Graph Neural Networks. $SO(3)$- and $SE(3)$-equivariant graph neural networks (GNNs) [19, 22, 26, 45, 46, 62, 74] are widely used to model the 3-dimensional roto-translation symmetry in various domains, including bioinformatics [17, 18, 27, 43, 84], chemistry [2, 26, 45, 74], computer vision [9, 20, 44, 48, 89], and robotics [25, 33, 61, 67].

Diffusion Models. Diffusion models are rapidly replacing previous generative models in various fields including computer vision [21, 30, 54, 59, 60, 70, 72], bioinformatics [17, 23, 81, 84], and robotics [1, 4, 6, 7, 10, 11, 24, 35, 50, 52, 56, 69, 75]. Recent works studied diffusion models on Riemannian manifolds [5, 31] such as Lie groups [17, 34, 42, 69, 75, 84]. In robotics, Simeonov et al. [69], Urain et al. [75] utilized diffusion models to generate end-effector poses from $SE(3)$. Several works also explore reward-guided diffusion policy [1, 35, 52, 75]. Equivariant diffusion models on the $SE(3)$ manifold have been partially explored in bioinformatics [17, 84] but not yet in robotics.

7. 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 [65, 66] or guided diffusion with motion planning cost [35, 75] 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, 8

[2] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):2453, 2022. 8

[3] 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. 6

[4] 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, 8

[5] 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. 3, 8

[6] 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, 8

[7] Johann Brehmer, Pim De Haan, Sönke Behrends, and Taco Cohen. Geometric algebra transformers. In *RSS 2023 Workshop on Symmetries in Robot Learning*, 2023. 1, 8

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

[9] 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. 5, 8

[10] Hongyi Chen, Yilun Du, Yiye Chen, Joshua B Tenenbaum, and Patricio A Vela. Planning with sequence models through iterative energy minimization. In *International Conference on Learning Representations*, 2023. 8

[11] 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, 8

[12] Gregory S Chirikjian. *Engineering applications of noncommutative harmonic analysis: with emphasis on rotation and motion groups*. CRC press, 2000. 15, 17

[13] Gregory S Chirikjian. *Stochastic models, information theory, and Lie groups, volume 2: Analytic methods and modern applications*. Springer Science & Business Media, 2011. 2, 3, 15

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

[15] 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, 5, 6, 8

[16] 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. 26

[17] 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, 4, 8, 14, 23

[18] Patrick Cramer. Alphafold2 and the future of structural biology. *Nature structural & molecular biology*, 28(9):704–705, 2021. 8

[19] 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. 8, 25

[20] 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. 5, 8

[21] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 8

[22] Weitao Du, He Zhang, Yuanqi Du, Qi Meng, Wei Chen, Nanning Zheng, Bin Shao, and Tie-Yan Liu. SE(3) equivariant graph neural networks with complete local frames. In *International Conference on Machine Learning*, pages 5583–5608. PMLR, 2022. 8

[23] Yilun Du, Conor Durkan, Robin Strudel, Joshua B Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Sussman Grathwohl. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. In *International Conference on Machine Learning*, pages 8489–8510. PMLR, 2023. 8

[24] Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in neural information processing systems*, 37, 2023. 8

[25] Jiahui Fu, Yilun Du, Kurran Singh, Joshua B Tenenbaum, and John J Leonard. Neuse: Neural se (3)-equivariant embedding for consistent spatial understanding with objects. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023. 8

[26] 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, 8, 23

[27] Octavian-Eugen Ganea, Xinyuan Huang, Charlotte Bunne, Yatao Bian, Regina Barzilay, Tommi S. Jaakkola, and Andreas Krause. Independent SE(3)-equivariant models for

end-to-end rigid protein docking. In *International Conference on Learning Representations*, 2022. 8

[28] 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. 26

[29] Mario Geiger and Tess Smidt. e3nn: Euclidean neural networks, 2022. 2

[30] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 8, 22

[31] 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. 3, 8

[32] 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, 7

[33] 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, 8

[34] Yesukhei Jagvaral, Francois Lanusse, and Rachel Mandelbaum. Diffusion generative models on SO(3). 2022. 2, 3, 4, 8, 14

[35] Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning*, 2022. 1, 8

[36] Mingxi Jia, Dian Wang, Guanang Su, David Klee, Xupeng Zhu, Robin Walters, and Robert Platt. Seil: Simulation-augmented equivariant imitation learning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1845–1851. IEEE, 2023. 7

[37] 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, 5, 6, 8, 13, 23

[38] 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. 22

[39] Colin Kohler, Anuj Shrivatsav Srikanth, Eshan Arora, and Robert Platt. Symmetric models for visual force policy learning. *arXiv preprint arXiv:2308.14670*, 2023. 8

[40] 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. 17

[41] 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. 26

[42] 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, 4, 8, 14

[43] Jae Hyeon Lee, Payman Yadollahpour, Andrew Watkins, Nathan C Frey, Andrew Leaver-Fay, Stephen Ra, Kyunghyun Cho, Vladimir Gligorijević, Aviv Regev, and Richard Bonneau. Equifold: Protein structure prediction with a novel coarse-grained structure representation. *bioRxiv*, pages 2022–10, 2022. 8

[44] Jiahui Lei, Congyue Deng, Karl Schmeckpeper, Leonidas Guibas, and Kostas Daniilidis. Efem: Equivariant neural field expectation maximization for 3d object segmentation without scene supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4902–4912, 2023. 8

[45] Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. In *The Eleventh International Conference on Learning Representations*, 2023. 5, 8, 23

[46] Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. *arXiv preprint arXiv:2306.12059*, 2023. 8

[47] Michael H Lim, Andy Zeng, Brian Ichter, Maryam Bandari, Erwin Coumans, Claire Tomlin, Stefan Schaal, and Aleksandra Faust. Multi-task learning with sequence-conditioned transporter networks. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2489–2496. IEEE, 2022. 7

[48] Chien Erh Lin, Jingwei Song, Ray Zhang, Minghan Zhu, and Maani Ghaffari. SE(3)-equivariant point cloud-based place recognition. In *Conference on Robot Learning*, pages 1520–1530. PMLR, 2023. 8

[49] 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, 8

[50] 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, 8

[51] Kevin M Lynch and Frank C Park. *Modern robotics*. Cambridge University Press, 2017. 3, 19

[52] 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, 8

[53] Richard M Murray, Zexiang Li, and S Shankar Sastry. *A mathematical introduction to robotic manipulation*. CRC press, 2017. 3, 15, 19

[54] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion

models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 8

[55] Dmitry I Nikolayev and Tatjana I Savyolov. Normal distribution on the rotation group SO(3). *Textures and Microstructures*, 29, 1970. 2, 15

[56] 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, 8

[57] 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. 26

[58] 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. 5

[59] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 8

[60] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 8

[61] 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, 4, 5, 6, 8, 13, 14, 15, 22, 23, 25, 26, 27

[62] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *International conference on machine learning*, pages 9323–9332. PMLR, 2021. 8

[63] 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

[64] Daniel Seita, Pete Florence, Jonathan Tompson, Erwin Coumans, Vikas Sindhwani, Ken Goldberg, and Andy Zeng. Learning to rearrange deformable cables, fabrics, and bags with goal-conditioned transporter networks. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4568–4575. IEEE, 2021. 7

[65] Joohwan Seo, Nikhil Potu Surya Prakash, Alexander Rose, and Roberto Horowitz. Geometric impedance control on SE(3) for robotic manipulators. *IFAC World Congress*, 2023. 8

[66] 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. 8

[67] 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, 6, 8, 25

[68] 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, 6, 7, 8, 25

[69] 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, 6, 8

[70] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 8

[71] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 3, 4, 22

[72] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 8

[73] Yadong Teng, Huimin Lu, Yujie Li, Tohru Kamiya, Yoshihisa Nakatoh, Seiichi Serikawa, and Pengxiang Gao. Multidimensional deformable object manipulation based on dn-transporter networks. *IEEE Transactions on Intelligent Transportation Systems*, 24(4):4532–4540, 2022. 7

[74] 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, 8, 23

[75] 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, 4, 6, 7, 8, 14, 22, 25, 26, 27

[76] Dian Wang, Mingxi Jia, Xupeng Zhu, Robin Walters, and Robert Platt. On-robot learning with equivariant models. In *6th Annual Conference on Robot Learning*, 2022. 8

[77] Dian Wang, Robin Walters, and Robert Platt. SO(2)-equivariant reinforcement learning. In *International Conference on Learning Representations*, 2022.

[78] 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, 8

[79] Dian Wang, Jung Yeon Park, Neel Sortur, Lawson L.S. Wong, Robin Walters, and Robert Platt. The surprising effectiveness of equivariant models in domains with latent symmetry. In *International Conference on Learning Representations*, 2023. 8

[80] Dian Wang, Xupeng Zhu, Jung Yeon Park, Mingxi Jia, Guanang Su, Robert Platt, and Robin Walters. A general theory

of correct, incorrect, and extrinsic equivariance. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 8

[81] Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *BioRxiv*, pages 2022–12, 2022. 8

[82] Hongtao Wu, Jikai Ye, Xin Meng, Chris Paxton, and Gregory S Chirikjian. Transporters with visual foresight for solving unseen rearrangement tasks. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10756–10763. IEEE, 2022. 7

[83] 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. 24

[84] 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, 4, 8, 14, 23

[85] Anthony Zee. *Group theory in a nutshell for physicists*. Princeton University Press, 2016. 3, 14

[86] 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, 7

[87] Linfeng Zhao, Jung Yeon Park, Xupeng Zhu, Robin Walters, and Lawson LS Wong. SE(3) frame equivariance in dynamics modeling and reinforcement learning. In *ICLR 2023 Workshop on Physics for Machine Learning*, 2023. 8

[88] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. 26

[89] Minghan Zhu, Maani Ghaffari, William A Clark, and Huei Peng. E2pn: Efficient SE(3)-equivariant point network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1223–1232, 2023. 8

[90] Xupeng Zhu, Dian Wang, Ondrej Biza, Guanang Su, Robin Walters, and Robert Platt. Sample efficient grasp learning using equivariant models. *Proceedings of Robotics: Science and Systems (RSS)*, 2022. 8