3D Equivariant Visuomotor Policy Learning via Spherical Projection

Boce Hu[1] **Dian Wang**[✉,2] **David Klee**[1] **Heng Tian**[1] **Xupeng Zhu**[1] **Haojie Huang**[1]
Robert Platt[†,1] **Robin Walters**[†,1]
[1]Northeastern University [2]Stanford University [†]Equal Advising

https://isp-3d.github.io/

Abstract

Equivariant models have recently been shown to improve the data efficiency of diffusion policy by a significant margin. However, prior work that explored this direction focused primarily on point cloud inputs generated by multiple cameras fixed in the workspace. This type of point cloud input is not compatible with the now-common setting where the primary input modality is an eye-in-hand RGB camera like a GoPro. This paper closes this gap by incorporating into the diffusion policy model a process that projects features from the 2D RGB camera image onto a sphere. This enables us to reason about symmetries in $\mathrm{SO}(3)$ without explicitly reconstructing a point cloud. We perform extensive experiments in both simulation and the real world that demonstrate that our method consistently outperforms strong baselines in terms of both performance and sample efficiency. Our work, Image-to-Sphere Policy (**ISP**), is the first $\mathrm{SO}(3)$-equivariant policy learning framework for robotic manipulation that works using only monocular RGB inputs.

1 Introduction

The eye-in-hand configuration, where the primary perception modality is a camera mounted near the wrist of the robot, is an important setting for robotic policy learning. This setup avoids the need for carefully calibrated external camera systems, is easier to integrate onto mobile robot platforms, and provides fine-grained visual details in the regions where the end-effector interacts with the environment. Moreover, it is used in a growing number of large robot datasets [27, 28, 42, 29, 2, 46].

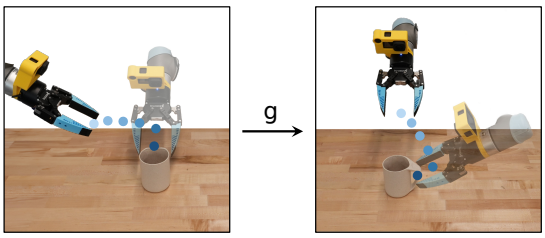


Figure 1: We propose the first $\mathrm{SO}(3)$-equivariant policy learning framework based on a single eye-in-hand RGB image, where the predicted action sequence transforms equivariantly under the same group action $g \in \mathrm{SO}(3)$ applied to the whole scene.

Despite recent advances in equivariant learning [66, 63], there remains a lack of effective network architectures for leveraging equivariant structure in this setting using only RGB input. Equivariant neural networks improve data efficiency and generalization by incorporating prior knowledge of domain symmetries directly into the model [34, 60, 41], and have recently been shown to enhance the performance of diffusion policy [1, 70]. However, existing equivariant diffusion policy frameworks perform best with point cloud data captured from multiple depth cameras [58]. When used with RGB data, current equivariant policies are unable to fully leverage the $\mathrm{SO}(3)$ structure present in the problem and underperform the point cloud version significantly [65]. This naturally raises the question: can $\mathrm{SO}(3)$-equivariance be achieved directly from monocular RGB images to support

✉Corresponding to dianwang@stanford.edu

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

https://doi.org/10.52202/085713-0201

data-efficient visuomotor learning? Such a capability should also have the potential to serve as a modular, plug-and-play component that generalizes seamlessly to richer sensing setups.

This paper addresses this challenge by introducing a novel diffusion policy framework that incorporates $\mathrm{SO}(3)$-equivariance into eye-in-hand visuomotor learning. Our method first projects features extracted from 2D RGB observations onto a sphere and then rotates the resulting spherical signal to compensate for camera motion. This yields a stable, $\mathrm{SO}(3)$-equivariant representation that is well-suited for downstream equivariant architectures. Unlike prior work that relies on segmented point clouds [58, 70] or calibrated multi-camera systems [65], our approach maintains equivariance throughout the entire policy and supports robust, sample-efficient closed-loop control directly from raw eye-in-hand inputs. To the best of our knowledge, this is the first framework to learn $\mathrm{SO}(3)$-equivariant visuomotor policies from monocular RGB observations in eye-in-hand settings.

Our key contributions are summarized as follows:

- We introduce **Image-to-Sphere Policy** (**ISP**), the first $\mathrm{SO}(3)$-equivariant policy learning framework that uses spherical projection from 2D RGB inputs to model 3D symmetries.
- We theoretically prove that our method achieves global $\mathrm{SO}(3)$-equivariance and local $\mathrm{SO}(2)$-invariance, facilitating policy learning.
- We validate our method through extensive experiments, achieving an average success rate improvement of 11.6% over twelve simulation tasks and 42.5% across four real-world tasks.

2 Related Work

Eye-in-hand Policy Learning Eye-in-hand policy learning [22, 26, 69, 67, 11, 43] has become a flexible and scalable alternative to traditional systems that rely on multiple fixed, externally mounted cameras with precise calibration [65, 25, 58, 71, 5]. By mounting cameras on the robot's wrist, these methods simplify deployment, avoid explicit calibration, and ease demonstration collection [6, 36, 52, 23]. However, the constantly shifting viewpoint introduces challenges like partial observability, which motivates the use of closed-loop policies that can handle local, viewpoint-dependent observations [72, 23, 4, 47, 73]. Recent work has explored transformer-based [59] or diffusion-based [15] architectures for eye-in-hand-control [74, 11], showing promising results across diverse manipulation tasks. Despite this progress, existing approaches often require large-scale demonstration data [35, 37], and often overlook symmetry structures inherently present in observations. Our method addresses this gap by introducing equivariant representations that encode geometric structure for eye-in-hand settings.

Closed-loop Visuomotor Policy Learning Early approaches to closed-loop visuomotor policy learning relied on reinforcement learning and CNN-based policies to map visual inputs to single-step actions [31, 72, 27]. While effective in simple tasks, these methods were sample-inefficient and struggled to capture multimodal behaviors, as each action was predicted independently without considering temporal context. To address this, subsequent work introduced temporal modeling into behavior cloning frameworks, such as BCRNN [39] and BeT [53], to improve sequential consistency and planning horizons. Building on this direction, recent advances have adopted generative policy models [5, 44, 71, 65], which model multi-step action sequences as a denoising process conditioned on observations. These approaches offer stronger expressiveness and improved multimodal behavior modeling. ISP extends this line of work by further integrating structural inductive biases, which enable more generalizable closed-loop control in complex manipulation settings.

Equivariance in Robotic Manipulation Equivariant and invariant representations have been shown to improve performance and sample efficiency [13, 54, 10, 56, 33, 18, 14, 32, 7]. Prior work has incorporated equivariant architectures for open-loop pick-and-place tasks [76, 64, 17, 19, 9, 45, 61, 21], showing strong performance with fewer demonstrations. Recently, equivariance has been extended to closed-loop diffusion policies [65, 70, 58]. EquiDiff [65] employs an $\mathrm{SO}(2)$-equivariant architecture to enhance Diffusion Policy [5]. EquiBot [70] adopts an $\mathrm{SIM}(3)$-equivariant structure, and ET-SEED [58] performs trajectory-level $\mathrm{SE}(3)$-equivariant diffusion, both leveraging segmented point cloud inputs to model spatial symmetries, thereby improving policy generalization. However, these approaches typically rely on multi-camera setups with fixed viewpoints or preprocessed 3D inputs. These constraints reduce their practicality in eye-in-hand settings, where the continuously shifting viewpoint and monocular RGB input violate the assumptions of existing equivariant models. To fill this gap, ISP models symmetry in the eye-in-hand RGB setting, preserves $\mathrm{SO}(3)$-equivariance, and can integrate with other frameworks to enhance their effectiveness without additional preprocessing.

3 Background

3.1 Representations of $\mathrm{SO}(3)$

A *group representation* encodes symmetry by mapping elements of a group to linear transformations. In this work, we focus on the special orthogonal group $\mathrm{SO}(3)$ of 3D rotations. A representation of $\mathrm{SO}(3)$ is a homomorphism $\rho : \mathrm{SO}(3) \to \mathrm{GL}(V)$, where V is a finite-dimensional vector space and $\mathrm{GL}(V)$ denotes the group of invertible linear transformations on V. We highlight three commonly used representations in robotics and geometric deep learning:

- **Degree-0 trivial representation** ρ_0: Maps every $g \in \mathrm{SO}(3)$ to the identity transformation on $\mathbb{R}$. This is used for rotation-invariant quantities, such as scalar sensor readings or gripper states.
- **Degree-1 standard representation** ρ_1: Maps $g \in \mathrm{SO}(3)$ to a 3×3 rotation matrix acting on $v \in \mathbb{R}^3$ via $\rho_1(g)v = gv$, capturing vector features like positions and directions.
- **Higher degree irreducible representations** ρ_ℓ: For $\ell \in \mathbb{N}$, the representation $\rho_\ell \colon \mathrm{SO}(3) \to \mathrm{GL}(\mathbb{R}^{2\ell+1})$ is given by the Wigner D-matrix of degree ℓ. It is used to describe higher degree features, such as relative poses, and is often used for latent features in equivariant neural networks.

3.2 Spherical Harmonics and Fourier Coefficients

Spherical harmonics $Y_\ell^m : \mathbb{S}^2 \to \mathbb{R}$ form an orthonormal basis for square-integrable functions on the 2-sphere and realize the irreducible representations of $\mathrm{SO}(3)$. Any spherical function $\Phi : \mathbb{S}^2 \to \mathbb{R}^d$ can thus be expanded as:

$$\Phi(\theta, \phi) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} c_\ell^m \, Y_\ell^m(\theta, \phi), \tag{1}$$

where c_ℓ^m are the corresponding Fourier coefficients. The mapping $\Phi \mapsto \{c_\ell^m\}$ is known as the Spherical Fourier Transform. Under a rotation $R \in \mathrm{SO}(3)$, each coefficient vector $c_\ell \in \mathbb{R}^{2\ell+1}$ transforms linearly via the representation ρ_ℓ:

$$c'_\ell = \rho_\ell(R) \cdot c_\ell. \tag{2}$$

This enables efficient rotation-equivariant operations on spherical signals in the spectral domain.

3.3 Diffusion Policy

Diffusion-based policy learning [24, 48] is a class of imitation learning methods that model distributions over action trajectories using denoising diffusion probabilistic models (DDPMs) [15]. These methods iteratively denoise sequences of noisy actions, conditioned on observations, to recover expert-like behavior. Formally, given an observation $\mathcal{O}$ and diffusion timestep k, the policy predicts a noise estimate ϵ^k from a corrupted action sequence $\mathbf{a}^k = \mathbf{a}^0 + \epsilon^k$ using a denoiser network Γ. The model is trained to minimize the denoising objective $\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{a}^0, k, \epsilon^k}\left[\|\Gamma(\mathcal{O}, \mathbf{a}^k, k) - \epsilon^k\|^2\right]$. At test time, the policy generates actions by iteratively denoising a randomly initialized sequence from Gaussian noise. Recent extensions [65] incorporate symmetry priors by designing the denoiser to be equivariant with respect to a transformation group G. Specifically, for compact groups such as $\mathrm{SO}(3)$, the denoiser Γ is required to satisfy the equivariance constraint:

$$\Gamma(g \cdot \mathcal{O},\, g \cdot \mathbf{a}^k,\, k) = g \cdot \Gamma(\mathcal{O}, \mathbf{a}^k, k) \quad \forall g \in G. \tag{3}$$

This formulation ensures that the denoising process respects the symmetry of the environment.

3.4 Problem formulation

We study closed-loop robotic visuomotor policy learning through behavior cloning, where a policy is trained to imitate expert demonstrations. Given an observation sequence $\mathcal{O} = \{o_{t-k+1}, ..., o_t\}$ at timestep t, the learned policy predicts an action chunk $A = \{\mathbf{a}_{t+1}, ..., \mathbf{a}_{t+n}\}$, where k and n are the observation length and prediction horizon, respectively. Each observation $o = (I, P)$ consists of an RGB image I from the wrist camera, proprioceptive input P describing the end-effector pose. Prior work has shown that higher performance is achieved when using absolute action representations, i.e., actions expressed in the world frame [5, 65]. Following this, we represent each predicted action

https://doi.org/10.52202/085713-0201

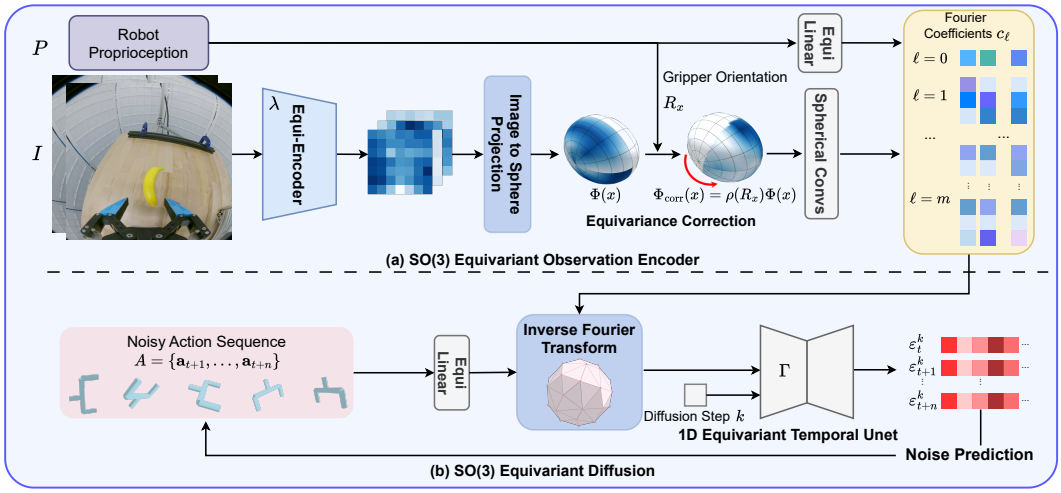


Figure 2: **Overview of Image-to-Sphere Policy (ISP)** (a) An $\mathrm{SO}(3)$-equivariant observation encoder extracts features from the RGB input, projects them onto the sphere, and applies an equivariance correction using the gripper orientation R_x to account for the camera's dynamic viewpoint (red arrow). The corrected spherical signal $\Phi_{\text{corr}}(x)$ is then processed by spherical convolution layers to extract $\mathrm{SO}(3)$ signals. Proprioceptive inputs are embedded via equivariant linear layers. Both image and proprioceptive features are represented as a set of Fourier coefficients c_ℓ on $\mathrm{SO}(3)$ and fused (yellow block). (b) The encoded spherical signals are transformed back to the spatial domain via inverse Fourier transform, sampling finite group elements as the conditioning vector for $\mathrm{SO}(3)$-equivariant denoising. The noisy action sequence is processed in the same way, through equivariant linear layers and projected onto the same group elements.

$a_t \in \mathbb{R}^{10}$ as the absolute end-effector pose including a position in $\mathbb{R}^3$, an orientation represented as a 6D rotation vector in $\mathbb{R}^6$ (see [75]), and a gripper open state in $\mathbb{R}^1$. As noted by [65], the absolute action parametrization has a symmetry: 3D transformations of the world frame result in the same 3D transformations to the action. We formalize the equivariance properties of ISP in Section 4.1.

4 Method

Figure 2 illustrates an overview of our proposed method, which consists of two key components, an $\mathrm{SO}(3)$-equivariant observation encoder (Figure 2a) and an $\mathrm{SO}(3)$-equivariant diffusion module (Figure 2b). The observation encoder uses spherical projection to map image-extracted features onto a hemisphere and applies spherical convolutions to ensure $\mathrm{SO}(3)$-equivariance, producing the conditioning vector for the diffusion process. The diffusion module is designed as an $\mathrm{SO}(3)$-equivariant function of the conditioning vectors and noisy inputs. As a result, the entire policy is end-to-end $\mathrm{SO}(3)$-equivariant. In the following subsections, we first describe our observation encoder, which extracts $\mathrm{SO}(3)$-equivariant features from 2D images, and then our equivariant diffusion module.

4.1 $\mathrm{SO}(3)$ Equivariant Observation Encoder

This section describes how we construct an $\mathrm{SO}(3)$-equivariant observation encoder that maps 2D images and robot proprioception into a 3D feature representation. The observation $x \in X$ consists of two parts, an eye-in-hand RGB image I, that captures visual information, and proprioceptive data, $P \in \mathbb{R}^7$, including the end-effector's 6D pose (position and orientation) and gripper state. Both these signals need to be represented in a way that encodes equivariance. Representing P is relatively easy. Following [65, 70, 51], we embed end-effector pose in $\mathrm{SO}(3)$ using the standard representation and gripper state using the trivial representation (Section 3.1). In contrast, encoding the 2D image input I into $\mathrm{SO}(3)$-equivariant features is harder because changes in the pose of the wrist-mounted camera induce out-of-plane viewpoint variations that are hard to model. We address this by projecting a standard 2D encoding of the image onto the sphere, as described below and first proposed in the context of object pose estimation [30, 16]. This enables us to reason about $\mathrm{SO}(3)$ action using its irreducible representations encoded as Wigner D-matrices (Section 3.1).

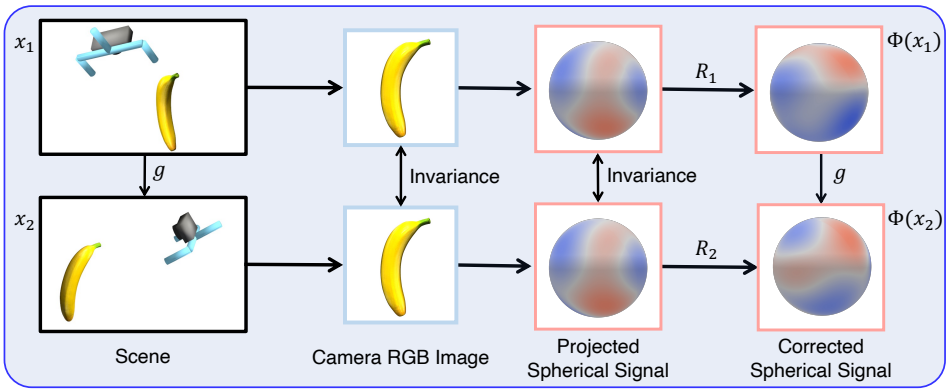


Figure 3: **Illustration of Equivariance Correction.** The left side shows two identical scenes under different global transformations. Since the wrist-mounted camera captures images in its local frame, the resulting images, and thus the projected spherical signals, remain identical across both scenes. By applying the gripper orientation R as an equivariance correction, we align these spherical signals to a common world frame, ensuring their equivariant transformation under global scene rotations.

Image Encoder Our image encoder is detailed in Figure 2a. First, we encode the input image I from the observation x using a standard $\mathrm{SO}(2)$-equivariant image encoder λ. Next, the resulting feature map $\lambda(I)$ is mapped onto the sphere using a learnable orthographic projection (see Appendix A for details). This converts a "flat" image into a spherical signal $\Phi(x) : \mathbb{S}^2 \to \mathbb{R}^d$ that is easier to manipulate in $\mathrm{SO}(3)$. We represent this spherical signal in the spectral domain as truncated Fourier coefficients calculated using the spherical Fourier transform (Equation 1).

Equivariance Correction At this point, the image encoding has been projected onto the sphere and represented using spherical harmonics. However, there is a problem. Since global 3D transformations of the world transform the camera and objects equally, the observed image and the projected signal $\Phi(x)$ would be invariant. This introduces a mismatch in that $\Phi(x)$ stays constant while the world and actions rotate, thereby breaking global $\mathrm{SO}(3)$-equivariance. We accommodate this by rotating the spherical signal by an amount corresponding to the $\mathrm{SO}(3)$ orientation of the gripper. We call this the *equivariance correction* factor, and it is illustrated in Figure 3. On the left of Figure 3, we see two scenes that are the same except for an $\mathrm{SO}(3)$ rotation of g. The eye-in-hand camera image (of the banana) is the same in both situations, even though the scene is rotated. This results in the identical projected spherical signal. It is only by applying the equivariance correction factor to the two respective signals (R_1 and R_2) that we recover the camera pose in the spherical signal. This ensures that the spherical signals produced in different camera poses are represented in a consistent global frame. We analyze this approach below.

Definition 1 (Equivariance Correction)**.** *Let G be a group acting on the input space X and output space Y. For a function $f : X \to Y$, an* Equivariance-Correction map *is any $\mathcal{C} : X \to G$ satisfying $\mathcal{C}(g{\cdot}x)\, f(g{\cdot}x) = g\,\mathcal{C}(x)\, f(x)$ for all $g \in G$, and $x \in X$. The corrected function $f_{corr}(x) = \mathcal{C}(x)\, f(x)$ is therefore G-equivariant.*

Notice that Definition 1 implies $f(x)$ and $f(gx)$ are in the same orbit. Equivariance Correction is similar to a *canonicalization map* $c : X \to G$, where $f_{\text{cano}} = c(x) f(c(x)^{-1}x)$ transforms the input to a canonical frame, then transforms the output back to the original frame. When $f(x) = f(gx)$, Equivariance Correction is a special case of canonicalization where $f(c(x)^{-1}x) = f(x)$ is invariant, so it only transforms the output to restore equivariance without altering the input.

We now show that Definition 1 is satisfied when the correction map is chosen to be the end-effector rotation. Let $x \in X$ denote the robot observation at a given timestep, which includes an eye-in-hand RGB image I and the corresponding camera (end-effector) pose $R_x \in \mathrm{SO}(3)$ in the world frame. Let $\Phi(x)$ denote the spherical signal derived from the image I, expressed in the camera frame, and let ρ be a representation of $\mathrm{SO}(3)$ acting on $\Phi(x)$.

Proposition 1 (Equivariance Correction via End-Effector Pose)**.** *The map $\mathcal{C} : (I, R_x) \mapsto R_x$, which assigns each camera image to its corresponding camera pose $R_x \in \mathrm{SO}(3)$ is an equivariance correction. The corrected signal $\Phi_{corr}(x) = \rho(\mathcal{C}(x))\Phi(x) = \rho(R_x)\Phi(x)$ is in a world-aligned frame. Thus, the mapping Φ_{corr} is $\mathrm{SO}(3)$-equivariant: for any global rotation $g \in \mathrm{SO}(3)$, we have $\Phi_{corr}(g{\cdot}x) = \rho(g)\, \Phi_{corr}(x)$.*

https://doi.org/10.52202/085713-0201

Proof. Let $g \in \mathrm{SO}(3)$ be a global rotation applied simultaneously to the scene and the camera. Since the image is recorded in the camera frame, the spherical signal is unaffected, i.e. $\Phi(g \cdot x) = \Phi(x)$, while the camera pose updates as $R_x \mapsto R_{gx} = gR_x$. For the corrected signal, we therefore obtain

$$\Phi_{\text{corr}}(gx) = \rho(R_{gx})\Phi(gx) = \rho(g)\rho(R_x)\Phi(x) = \rho(g)\Phi_{\text{corr}}(x), \tag{4}$$

where the second equality follows from the homomorphism property $\rho(gR_x) = \rho(g)\rho(R_x)$ of the representation ρ. Hence the map Φ_{corr} is $\mathrm{SO}(3)$-equivariant. □

A concrete realization of ρ with the spherical–harmonic coefficients and Wigner D-matrices is given in Appendix B, where the proposition reduces to the rotation of coefficient vectors in Eq. 2.

Camera-rotation invariance Our model also enforces an additional symmetry, rotations of the camera around its optical axis while the object remains stationary. These rotations form an $\mathrm{SO}(2)$ subgroup. Such rotations transform both the image and the camera pose, but their effects cancel out in the corrected world-frame signal. We now formalize the invariance of the corrected world-frame signal under such transformations.

Proposition 2 (Invariance to $\mathrm{SO}(2)$ Rotation of the Eye-in-hand Camera)**.** *Let $g \in \mathrm{SO}(2)$ be a rotation about the camera's optical axis. Then, under the transformation $(I, R_x) \mapsto (g \cdot I, R_x g^{-1})$, the corrected signal defined in Proposition 1 remains invariant: $\Phi_{corr}(g \cdot x) = \Phi_{corr}(x)$.*

Proof. Assume the image encoder λ is $\mathrm{SO}(2)$-equivariant, i.e., $\lambda(g \cdot I) = g \cdot \lambda(I)$ for all $g \in \mathrm{SO}(2)$. Because spherical projection and spherical Fourier transform preserve equivariance, the spherical signal satisfies $\Phi(g \cdot x) = g \cdot \Phi(x)$. Meanwhile, the camera pose transforms as $R_x \mapsto R_x g^{-1}$, since applying an $\mathrm{SO}(2)$ rotation g in the camera frame corresponds to right-multiplying its world-frame orientation R_x by g^{-1} (i.e., rotating the camera relative to itself). The corrected signal is:

$$\Phi_{\text{corr}}(g \cdot x) = \rho(R_x g^{-1})\, \Phi(g \cdot x) = \rho(R_x g^{-1})\, \rho(g)\, \Phi(x) = \rho(R_x)\, \Phi(x) = \Phi_{\text{corr}}(x). \tag{5}$$

Thus, the corrected signal is invariant under any $\mathrm{SO}(2)$ rotations of the eye-in-hand camera. □

By combining Propositions 1 and 2, we obtain a two-level symmetry in the encoder: the features are globally $\mathrm{SO}(3)$-equivariant and locally $\mathrm{SO}(2)$-invariant to rotations of the camera around its optical axis. These properties are inherently preserved without requiring additional constraints. As shown in Section 5, encoding these properties into the network leads to empirically improved performance.

4.2 $\mathrm{SO}(3)$ Equivariant Diffusion

As described in Section 3.3, we enforce end-to-end $\mathrm{SO}(3)$-equivariance by requiring the denoising network Γ to satisfy: $\Gamma(g \cdot \mathcal{O}, g \cdot \mathbf{a}^k, k) = g \cdot \Gamma(\mathcal{O}, \mathbf{a}^k, k)$ for all $g \in \mathrm{SO}(3)$. To achieve this, we extend the 2D denoising model from *EquiDiff* [65] to 3D. EquiDiff applies a shared 1D temporal U-Net [49] independently to each group element in $C_n \subset \mathrm{SO}(2)$. This element-wise weight sharing guarantees that the same parameters act on every group element, resulting in a noise embedding in the regular representation. To generalize to 3D, we approximate the continuous symmetry group $\mathrm{SO}(3)$ with a finite subgroup and perform sampling accordingly. Denote $H \subset \mathrm{SO}(3)$ the subgroup that the diffusion process is equivariant to (e.g., the icosahedral group I_{60}). Denote $S \subset \mathrm{SO}(3)$ a set that is closed under H, i.e., $HS = S$. Intuitively, S could be viewed as copies of the rotations in H, each with different offset angles to capture a denser discrete signal. Given a signal $\Psi : \mathrm{SO}(3) \to \mathbb{R}^d$, we first sample $\Psi(S) = \{\Psi(s_i) : s_i \in S\}$ and then evaluate the U-Net pointwise on each sample $\Gamma(\Psi(S)) = \{\Gamma(\Psi(s_i)) : s_i \in S\}$, where both the input and output can be treated as copies of the regular representations of H. Since $g \in H$ permutes the order of $\Psi(S)$ and $\Gamma(\Psi(S))$ identically, the entire process is H-equivariant. Because the spherical convolution layers output a signal on $\mathrm{SO}(3)$, we can flexibly choose any finite group H and sampling set S for discretization. In our implementation, we use both $C_8 \subset \mathrm{SO}(2)$ and $I_{60} \subset \mathrm{SO}(3)$ as choices of H. We refer readers to Appendix C for further details.

4.3 End-to-End Symmetry Analysis

In this section, we analyze the equivariant properties of our method. First, due to the $\mathrm{SO}(3)$-equivariant encoder (Proposition 1) and the $\mathrm{SO}(3)$-equivariant diffusion model (Section 4.2), our policy has end-to-end symmetry to global scene $\mathrm{SO}(3)$ rotations. This significantly improves its sample efficiency and generalizability to world coordinate frame changes.

Figure 5: A subset of experimental environments from MimicGen. Left: external view of the task. Right: eye-in-hand observation used in the experiments. The full set of tasks is shown in Appendix D.

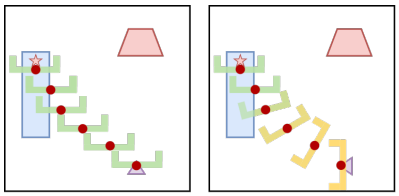


Figure 4: Illustration of translation invariance and rotation equivariance-to-invariance transition.

The benefit of $\mathrm{SO}(2)$ camera-rotation invariance (Proposition 2) is subtle. Under a rotation of the gripper with respect to the workspace, no a priori constraint can be placed on how the action trajectory should transform. However, our diffusion model receives a representation from the observation encoder that is equivariant to this rotation because it is constructed from invariant features (from the spherical signals) and equivariant features (from the end-effector rotation), thus providing a structured geometric bias. Figure 4 illustrates the benefit of this design. In both states (a) and (b), the gripper (triangle) aims to reach the same goal pose (star), but in (b) it is rotated by 90° around its optical axis. Translationally, the action in (b) should remain invariant (red dots), while rotationally, it should gradually transition from equivariant (yellow) to invariant (green) behavior. The equivariant component in the representation ensures that the model can correctly handle the initial 90° rotation through its symmetry, while the invariant component provides stability and goal alignment. Together, this representation offers a geometric inductive bias for learning such trajectories, whereas non-equivariant models must infer these patterns purely from data. The advantage is empirically validated in Section 5.

5 Experiments

5.1 Simulation

Experiment Setting We evaluate ISP on twelve robotic manipulation tasks from the MimicGen benchmark [40], which is widely used in previous work on closed-loop policy learning [8, 65]. A representative subset of these simulation tasks is shown in Figure 5 (see Appendix D for a full description of all twelve MimicGen tasks). Policies are trained and evaluated exclusively using eye-in-hand RGB observations (right image in each subfigure of Figure 5). To capture sufficient scene context, we enlarge the camera's field of view (FOV) to approximate a typical fisheye camera setup and re-generate the enlarged FOV observations using the original Mimicgen demonstrations for our method and baselines. For each task, we train three independent models with different random seeds (0, 1, and 2) for each of the 100- and 200-demonstration settings. The models are evaluated 60 times throughout training using 50 fixed rollouts per evaluation. We report the average of the best success rates from the three runs. Task and training details are provided in Appendix D and Appendix E.

Baselines Our experiments aim to validate the benefits of explicitly modeling equivariance in eye-in-hand visuomotor policies. We evaluate two versions of ISP with different symmetry levels, an **$\mathrm{SO}(3)$-equivariant** version and an **$\mathrm{SO}(2)$-equivariant** variant, which is symmetric only about rotations in the plane of the table. Although the $\mathrm{SO}(3)$ version has more symmetry, the $\mathrm{SO}(2)$ version is more lightweight, which may be preferable in some settings. We compare against three strong baselines: **(1) Diffusion Policy** [5]: A diffusion-based policy without any equivariance, serving as the primary reference. **(2) EquiDiff (modified)** [65]: Designed for fixed-camera settings, it achieves $\mathrm{SO}(2)$ equivariance via an equivariant image encoder and an equivariant temporal U-Net. For eye-in-hand control, we replace its image encoder with a standard ResNet [12], so only proprioception and denoising remain equivariant. **(3) ACT** [74]: A transformer-based behavior cloning method. To ensure a fair comparison, all experiments in the following sections, including ablations and method variants, consistently **apply** $\mathrm{SO}(2)$ data augmentation during training by rotating the end-effector pose in both proprioception and actions, equivalent to jointly rotating the gripper and scene.

https://doi.org/10.52202/085713-0201

Table 1: Success rates (%) on MimicGen tasks with 100 and 200 demonstrations, averaged over 3 seeds. We report both overall mean and per-task performance. The best result is highlighted in **bold**, and the second best is underlined. Full results with standard deviations are in Appendix F.

	Mean		Stack D1		Stack Three D1		Square D2		Threading D0		Three Pc. D0		Hammer Cl. D1	
Method	100	200	100	200	100	200	100	200	100	200	100	200	100	200
ISP-SO(3)	65.2 (+11.6)	75.0 (+10.5)	**99**	**100**	70	**88**	**35**	**51**	**90**	**92**	71	79	66	73
ISP-SO(2)	65.0 (+11.4)	73.1 (+8.6)	98	**100**	**75**	**88**	32	**51**	85	87	**75**	**80**	**71**	73
DiffPo	53.6	64.1	91	96	43	77	12	25	77	87	73	73	59	63
EquiDiff	53.0	64.5	96	99	61	80	9	19	89	**92**	74	79	59	**74**
ACT	23.0	40.9	45	77	12	37	3	10	36	53	28	50	35	63

	Mug Cl. D1		Coffee D2		Kitchen D1		Pick Place D0		Coffee Prep. D1		Nut Asse. D0	
Method	100	200	100	200	100	200	100	200	100	200	100	200
ISP-SO(3)	54	59	**64**	**69**	**75**	**79**	42	**66**	41	**61**	**75**	82
ISP-SO(2)	**56**	61	59	63	65	72	**46**	61	**47**	56	74	**84**
DiffPo	49	61	53	55	61	71	36	48	37	52	51	62
EquiDiff	51	**62**	47	61	55	67	28	46	27	39	40	56
ACT	25	37	21	35	21	51	9	14	8	16	37	49

Results Table 1 reports the maximum success rates across all methods and configurations. In terms of performance, ISP-SO(3) achieves the best results in 21 out of 24 task settings, consistently outperforming the baselines. The remaining three settings show only marginal differences (within 1-2%), all within the standard error margins. Similarly, ISP-SO(2) outperforms baselines in 20 settings, which further validates the effectiveness of our design. With only 100 demonstrations, our model exceeds the best-performing baseline by an average of 11.6%. With 200 demonstrations, the advantage remains similar at 10.5%. Importantly, our model trained with 100 demonstrations surpasses all baselines trained with 200 demonstrations and additional data augmentation, clearly demonstrating superior data efficiency. These results collectively highlight that the explicit modeling of equivariance is the key factor driving both the improved performance and enhanced sample efficiency of our method. Appendix F provides the full experimental results with standard deviations across three random seeds.

Ablation Study To assess the contribution of each component of our method, we conduct an ablation study on four representative tasks with 100 demonstrations: Stack Three D1, Square D2, Coffee D2, and Nut assembly D0. We evaluate the following variants of ISP-SO(3), each corresponding to a core module in our design: (1) **Sphere**: With or without the spherical projection and spherical convolutions for extracting SO(3)-equivariant features from images. (2) **EquiEnc**: With or without the proposed equivariant image encoder that captures SO(2)-invariant features (Proposition 2). (3) **EquiU**: With or without an equivariant temporal denoising U-Net in the diffusion module. The results are summarized in Table 2. Removing the spherical projection leads to the largest performance drop of 9.2%, highlighting its critical role in capturing symmetries, despite the use of data augmentation. Disabling the equivariant image encoder and the equivariant U-Net results in drops of 6.8% and 6.7%, respectively. These results demonstrate that all three components: spherical lifting, invariant encoding, and equivariant denoising, are essential for the overall effectiveness of our method. In addition, we further investigate the role of Equivariance Correction (Proposition 1) by comparing delta and absolute control strategies in Appendix G.

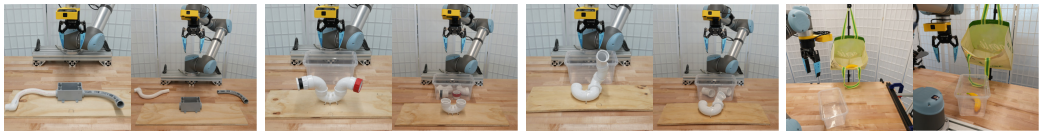
Table 2: Ablation study results. A red cross indicates that the corresponding module is removed in that variant.

Sphere	EquiEnc	EquiU	Sta.	Cof.	Nut.	Squ.	Mean
✗	✓	✓	63.3	61.3	59.0	23.3	51.8 (-9.2)
✓	✗	✓	66.0	57.3	61.3	32.0	54.2 (-6.8)
✓	✓	✗	68.7	58.7	58.0	32.0	54.3 (-6.7)
✓	✓	✓	**70.0**	**64.0**	**75.3**	**34.7**	**61.0**

The Benefits of Pretraining While our method already benefits from explicit equivariance, we further explore whether incorporating a pretrained image encoder can provide additional performance gains. Intuitively, pretraining can introduce stronger geometric priors and yield higher-quality visual features, especially beneficial in data-limited regimes. To evaluate this effect, we conduct experiments on the MimicGen using 100 demonstrations and the same evaluation protocol described above. We compare the ISP-SO(2) with two variants: *Pretraining*, which initializes the image encoder with an

Table 3: Success rates (%) on MimicGen tasks with 100 demonstrations, comparing pretrained and scratch initialization of the equivariant image encoder. Results are averaged over three seeds. Values in parentheses indicate the performance difference between the two settings.

Method	Mean	Stack D1	Stack Three D1	Square D2	Threading D0	Three Pc. D0	Hammer Cl. D1
ISP-SO(2) (Pretraining)	72.1(+7.1)	98.0 (=)	81.3 (+6.6)	56.0 (+24.0)	91.3 (+6.6)	76.7 (+2.0)	72.7 (+2.0)
ISP-SO(2) (Scratch)	65.0	98.0	74.7	32.0	84.7	74.7	70.7
		Mug Cl. D1	Coffee D2	Kitchen D1	Pick Place D0	Coffee Pre. D1	Nut Assembly D0
ISP-SO(2) (Pretraining)		54.0 (-2.0)	66.7 (+8.0)	64.0 (-0.7)	56.3 (+10.6)	63.3 (+16.6)	85.0 (+11.3)
ISP-SO(2) (Scratch)		56.0	58.7	64.7	45.7	46.7	73.7



(1) Box-Pipe Disassembly (2) U-Pipe Disassembly (3) 3D-Pipe Disassembly (4) Grocery Bag Retrieval

Figure 6: Real-world environments for evaluation. A GoPro camera is mounted on the robot's wrist to capture eye-in-hand observations. In each subfigure, the left image shows the initial state, while the right image shows the goal state. See Appendix H for detailed task descriptions.

ImageNet-1k [50]-pretrained equivariant ResNet-18, and *Scratch*, which trains the entire model from random initialization. Table 3 reports the maximum evaluation success rates.

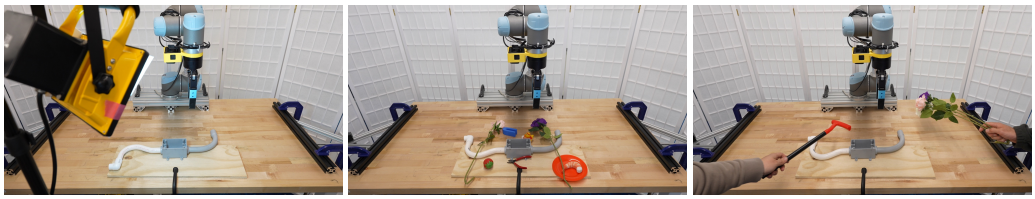
Results show that ISP-SO(2) (Pretraining) surpasses ISP-SO(2) (Scratch) by 7.1%, indicating consistently improved final performance across most tasks. Moreover, the pretrained version with only 100 demonstrations achieves comparable performance to training from scratch with 200 demonstrations, further highlighting its data efficiency. These findings demonstrate the effectiveness of pretraining in providing richer and more stable visuomotor representations. Although the performance gains are marginal or absent in a few tasks, this suggests that naive pretraining may not always align perfectly with the downstream visuomotor learning objective. Developing pretraining strategies that are tailored to equivariant visuomotor policy representations is, therefore, a promising future direction.

5.2 Real World

Physical Setups Our real robot experiments use a Universal Robot UR5 equipped with a Robotiq-85 Gripper and custom-designed soft fingers. A GoPro camera is mounted on the wrist, following prior setups [6, 11, 35]. Demonstrations are collected via the Gello teleoperation interface [68], with observations and actions recorded at 5 Hz. Following [5, 65], we employ DDIM [55] to reduce the number of denoising steps to 16. Figure 6 illustrates the four real-world manipulation tasks. The first three tasks involve pipe disassembly, each focusing on different challenges in closed-loop control: background-object segmentation (Box-Pipe), long-horizon control (U-Pipe), and handling complex 3D geometries (3D-Pipe). The fourth task involves retrieving objects from a deformable grocery bag, for which wrist-mounted camera observations are the only reliable source of visual information due to severe occlusions and limited external visibility. We compare ISP-SO(3) against the Diffusion Policy [5]. Further details on the physical setup, task visualization, goal specification, and practical guidelines for data collection are provided in Appendix H and Appendix I.

Results Table 4 reports success rates over 20 trials per task. Our method consistently outperforms the Diffusion Policy [5] baseline, with significant improvements on Box-Pipe (80% vs. 10%) and 3D-Pipe (75% vs. 15%). The former benefits from more precise visual representations that distinguish the gray pipe from the gray box background, while the latter showcases the advantage of SO(3)-equivariant features for reasoning over complex 3D geometries. The U-Pipe task also shows a notable gain (85% vs. 65%), demonstrating the sustained and stable performance of our equivariant method in the long-horizon task. On the Grocery Bag task, which heavily relies on eye-in-hand perception, our method achieves a 95% success rate. This shows its high stability and robustness. These results confirm the effectiveness of our equivariant design in addressing diverse manipulation challenges in the real world. See Appendix J for a detailed failure analysis. We further evaluate the computational efficiency of ISP in real-world settings, with a comprehensive discussion provided in Appendix K. In addition, we discuss potential limitations and practical considerations of equivariance in Appendix L.

https://doi.org/10.52202/085713-0201



(a) Lighting Change (b) Background Clutter (c) Partial Camera Occlusion

Figure 7: Real-world perturbation scenarios used to evaluate the robustness and generalization of our method on the Box-Pipe Disassembly task.

Robustness to Real-World Perturbations To further evaluate the robustness and generalization ability of our policy, we conducted additional real-world experiments on the Box-Pipe Disassembly task under various domain shifts. First, we altered the lighting conditions by introducing a strong white point light source near the workspace, which substantially changed the shadows and color temperature of the scene (Figure 7a). Second, we perturbed the background by placing multiple household objects on the table to create clutter (Figure 7b). Finally, to test robustness against partial occlusion, we repeatedly and briefly blocked the eye-in-hand camera by rapidly waving different objects (e.g., a toy golf club and a flower) in front of it during policy rollout (Figure 7c). Using the same initial states and 20 rollouts per condition as in the previous real-world experiments, ISP-$\mathrm{SO}(2)$ achieves success rates of 85% under lighting changes, 75% with background clutter, and 85% under partial camera occlusion. For reference, the performance of ISP-$\mathrm{SO}(3)$ without perturbations is 80%. These results demonstrate that the proposed method generalizes well to real-world disturbances and maintains strong task performance under challenging visual conditions.

Table 4: Real-world task performance over 20 trials. The number of demonstrations used for training each task is shown in the second row.

	Box-Pipe	U-Pipe	3D-Pipe	Grocery Bag
# Demos	65	65	65	60
ISP-$\mathrm{SO}(3)$	80%(16/20)	85%(17/20)	75%(15/20)	95%(19/20)
DiffPo [5]	10%(2/20)	65%(13/20)	15%(3/20)	75%(15/20)

6 Conclusion

In this paper, we propose **Image-to-Sphere Policy (ISP)**, the first $\mathrm{SO}(3)$-equivariant policy learning framework for eye-in-hand visuomotor control using only monocular RGB inputs. By lifting 2D image features onto the sphere and introducing an equivariance correction mechanism to compensate for dynamic camera viewpoints, our method achieves global $\mathrm{SO}(3)$-equivariance and local $\mathrm{SO}(2)$-invariance without relying on depth sensors or multi-camera setups. This design enables robust and sample-efficient policy learning in dynamic, real-world settings. Extensive experiments in both simulation and real-world tasks demonstrate that ISP consistently outperforms strong baselines, achieving higher success rates with fewer demonstrations. Our work provides a general and effective algorithmic solution that is both deployable and scalable for eye-in-hand visuomotor learning.

Limitations Our method has several limitations for future investigation. First, we only consider a single wrist-mounted RGB camera. While this view provides fine-grained local information, it lacks the global scene context that an agent-view camera could offer. Effectively combining these complementary perspectives remains an important challenge. Second, our approach models rotational equivariance but does not address translational equivariance. This limits the model's ability to generalize to object translations within the scene. Extending the equivariance correction to handle camera translations is a promising direction for future work. Third, the use of equivariant networks increases training time. Although inference remains efficient, reducing training overhead through more lightweight architectures would further enhance practicality. Fourth, our current method focuses on single-arm manipulation. Extending the framework to bimanual systems, where coordination between two arms is required, is a natural next step. Finally, our method does not yet leverage vision-language models. Integrating high-level semantic understanding through vision language models could further improve generalization and task understanding in more diverse environments.

https://doi.org/10.52202/085713-0201

Acknowledgments

We would like to thank Rachel Lim and Andrew Cole for their assistance with the real-world experiments as well as all members of the Helping Hands Lab for their valuable discussions and feedback on the manuscript. This work was supported in part by NSF grants 2107256, 2134178, 2314182, 2409351, 2442658, and NASA grant 80NSSC19K1474.

References

[1] Johann Brehmer, Joey Bose, Pim De Haan, and Taco Cohen. EDGI: Equivariant Diffusion for Planning with Embodied Agents. *arXiv preprint arXiv:2303.12410*, 2023.

[2] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[3] Gabriele Cesa, Leon Lang, and Maurice Weiler. A program to build E(N)-equivariant steerable CNNs. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=WE4qe9xlnQw.

[4] Ricson Cheng, Arpit Agarwal, and Katerina Fragkiadaki. Reinforcement learning of active vision for manipulating objects under occlusions. In *Conference on robot learning*, pages 422–431. PMLR, 2018.

[5] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.

[6] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*, 2024.

[7] Zihao Dong, Alan Papalia, Leonard Jung, Alenna Spiro, Philip R Osteen, Christa S Robison, and Michael Everett. Learning smooth state-dependent traversability from dense point clouds. *arXiv preprint arXiv:2506.04362*, 2025.

[8] Niklas Funk, Julen Urain, Joao Carvalho, Vignesh Prasad, Georgia Chalvatzaki, and Jan Peters. Actionflow: Equivariant, accurate, and efficient policies with spatially symmetric flow matching. *arXiv preprint arXiv:2409.04576*, 2024.

[9] Chongkai Gao, Zhengrong Xue, Shuying Deng, Tianhai Liang, Siqi Yang, Lin Shao, and Huazhe Xu. Riemann: Near real-time se (3)-equivariant robot manipulation without point cloud segmentation. *arXiv preprint arXiv:2403.19460*, 2024.

[10] Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, and Jianzhu Ma. 3D Equivariant Diffusion for Target-Aware Molecule Generation and Affinity Prediction. In *The Eleventh International Conference on Learning Representations*, 2023.

[11] Huy Ha, Yihuai Gao, Zipeng Fu, Jie Tan, and Shuran Song. Umi on legs: Making manipulation policies mobile with manipulation-centric whole-body controllers. *arXiv preprint arXiv:2407.10353*, 2024.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[13] Lingshen He, Yuxuan Chen, Zhengyang Shen, Yiming Dong, Yisen Wang, and Zhouchen Lin. Efficient equivariant network. In *Advances in Neural Information Processing Systems*, 2021.

[14] Lingshen He, Yuxuan Chen, Zhengyang Shen, Yibo Yang, and Zhouchen Lin. Neural epdos: Spatially adaptive equivariant partial differential operator based networks. In *The international conference on learning representations*, 2022.

 https://doi.org/10.52202/085713-0201

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

[16] Owen Howell, David Klee, Ondrej Biza, Linfeng Zhao, and Robin Walters. Equivariant single view pose prediction via induced and restriction representations. *Advances in Neural Information Processing Systems*, 36:47251–47263, 2023.

[17] Boce Hu, Xupeng Zhu, Dian Wang, Zihao Dong, Haojie Huang, Chenghao Wang, Robin Walters, and Robert Platt. Orbitgrasp: $SE(3)$-equivariant grasp learning. *arXiv preprint arXiv:2407.03531*, 2024.

[18] Boce Hu, Heng Tian, Dian Wang, Haojie Huang, Xupeng Zhu, Robin Walters, and Robert Platt. Push-grasp policy learning using equivariant models and grasp score optimization. *IEEE Robotics and Automation Letters*, 10(11):11180–11187, 2025. doi: 10.1109/LRA.2025.3606392.

[19] Haojie Huang, Dian Wang, Robin Walters, and Robert Platt. Equivariant transporter network. *arXiv preprint arXiv:2202.09400*, 2022.

[20] Haojie Huang, Owen Howell, Dian Wang, Xupeng Zhu, Robin Walters, and Robert Platt. Fourier transporter: Bi-equivariant robotic manipulation in 3d. *arXiv preprint arXiv:2401.12046*, 2024.

[21] Haojie Huang, Dian Wang, Arsh Tangri, Robin Walters, and Robert Platt. Leveraging symmetries in pick and place. *The International Journal of Robotics Research*, 43(4):550–571, 2024.

[22] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

[23] Rishabh Jangir, Nicklas Hansen, Sambaran Ghosal, Mohit Jain, and Xiaolong Wang. Look closer: Bridging egocentric and third-person views with transformers for robotic manipulation. *IEEE Robotics and Automation Letters*, 7(2):3046–3053, 2022.

[24] Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with Diffusion for Flexible Behavior Synthesis. In *International Conference on Machine Learning*, pages 9902–9915. PMLR, 2022.

[25] Mingxi Jia, Dian Wang, Guanang Su, David Klee, Xupeng Zhu, Robin Walters, and Robert Platt. Seil: Simulation-augmented equivariant imitation learning. *arXiv preprint arXiv:2211.00194*, 2022.

[26] Yunfan Jiang, Ruohan Zhang, Josiah Wong, Chen Wang, Yanjie Ze, Hang Yin, Cem Gokmen, Shuran Song, Jiajun Wu, and Li Fei-Fei. Behavior robot suite: Streamlining real-world whole-body manipulation for everyday household activities. *arXiv preprint arXiv:2503.05652*, 2025.

[27] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on robot learning*, pages 651–673. PMLR, 2018.

[28] Dmitry Kalashnikov, Jacob Varley, Yevgen Chebotar, Benjamin Swanson, Rico Jonschkowski, Chelsea Finn, Sergey Levine, and Karol Hausman. Mt-opt: Continuous multi-task robotic reinforcement learning at scale. *arXiv preprint arXiv:2104.08212*, 2021.

[29] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

[30] David M Klee, Ondrej Biza, Robert Platt, and Robin Walters. Image to sphere: Learning equivariant features for efficient pose prediction. *arXiv preprint arXiv:2302.13926*, 2023.

[31] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.

[32] Yikang Li, Yeqing Qiu, Yuxuan Chen, Lingshen He, and Zhouchen Lin. Affine equivariant networks based on differential invariants. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5546–5556, 2024.

[33] Yikang Li, Yeqing Qiu, Yuxuan Chen, and Zhouchen Lin. Affine steerable equivariant layer for canonicalization of neural networks. In *The international conference on learning representations*, 2025.

[34] Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KwmPfARgOTD.

[35] Fanqi Lin, Yingdong Hu, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. Data scaling laws in imitation learning for robotic manipulation. *arXiv preprint arXiv:2410.18647*, 2024.

[36] Kehui Liu, Chuyue Guan, Zhongjie Jia, Ziniu Wu, Xin Liu, Tianyu Wang, Shuai Liang, Pengan Chen, Pingrui Zhang, Haoming Song, et al. Fastumi: A scalable and hardware-independent universal manipulation interface with dataset. *arXiv e-prints*, pages arXiv–2409, 2024.

[37] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024.

[38] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[39] Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. *arXiv preprint arXiv:2108.03298*, 2021.

[40] Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations. *arXiv preprint arXiv:2310.17596*, 2023.

[41] Arnab Kumar Mondal, Pratheeksha Nair, and Kaleem Siddiqi. Group equivariant deep reinforcement learning. *arXiv preprint arXiv:2007.03437*, 2020.

[42] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.

[43] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025.

[44] Aaditya Prasad, Kevin Lin, Jimmy Wu, Linqi Zhou, and Jeannette Bohg. Consistency policy: Accelerated visuomotor policies via consistency distillation. *arXiv preprint arXiv:2405.07503*, 2024.

[45] Yu Qi, Yuanchen Ju, Tianming Wei, Chi Chu, Lawson LS Wong, and Huazhe Xu. Two by two: Learning multi-task pairwise objects assembly for generalizable robot manipulation. *CVPR 2025*, 2025.

[46] Yu Qi, Haibo Zhao, Ziyu Guo, Siyuan Ma, Ziyan Chen, Yaokun Han, Renrui Zhang, Zitiantao Lin, Shiji Xin, Yijian Huang, et al. Bear: Benchmarking and enhancing multimodal language models for atomic embodied capabilities. *arXiv preprint arXiv:2510.08759*, 2025.

 https://doi.org/10.52202/085713-0201

[47] Allen Z Ren, Justin Lidard, Lars L Ankile, Anthony Simeonov, Pulkit Agrawal, Anirudha Majumdar, Benjamin Burchfiel, Hongkai Dai, and Max Simchowitz. Diffusion policy policy optimization. *arXiv preprint arXiv:2409.00588*, 2024.

[48] Moritz Reuss, Maximilian Li, Xiaogang Jia, and Rudolf Lioutikov. Goal conditioned imitation learning using score-based diffusion policies. In *Robotics: Science and Systems*, 2023.

[49] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[50] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[51] Hyunwoo Ryu, Jiwoo Kim, Hyunseok An, Junwoo Chang, Joohwan Seo, Taehan Kim, Yubin Kim, Chaewon Hwang, Jongeun Choi, and Roberto Horowitz. Diffusion-edfs: Bi-equivariant denoising generative modeling on se (3) for visual robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18007–18018, 2024.

[52] Mingyo Seo, H. Andy Park, Shenli Yuan, Yuke Zhu, , and Luis Sentis. Legato: Cross-embodiment imitation using a grasping tool. *IEEE Robotics and Automation Letters (RA-L)*, 2025.

[53] Nur Muhammad Shafiullah, Zichen Cui, Ariuntuya Arty Altanzaya, and Lerrel Pinto. Behavior transformers: Cloning k modes with one stone. *Advances in neural information processing systems*, 35:22955–22968, 2022.

[54] 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.

[55] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[56] Sangli Teng, William Clark, Anthony Bloch, Ram Vasudevan, and Maani Ghaffari. Lie algebraic cost function design for control on lie groups. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 1867–1874. IEEE, 2022.

[57] 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.

[58] Chenrui Tie, Yue Chen, Ruihai Wu, Boxuan Dong, Zeyi Li, Chongkai Gao, and Hao Dong. Et-seed: Efficient trajectory-level se (3) equivariant diffusion policy. *arXiv preprint arXiv:2411.03990*, 2024.

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[60] Robin Walters, Jinxi Li, and Rose Yu. Trajectory prediction using equivariant continuous convolution. *arXiv preprint arXiv:2010.11344*, 2020.

[61] Dian Wang, Mingxi Jia, Xupeng Zhu, Robin Walters, and Robert Platt. On-robot learning with equivariant models. *arXiv preprint arXiv:2203.04923*, 2022.

[62] Dian Wang, Jung Yeon Park, Neel Sortur, Lawson LS Wong, Robin Walters, and Robert Platt. The surprising effectiveness of equivariant models in domains with latent symmetry. *arXiv preprint arXiv:2211.09231*, 2022.

[63] Dian Wang, Robin Walters, and Robert Platt. $so(2)$-equivariant reinforcement learning. *arXiv preprint arXiv:2203.04439*, 2022.

[64] 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.

[65] Dian Wang, Stephen Hart, David Surovik, Tarik Kelestemur, Haojie Huang, Haibo Zhao, Mark Yeatman, Jiuguang Wang, Robin Walters, and Robert Platt. Equivariant diffusion policy. *arXiv preprint arXiv:2407.01812*, 2024.

[66] Maurice Weiler and Gabriele Cesa. General e (2)-equivariant steerable cnns. *Advances in neural information processing systems*, 32, 2019.

[67] Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots*, 47(8):1087–1102, 2023.

[68] Philipp Wu, Yide Shentu, Zhongke Yi, Xingyu Lin, and Pieter Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12156–12163. IEEE, 2024.

[69] Xiaomeng Xu, Dominik Bauer, and Shuran Song. Robopanoptes: The all-seeing robot with whole-body dexterity. *arXiv preprint arXiv:2501.05420*, 2025.

[70] Jingyun Yang, Zi-ang Cao, Congyue Deng, Rika Antonova, Shuran Song, and Jeannette Bohg. Equibot: Sim (3)-equivariant diffusion policy for generalizable and data efficient learning. *arXiv preprint arXiv:2407.01479*, 2024.

[71] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. *arXiv preprint arXiv:2403.03954*, 2024.

[72] Tianhao Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Xi Chen, Ken Goldberg, and Pieter Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 5628–5635. Ieee, 2018.

[73] Haibo Zhao, Dian Wang, Yizhe Zhu, Xupeng Zhu, Owen Howell, Linfeng Zhao, Yaoyao Qian, Robin Walters, and Robert Platt. Hierarchical equivariant policy via frame transfer. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=nAv5ketrHq.

[74] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

[75] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the Continuity of Rotation Representations in Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019.

[76] Xupeng Zhu, Dian Wang, Ondrej Biza, Guanang Su, Robin Walters, and Robert Platt. Sample efficient grasp learning using equivariant models. *arXiv preprint arXiv:2202.09468*, 2022.

 https://doi.org/10.52202/085713-0201

A Orthographic Projection Details

The orthographic projection in our method follows the approach of [30], which lifts 2D feature maps onto the unit sphere in the camera frame through a signal remapping operation. Unlike traditional geometric projections that rely on explicit camera calibration or depth information, our projection is entirely learned and does not depend on a predefined 3D center or physical camera parameters.

In practice, the spherical signal is sampled on a HEALPix grid, which provides an equal-area, hierarchical discretization of the sphere. For each point on the sphere, we apply a learnable weighted aggregation over the entire 2D feature map to compute its corresponding signal value. This design allows the network to flexibly determine how spatial features are mapped onto the sphere, rather than relying on fixed projection kernels. This is particularly useful for wide FOV images, where classical orthographic lifting can introduce distortions near image boundaries. By learning the mapping, the model can implicitly compensate for such distortions. However, this also means that robustness to changes in camera intrinsics (e.g., different FOVs or lens distortions) is not explicitly enforced. A promising future direction is to train the projection module under diverse intrinsic settings to support the model to learn a more general and transferable projection function.

B Spectral Realization of the Equivariance Correction

In this section, we provide a concrete spectral realization of the equivariance correction introduced in Proposition 1, using the spherical-harmonic coefficients and Wigner D-matrices.

Proof. Let x be the observation with camera pose $R_x \in \mathrm{SO}(3)$ and let $c_\ell(x) \in \mathbb{R}^{2\ell+1}$ denotes spherical harmonic coefficients. Under a global rotation $g \in \mathrm{SO}(3)$ applied to both the scene and the camera, the camera pose transforms as $R_x \mapsto R_{gx} = gR_x$. Since the signal $\Phi(x)$ is expressed in the local (camera) frame, the spherical coefficients remain unchanged under the global transformation, so $c_\ell(gx) = c_\ell(x)$. Applying Equation 2 with the updated camera pose, the corrected coefficients at gx are:

$$c_{\ell,\mathrm{corr}}(gx) = D^\ell(R_{gx})\, c_\ell(gx) = D^\ell(gR_x)\, c_\ell(x). \tag{6}$$

Since the Wigner D-matrices D^ℓ form a group representation of $\mathrm{SO}(3)$, they satisfy the homomorphism property: $D^\ell(gR_x) = D^\ell(g)\, D^\ell(R_x)$. Substituting this, we obtain:

$$c_{\ell,\mathrm{corr}}(gx) = D^\ell(g)\, D^\ell(R_x)\, c_\ell(x) \tag{7}$$

Recognizing that $c_{\ell,\mathrm{corr}}(x) = D^\ell(R_x)\, c_\ell(x)$ by Proposition 1, we conclude:

$$c_{\ell,\mathrm{corr}}(gx) = D^\ell(g)\, c_{\ell,\mathrm{corr}}(x) \tag{8}$$

Thus, the corrected coefficients $c_{\ell,\mathrm{corr}}(x)$ transform equivariantly under the group action $g \in \mathrm{SO}(3)$. □

This result shows that equivariance correction can be implemented spectrally by left-multiplying the spherical harmonic coefficients with Wigner D-matrices according to the camera orientation. This aligns the signal, originally expressed in the camera frame, to a common world frame for consistent and equivariant downstream processing across varying viewpoints.

C Implementation of Our Policy

Our model consists of an $\mathrm{SO}(3)$-equivariant observation encoder followed by an $\mathrm{SO}(3)$-equivariant diffusion module, both implemented using `escnn` [3] and `e3nn` [57].

Given an observation $x \in X$, the $\mathrm{SO}(2)$-equivariant image encoder λ first maps the RGB image I into a regular representation, which is then mapped to a trivial representation $\lambda(I) \in \mathbb{R}^{n\times h\times w}$, where n, h, and w denote the number of channels, height, and width, respectively. These 2D features are lifted to the sphere via orthographic projection, producing a signal $\Phi(x)$ on S^2. To account for varying viewpoints, we use the gripper orientation R_x as an *equivariance correction* factor to align the spherical signal into a common reference frame. In our setup, the wrist-mounted camera is rigidly attached to the gripper, so the gripper orientation provides a fixed proxy for the camera pose. This

approximation is sufficient for aligning the image features with the proprioceptive signals, and any minor misalignments can be further handled by the equivariant convolution layers. The corrected signal is then processed by a sequence of $\mathrm{S}^2 \to \mathrm{SO}(3)$ and $\mathrm{SO}(3) \to \mathrm{SO}(3)$ spherical convolution layers to generate the signal $\Psi(x)$ on $\mathrm{SO}(3)$. The proprioceptive state is encoded using the irreps ρ_0 and ρ_1, and passed through $\mathrm{SO}(3)$-equivariant linear layers to yield Fourier coefficients of the same type as $\Psi(x)$. These are then concatenated with the image signal to form the global conditioning vector $e_{\mathbf{o}} \in \mathbb{R}^{u \times d_{\mathbf{o}}}$, where u is the number of channels and $d_{\mathbf{o}}$ is the feature dimension. Similarly, the noisy action chunk $\mathbf{a}^k$ is embedded into $e_{\mathbf{a}} \in \mathbb{R}^{u \times d_{\mathbf{a}} \times n}$, where $d_{\mathbf{a}}$ denotes the number of action feature channels and n the number of time steps. An inverse FFT is applied to sample both $e_{\mathbf{o}}$ and $e_{\mathbf{a}}$ onto discrete subgroups, either the icosahedral group $I_{60} \subset \mathrm{SO}(3)$ or the cyclic group $C_8 \subset \mathrm{SO}(2)$, producing $e_{\mathbf{o}} \in \mathbb{R}^{p \times d_{\mathbf{o}}}$, $e_{\mathbf{a}} \in \mathbb{R}^{p \times d_{\mathbf{a}} \times n}$, where $p = 60$ or 8 is the number of group elements. For each group element $g \in I_{60}$ or $g \in C_8$, a shared $\mathrm{SO}(3)$- or $\mathrm{SO}(2)$-equivariant 1-D temporal U-Net processes the action sequence $e^g_{\mathbf{a}}$, conditioned on the observation $e^g_{\mathbf{o}}$ and diffusion step k. This design follows the point-wise equivariant processing strategy proposed in [65], ensuring equivariance across group elements. Finally, an equivariant decoder maps the denoised representation to the noise estimate ϵ^k.

D Simulation Settings

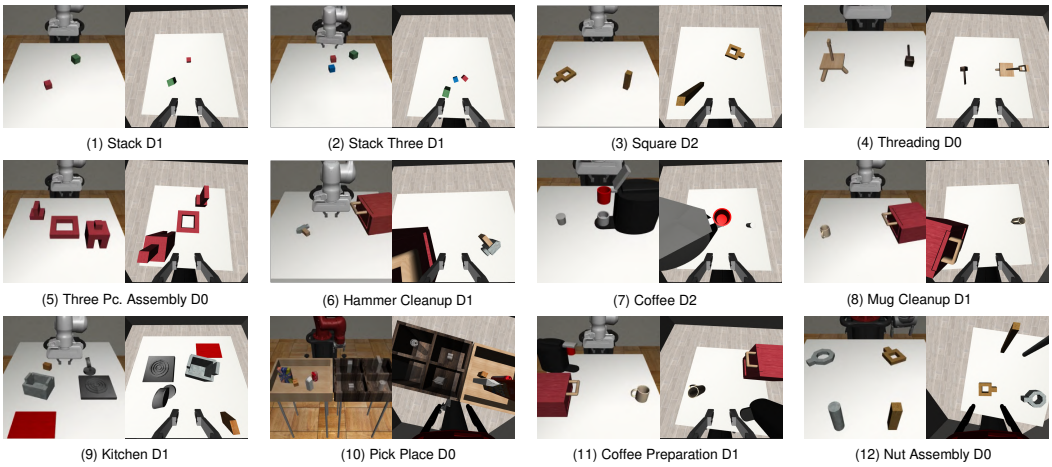


Figure 8: The twelve simulation tasks from the MimicGen [40] simulator. In each subfigure, the left image shows the task scene, while the right image shows the corresponding eye-in-hand view.

Figure 8 illustrates the twelve tasks in the MimicGen simulation. In each subfigure, the left image shows the full environment scene from the agent's view, while the right image is the eye-in-hand RGB observation used by the model. Following prior work [40, 65], we set the resolution of the eye-in-hand image to $3 \times 84 \times 84$ and adopt the same maximum episode length. To enable the wrist-mounted camera to capture more contextual information, we increase its FOV from 75 to 130 degrees, similar to that of a typical wide-angle camera.

E Training Details

For the simulation experiments, we follow the hyperparameter settings from prior work [65, 5]. In detail, we use an observation window of two history steps for ISP-$\mathrm{SO}(3)$ and one step for ISP-$\mathrm{SO}(2)$. In both cases, the denoising network outputs a sequence of 16 action steps, which are used for optimization during training, while only the first 8 steps are executed during evaluation. During training, input images are randomly cropped to a resolution of 76×76, while a center crop is applied at evaluation time. We train all models using the AdamW [38] optimizer with Exponential Moving Average, and adopt the DDPM [15] framework with 100 denoising steps for both training and evaluation. For all baselines, we retain their original hyperparameter settings for evaluation and only adjust the number of training steps to ensure consistency across methods. All methods are trained on the same dataset and evaluated using three random seeds.

https://doi.org/10.52202/085713-0201

For the real-world experiments, we use the same hyperparameters as in the simulation, except that we replace DDPM with DDIM [55] for both training and evaluation, and reduce the number of denoising steps to 16 at evaluation time. However, we find that using a resolution of 76×76 is insufficient for fine-grained manipulation in the real world, as the extremely wide FOV from the GoPro camera causes each pixel to correspond to a relatively large spatial region in the original setting. To address this, we increase the input resolution to 224×224. Specifically, starting from the original 720×720 RGB image captured using a GoPro with the Max Lens Mod, we apply a center crop of size 480×480, followed by resizing to 224×224. In addition, we apply standard data augmentations, including random cropping, rotation, and color jitter, to improve the robustness of both our method and the baselines.

All models are trained on single GPUs using compute clusters and workstations equipped with multiple high-performance consumer-grade GPUs.

F Full Simulation Experiment Results with Standard Deviations

Table 5 presents the same results as Table 1, with standard deviations included.

Table 5: Maximum success rates (%) on MimicGen tasks with 100 and 200 demonstrations across different methods, averaged over three random seeds. The $\pm$ indicates standard deviation.

	Stack D1		Stack Three D1		Square D2		Threading D0	
Method	100	200	100	200	100	200	100	200
ISP-SO(3)	99.3 ± 1.2	100.0 ± 0.0	70.0 ± 2.0	88.0 ± 2.0	34.7 ± 4.2	51.3 ± 2.3	90.0 ± 2.0	92.0 ± 0.0
ISP-SO(2)	98.0 ± 2.0	100.0 ± 0.0	74.7 ± 7.6	88.0 ± 2.0	32.0 ± 0.0	50.7 ± 5.0	84.7 ± 1.2	87.3 ± 3.1
DiffPo	90.7 ± 4.2	96.0 ± 2.0	43.3 ± 4.2	76.7 ± 4.2	12.0 ± 2.0	25.3 ± 3.1	77.3 ± 10.3	86.7 ± 7.0
EquiDiff	96.0 ± 0.0	98.7 ± 1.2	61.3 ± 5.0	80.0 ± 2.0	8.7 ± 1.2	19.3 ± 1.2	88.7 ± 5.8	92.0 ± 2.0
ACT	45.3 ± 7.6	77.3 ± 2.3	12.0 ± 2.0	36.7 ± 9.9	2.7 ± 1.2	10.0 ± 2.0	36.0 ± 6.9	53.3 ± 6.1
	Three Pc. Assembly D0		Hammer Cleanup D1		Mug Cleanup D1		Coffee D2	
Method	100	200	100	200	100	200	100	200
ISP-SO(3)	70.7 ± 1.2	79.3 ± 1.2	66.0 ± 0.0	73.3 ± 1.2	54.0 ± 8.7	58.7 ± 2.3	64.0 ± 0.0	68.7 ± 3.1
ISP-SO(2)	74.7 ± 1.2	80.0 ± 2.0	70.7 ± 1.2	73.3 ± 2.3	56.0 ± 2.0	60.7 ± 1.2	58.7 ± 3.1	63.3 ± 4.2
DiffPo	72.7 ± 3.1	73.3 ± 2.3	58.7 ± 7.6	63.3 ± 11.7	49.3 ± 8.3	61.0 ± 1.7	53.3 ± 3.1	54.7 ± 4.2
EquiDiff	74.0 ± 5.3	78.7 ± 1.2	59.3 ± 4.2	74.0 ± 2.0	50.7 ± 2.3	62.0 ± 0.0	47.3 ± 3.1	61.3 ± 2.3
ACT	28.0 ± 4.0	50.0 ± 5.3	34.7 ± 2.3	62.7 ± 5.8	24.7 ± 3.1	37.3 ± 5.8	20.7 ± 3.1	34.7 ± 2.3
	Kitchen D1		Pick Place D0		Coffee Preparation D1		Nut Assembly D0	
Method	100	200	100	200	100	200	100	200
ISP-SO(3)	75.3 ± 3.1	79.3 ± 4.2	42.0 ± 4.4	65.7 ± 5.5	40.7 ± 2.3	61.3 ± 2.3	75.3 ± 2.5	82.0 ± 7.5
ISP-SO(2)	64.7 ± 2.3	72.0 ± 2.0	45.7 ± 8.0	61.0 ± 5.6	46.7 ± 4.6	56.0 ± 0.0	73.7 ± 7.6	84.3 ± 1.5
DiffPo	60.7 ± 8.1	70.7 ± 3.1	36.3 ± 2.1	47.7 ± 1.5	37.3 ± 1.2	52.0 ± 5.3	51.3 ± 3.8	62.3 ± 1.5
EquiDiff	55.3 ± 1.2	66.7 ± 2.3	27.7 ± 2.9	46.3 ± 3.5	27.3 ± 1.2	38.7 ± 2.3	40.0 ± 4.0	56.3 ± 3.1
ACT	21.3 ± 1.2	50.7 ± 3.1	8.7 ± 1.5	13.7 ± 2.5	7.3 ± 2.3	16.0 ± 2.0	36.7 ± 1.2	49.0 ± 2.0

G Invariance via Delta Control vs. Equivariance via Rotation Correction

One of the core components of our method is the rotation correction step, which aligns the spherical signals to a common reference frame to preserve $\mathrm{SO}(3)$-equivariance throughout the policy pipeline. A natural alternative is to remove this step and instead express actions in the moving gripper frame, referred to as *delta actions* in [6], which can also be interpreted as a sequence of incremental transforms. This formulation leads to an $\mathrm{SE}(3)$-invariant system, as both perception and action are expressed relative to the gripper's local frame. This raises an important question: *Is rotation correction necessary if delta actions can achieve similar symmetry properties through invariance?*

Empirical Evidence To investigate this, we conducted additional experiments comparing absolute and delta action control on two MimicGen tasks: Square D2 and Nut Assembly D0. Specifically, we evaluated (a) a variant of our method without rotation correction that uses delta control, and (b) the original Diffusion Policy with delta control. Table 6 summarizes the results with 100 demonstrations.

Table 6: Comparison of absolute and delta control on two MimicGen tasks with 100 demonstrations. Values in parentheses indicate performance differences relative to ISP-SO(2) (Absolute).

Method	Square D2	Nut Assembly D0
ISP-SO(2) (Absolute)	32	74
ISP-SO(2) (Delta, No Rotation Correction)	22 (-10)	57 (-17)
DiffPo (Absolute)	12 (-20)	51 (-23)
DiffPo (Delta)	14 (-18)	22 (-52)

Figure 9: Real-world experimental setup. We use a UR5 robot equipped with a Robotiq-85 gripper and custom-designed soft fingers. A GoPro camera is mounted on the wrist to capture visual observations. Demonstrations are collected using the Gello teleoperation interface (bottom right).

We observe that absolute action consistently outperforms delta action. Similar trends have been reported in prior work, such as EquiDiff [65] and the Diffusion Policy [5], where delta or velocity control often results in inferior performance.

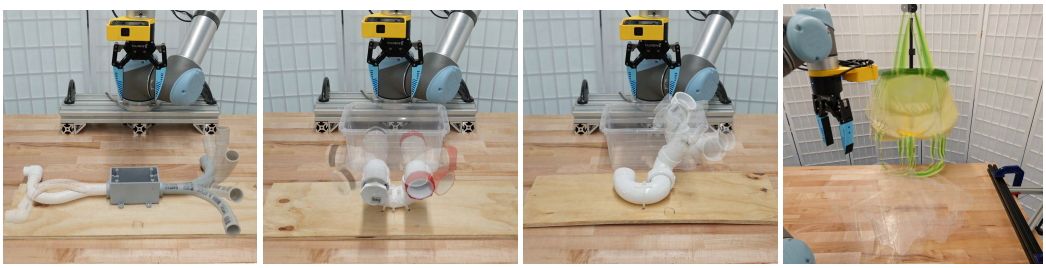
Generalization Perspective Although gripper-relative control guarantees invariance under single-camera setups, it does not generalize seamlessly to multi-camera or hybrid sensing configurations, where additional viewpoints can break this invariance assumption. In contrast, aligning both observations and actions to a shared world frame establishes a consistent global reference across all sensors. This property supports more flexible sensor integration and improved generalization in complex environments with multiple or moving viewpoints.

H Details of the Real-World Experiment

Figure 9 shows our real-world experimental setup. Demonstrations are collected using the Gello teleoperation interface [68]. While the robot is teleoperated in joint space, we record end-effector actions, including position, rotation, and gripper state. Visual observations and actions are recorded synchronously at each timestep.

Figure 10 illustrates the initial state distributions for each task. In *Box-Pipe Disassembly*, two pipes with different colors are connected to a junction box, where one pipe shares the same color as the box may confuse the policy. The orientations of the pipes are randomized. In *U-Pipe Disassembly*, four pipe fittings are arranged in a U-shape and initialized with random rotations. In *3D-Pipe Disassembly*, two pipes are connected with independently randomized 3D orientations. In *Grocery Bag Retrieval*, a toy banana is randomly placed inside a deformable plastic bag. The robot must reach into the bag, identify and retrieve the banana, and place it into a transparent container with minor positional

https://doi.org/10.52202/085713-0201



(a) Box-Pipe Disassembly (b) U-Pipe Disassembly (c) 3D-Pipe Disassembly (d) Grocery Bag Retrieval

Figure 10: Distribution of random initial states used in the real-world experiments.

variation. All subfigures in Figure 10 show averaged visualizations across multiple randomized initializations.

We visualize one episode for each task in Figure 11. These tasks emphasize different aspects. The pipe disassembly tasks require precise, closed-loop control to smoothly extract the pipes. This makes them particularly challenging for open-loop policies. The Grocery Bag Retrieval task highlights the importance of the eye-in-hand camera, as the target object is difficult to perceive and localize using only external views.

I Practical Guidelines for Data Collection

Before starting data collection on the real robot, it is critical to establish a **predefined task execution strategy** to ensure motion simplicity, efficiency, and cross-operator consistency. Such a strategy typically involves defining consistent action sequences, execution ranges, and task progression patterns, helping to avoid ambiguous or poorly structured scenarios that may lead to robotic indecision or an undesirably large amount of multimodal behavior during training.

Based on our experience, during data collection, demonstrations should:

1. Uniformly cover as many task-relevant initial states as possible.
2. Maintain a consistent end-effector speed within and between trajectories without interruption.
3. Avoid unnecessary stops, pauses, or other irregular motion patterns.
4. Synchronize sensing and control to minimize latency-induced artifacts.
5. Regularly verify alignment between the robot and sensors to prevent drift and maintain data consistency.

After data collection, all trajectories should be automatically or visually inspected to detect potential issues. In particular, segments exhibiting robotic hesitation or stalling, most commonly near the beginning and end of each demonstration, as well as episodes containing negative or low-quality behavior, should be identified and removed. Consistent inspection and pruning of low-quality data can significantly improve the stability and performance of policy learning.

These practical steps help ensure that the collected data are clean, diverse, and informative, which can ultimately enhance the robustness and generalization of learned visuomotor policies.

J Real World Experiment Failure Analysis

In the *Box-Pipe Disassembly* task, one of the primary failure cases arises from the inability to distinguish between the gray junction box and the gray pipe. For the original Diffusion Policy, the policy consistently misidentifies the box as the pipe to be disassembled, which triggers the robot's emergency stop. While our method occasionally encounters the same issue, the failure rate is significantly lower. This suggests that our method is more data-efficient and better at learning robust visual distinctions from limited demonstrations.

Table 7: Comparison of training and inference efficiency on a single RTX 4090 GPU.

Method	Training Speed (rel.)	Inference Time (ms)	Real-Time Capable
DiffPo	$1\times$	68	Yes
ISP-SO(2)	$2.6\times$ slower	63	Yes
ISP-SO(3)	$5.4\times$ slower	75	Yes

In the *U-Pipe Disassembly* task, a common failure mode for our method and the baseline occurs when pulling the red pipe inadvertently causes a connected pipe to be extracted as well. In such cases, the robot grasps both pipes simultaneously. We consider this a partial success. However, the baseline additionally suffers from incorrect orientation predictions under certain initial states, leading to more frequent failures.

In the *3D-Pipe Disassembly* task, our method occasionally fails to identify the correct grasp orientation. In contrast, the baseline struggles consistently with this issue and rarely completes the task successfully. One major contributing factor is the multimodality of the task. During data collection, it is difficult to maintain consistency in demonstration strategies because pipes can be grasped in multiple orientations. Nevertheless, by incorporating 3D symmetries, our method is more robust to such variations and generalizes better across diverse configurations.

In the *Grocery Bag Retrieval* task, failure cases primarily result from unsuccessful grasp attempts or inaccuracies during the placement phase. The deformable nature of the bag and the partial occlusion of the banana present additional challenges, especially under limited visual feedback.

K Computational Efficiency Analysis

In this section, we provide quantitative comparisons of the computational efficiency of our method during both training and inference. Our results are measured on a single RTX 4090 GPU.

Training Efficiency Compared to the original Diffusion Policy [5], ISP-SO(2) is approximately $2.6\times$ slower, and ISP-SO(3) is approximately $5.4\times$ slower during training. This increase is expected due to the added computational complexity of the equivariant layers. Nevertheless, the training speed remains practical for large-scale policy learning.

Inference Efficiency Despite the higher training cost, our method maintains high efficiency during inference, making it well-suited for real-time deployment. Table 7 summarizes the average inference time of each method in the real-world settings with 16-step DDIM sampling [55]. All methods exhibit comparable inference speeds. The SO(2) variant is slightly faster than the baseline, primarily due to its lighter-weight diffusion U-Net and the use of a smaller history observation window. Although the SO(3) variant is marginally slower, its inference time ($\sim$75 ms) remains close to DiffPo ($\sim$68 ms), well within real-time control requirements (e.g., 10 Hz).

L On Limitations and Practical Considerations of Equivariance

Interaction with Real-World Asymmetries Equivariance may face challenges in manipulation scenarios where asymmetries in the physical world are important. A representative example is tasks involving asymmetric robot kinematics, such as left-right differences in reachable workspace. Although equivariance allows the model to generalize across rotated scenes, joint limits are not preserved under rotation, which may lead to infeasible or suboptimal actions. Another example is manipulation involving heavy objects, where gravity breaks rotational symmetry in practice. An object that is easy to manipulate in one orientation may become unstable or infeasible to lift when rotated. Despite these challenges, prior work has shown that equivariant models can remain robust in the presence of symmetry-breaking factors such as visual appearance, camera pose, and shadows [62]. These asymmetries are already encoded in the input, allowing the model to learn appropriate behaviors without violating the equivariant structure. While cases where equivariance leads to performance degradation are relatively uncommon, they do highlight scenarios where symmetry-breaking mechanisms may be beneficial. A promising future direction is to augment

https://doi.org/10.52202/085713-0201

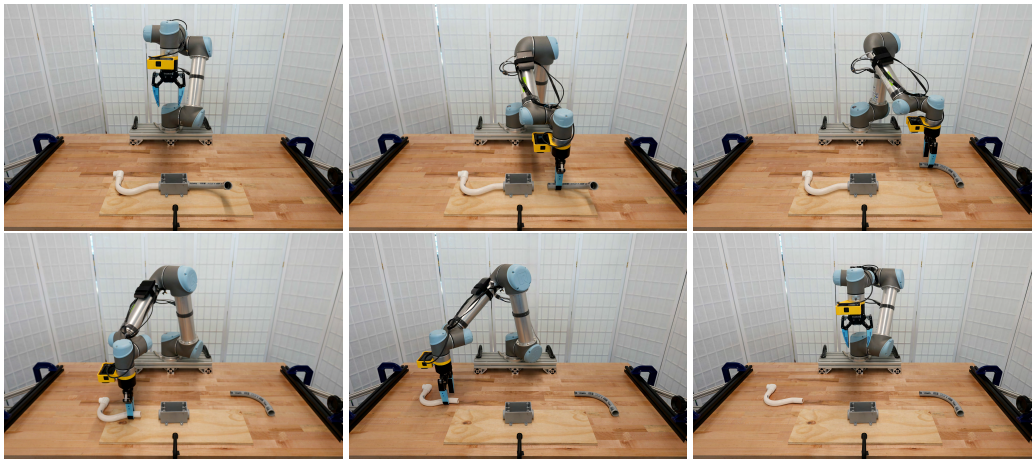
equivariant architectures with non-equivariant components or task-specific inductive biases (e.g., gravity-aware priors or joint-limit encodings) to better capture real-world asymmetries.

Discretization of Continuous Symmetry Groups In our framework, equivariance is enforced by sampling finite subgroups of $\mathrm{SO}(3)$ (e.g., I_{60} or C_8) and applying a shared U-Net across the sampled group elements. This approximation may, in principle, introduce discrepancies for rotations outside the sampled set. Empirically, however, no significant performance degradation was observed. In fact, discrete subgroups often lead to superior performance compared to continuous irreducible representations, consistent with prior findings in equivariant learning [3, 66]. While this approach reduces the theoretical degree of symmetry, it provides a more scalable and expressive modeling strategy by avoiding the computational overhead and activation function constraints associated with continuous irreps. Similar strategies have also demonstrated strong empirical effectiveness in other robotics applications [65, 20]. To further mitigate potential limitations and improve generalization, we apply random $\mathrm{SO}(2)$ rotations as a data augmentation strategy to the end-effector pose in both proprioceptive inputs and actions during training. Additionally, subgroup sampling is not restricted to a single set: multiple sets can be employed in practice to increase angular coverage when necessary. Finally, while equivariant architectures introduce structural inductive biases, they do not inherently limit the model's ability to generalize beyond the sampled rotations. With sufficient data diversity and augmentation, the network is able to interpolate smoothly across $\mathrm{SO}(3)$, thereby alleviating the potential impact of discretization.

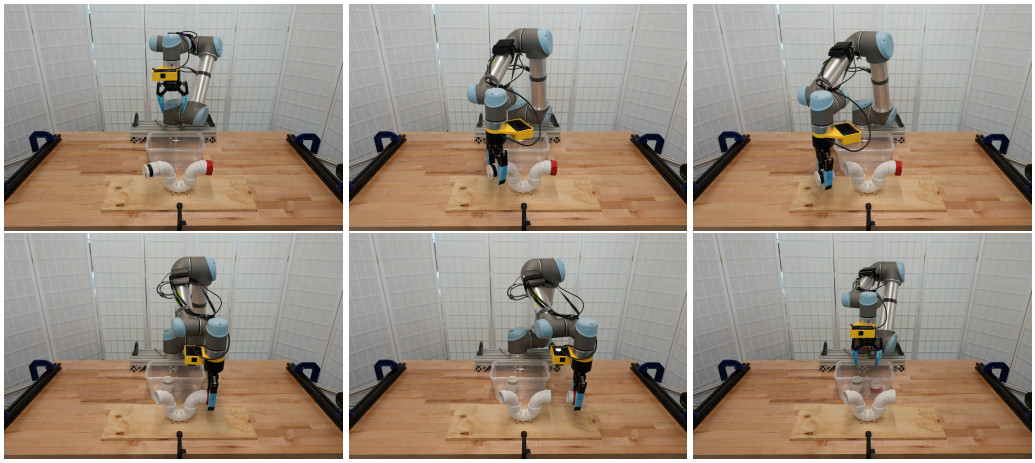
M Broader Impact

This work has several potential social impacts, both positive and negative. On the positive side, our proposed method enables more data-efficient and generalizable robot policy learning in 3D environments. This can facilitate the development of more robust and capable household robots, particularly in settings where labeled demonstrations are limited. Moreover, by leveraging geometric symmetries and closed-loop visuomotor control from wrist-mounted cameras, our method could lower the barrier for deploying autonomous robots in unstructured real-world environments, thereby expanding accessibility and utility.

However, as with many data-driven learning methods, our approach inherits limitations tied to the quality and intent of the training data. Since the robot policy is learned entirely through imitation, any unsafe, biased, or suboptimal behavior demonstrated during data collection may be reflected in the final policy. Furthermore, the increased autonomy enabled by our method underscores the importance of safety monitoring and responsible deployment, especially in applications involving human interaction.

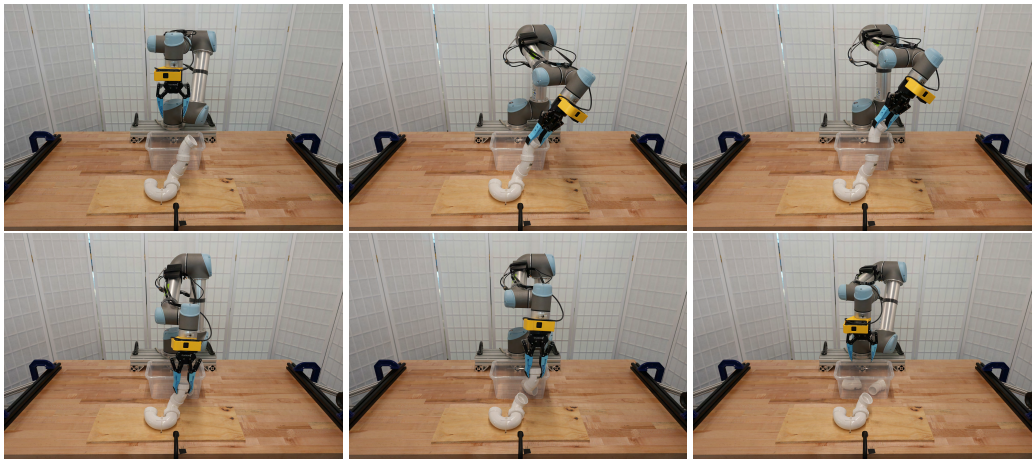


(a) Box-Pipe Disassembly

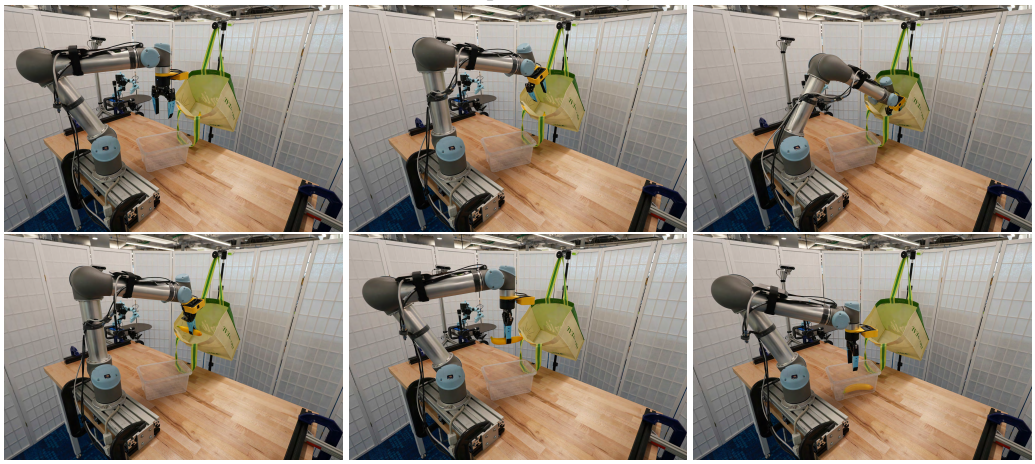


(b) U-pipe Disassembly

https://doi.org/10.52202/085713-0201



(c) 3D-Pipe Disassembly



(d) Grocery Bag Retrieval

Figure 11: Visualization of one episode for each task. Each subfigure illustrates a key action step in the trajectory.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the core claims of the paper, including our key contributions and the assumptions. These claims are well-supported by both the theoretical analysis and experimental results presented.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

https://doi.org/10.52202/085713-0201

Justification: The paper includes a discussion of the limitations of our method. Specifically, we clarify the computational efficiency issues introduced by using equivariant networks, and discuss limitations related to the task settings (e.g., single-arm manipulation and single-view observations). We also outline potential directions for future research to address these limitations and extend the applicability of our approach.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All definitions, propositions, and assumptions are clearly stated, numbered, and cross-referenced.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We will submit the code as supplementary material and release the code implementation for all experiments in this paper. Detailed descriptions of our method are also provided in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We include the complete code for data generation and all models in the supplementary material, ensuring full reproducibility of our results. A public GitHub repository will be provided with the final version of the paper.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.

https://doi.org/10.52202/085713-0201

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: In the Appendix, we provide the training details of our method, along with the code as well.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the full experimental results with standard errors computed over multiple random seeds.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide the details of the compute resources in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: Our work conforms, in every respect, with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss the broader impacts of our work in the appendix, which consists of both potential positive and negative societal implications.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

https://doi.org/10.52202/085713-0201

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The dataset used in this work is publicly available, and our method does not pose a high risk of misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have cited all relevant prior works, datasets, and software packages used in this paper. License, copyright information, and terms of use will be provided in our GitHub repository.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The related docs will be included in the supplementary material.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

https://doi.org/10.52202/085713-0201