A Practical Guide for Incorporating Symmetry in Diffusion Policy

Dian Wang[1] **Boce Hu**[2] **Shuran Song**[1] **Robin Walters**[2†] **Robert Platt**[2†]
[1]Stanford University [2]Northeastern University
https://sym-in-dp.github.io

Abstract

Recently, equivariant neural networks for policy learning have shown promising improvements in sample efficiency and generalization, however, their wide adoption faces substantial barriers due to implementation complexity. Equivariant architectures typically require specialized mathematical formulations and custom network design, posing significant challenges when integrating with modern policy frameworks like diffusion-based models. In this paper, we explore a number of straightforward and practical approaches to incorporate symmetry benefits into diffusion policies without the overhead of full equivariant designs. Specifically, we investigate (i) invariant representations via relative trajectory actions and eye-in-hand perception, (ii) integrating equivariant vision encoders, and (iii) symmetric feature extraction with pretrained encoders using Frame Averaging. We first prove that combining eye-in-hand perception with relative or delta action parameterization yields inherent SE(3)-invariance, thus improving policy generalization. We then perform a systematic experimental study on those design choices for integrating symmetry in diffusion policies, and conclude that an invariant representation with equivariant feature extraction significantly improves the policy performance. Our method achieves performance on par with or exceeding fully equivariant architectures while greatly simplifying implementation.

1 Introduction

Although recent advancements in incorporating symmetry in robotic policy learning have shown promising results [52, 60], the practical impact of this approach remains limited due to the significant implementation challenges. While equivariant models [62, 11] can substantially improve sample efficiency and generalization, integrating these symmetric properties into modern policy learning frameworks presents several obstacles. First, symmetry reasoning must be tailored specifically for each policy formulation, requiring different mathematical analyses and architectures across different policy frameworks like Q-learning [57], actor-critic methods [59], and diffusion models [3]. This creates a steep learning curve for practitioners seeking to leverage symmetry benefits. Second, state-of-the-art equivariant architectures often introduce considerable complexity with specialized layers [4, 15] and require structured inputs [11, 34], which do not naturally align well with modern policy components like diffusion-based action generation [5], eye-in-hand perception [7], or pretrained vision encoders. As a result, robotics researchers frequently face a difficult choice: adopting complex equivariant architectures with specialized implementation, or maintaining practical implementation concerns at the cost of symmetry benefits. This dilemma is particularly pronounced in diffusion-based visuomotor policies [5], which have emerged as a powerful paradigm for generating smooth, multimodal robot actions. Although prior works have tried to implement equivariant

[†]Equal Advising

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

diffusion models [51, 65, 60], the diffusion-denoising nature significantly enhances the difficulty of incorporating equivariant structure.

In this paper, we present a practical guide for incorporating symmetry into diffusion policies without requiring significant design overhead or sacrificing the advantages of modern policy formulations. We systematically investigate several straightforward approaches that achieve this balance: (i) invariant representations through relative trajectory actions and eye-in-hand perception, (ii) equivariant vision encoders that can be incorporated into standard diffusion frameworks, and (iii) Frame Averaging [48] techniques that enable symmetrization with pretrained encoders. Our approach offers a compelling alternative to end-to-end equivariant architectures: rather than choosing between symmetry benefits and implementation practicality, our methods demonstrate that these advantages can be achieved with minimal architectural changes and overhead. As shown in Figure 1, our method achieves excellent performance while maintaining implementation simplicity, positioning it at an optimal balance point in the symmetry-practicality tradeoff. Notably, our work using a single eye-in-hand image input reaches a similar performance as Wang et al. [60], which uses four cameras to reconstruct the 3D voxel grid input.

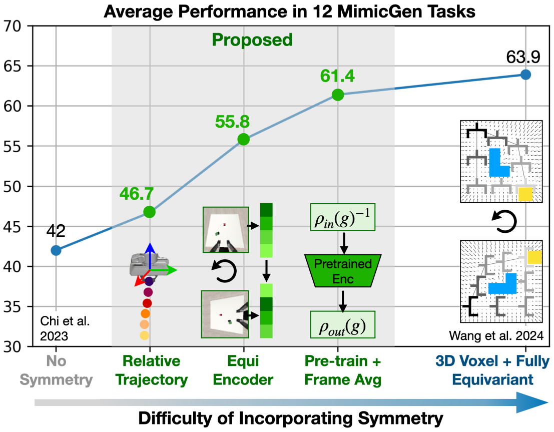

Figure 1: We propose a number of practical approaches for incorporating symmetry in Diffusion Policy, achieving comparable performance as fully-equivariant policies while maintaining simplicity.

Our contributions can be summarized as follows:

- We prove that eye-in-hand perception and relative trajectory action inherently possesses $\mathrm{SE}(3)$-invariance, significantly improving the policy's generalization.
- We demonstrate that transitioning from absolute action representations to relative trajectory actions provides a straightforward improvement in policy performance, both with eye-in-hand perception and extrinsic perception.
- We propose a novel approach that integrates a symmetric encoder into standard diffusion policy learning, achieved by either using equivariant network in end-to-end training or using Frame Averaging with pretrained encoders, and show that it can significantly improve performance.
- We show that combining the invariant perception and action representations and a pretrained encoder with Frame Averaging achieves the state-of-the-art results in the MimicGen [42] benchmark while maintaining low architectural complexity and computational overhead compared to fully equivariant methods.

2 Related Work

Diffusion Policies: Denoising diffusion models have transformed generative modeling in vision, achieving state-of-the-art results in image and video synthesis [18, 55] as well as planning [29, 33]. Recently, Chi et al. [5, 6] introduced *Diffusion Policy*, which extends diffusion models to robotic visuomotor control by denoising action trajectories conditioned on observations. Subsequent extensions include applications to reinforcement learning [61, 49], incorporation of 3D inputs [68, 31], hierarchical policies [64, 40, 71], and large vision-language action models [45, 2]. A key limitation of diffusion methods is their heavy demand for training data. To mitigate this, recent works have injected domain symmetries as equivariant constraints into the denoising network, thereby boosting sample efficiency and generalization [3, 51, 60, 65, 56]. However, they typically require complex equivariant denoising models. In contrast, our approach integrates symmetry by combining invariant observation and action representations with an equivariant vision encoder, greatly simplifying implementation.

Equivariant Policy Learning: Robotic policies often require generalizing across spatial transformations of the environment. Traditional methods often achieve this via extensive data augmentation [69, 70]. Recently, *equivariant models*, a class of methods that are mathematically constrained to be equivariant [8, 9, 62, 4, 11, 15, 34, 35], has been widely adapted to robotics to automatically instantiate spatial generalization. Such models have been widely applied across robot learning, including equivariant reinforcement learning [57, 59, 58, 44, 43, 32, 37, 19], imitation learning [30, 66, 14], grasp learning [74, 75, 24, 21, 36], and pick-place policies [52, 53, 46, 50, 22, 25, 23, 27, 13, 26]. However, these methods often demand complex symmetry reasoning and equivariant layers, which can hinder scalability. By contrast, our framework introduces symmetry in a more modular fashion, making it easier to implement and adapt.

Policy Learning using Eye-in-Hand Images: Eye-in-hand perception, using a camera mounted on the robot's end-effector, has been a popular choice in manipulation because it is simple and calibration-free. For instance, Jangir et al. [28] learn a shared latent space between egocentric and external views to train hybrid-input policies. Hsu et al. [20] show that eye-in-hand images (alone or combined with external cameras) yield higher success rates and better generalization, and similar findings have been captured in [41]. Consequently, many recent frameworks retain eye-in-hand imagery [72, 1, 73, 2, 38]. Another advantage is rapid data collection using a handheld gripper [54, 67, 47, 7, 63, 16]. In this work, we theoretically analyze how eye-in-hand observation, when paired with relative or delta trajectory actions, yields inherent symmetry advantages for diffusion-based policies.

3 Background

Problem Statement: We consider behavior cloning for visuomotor policy learning in robotic manipulation, where the goal is to learn a policy that maps an observation o to an action a, mimicking an expert policy. Both o and a may span multiple time steps, i.e., $o = \{o_{t-(m-1)}, \dots, o_{t-1}, o_t\}, a = \{a_t, a_{t+1}, \dots, a_{t+(n-1)}\}$, where m is the number of past observations and n is the number of future action steps. At time step t, the observation $o_t = (I_t, T_t, w_t)$ contains the visual information I_t, (e.g., images), the pose of the gripper in the world frame $T_t \in \mathrm{SE}(3)$, as well as the gripper aperture $w_t \in \mathbb{R}$. The action $a_t = (A_t, w_t)$ specifies a target pose $A_t \in \mathrm{SE}(3)$ of the gripper and an open-width command $w_t \in \mathbb{R}$. To simplify the notation for our analysis, we omit the gripper command w_t in the action and focus on the pose command by writing $a = \{A_t, A_{t+1} \dots, A_{t+n-1}\}$ (while in the actual implementation the policy controls both the pose and the aperture).

Diffusion Policy: Chi et al. [5] introduced Diffusion Policy, which formulates the behavior cloning problem as learning a Denoising Diffusion Probabilistic Model (DDPMs) [18] over action trajectories. Diffusion Policy learns a noise prediction function $\varepsilon_\theta(o, a+\varepsilon^k, k) = \varepsilon^k$ using a network ε_θ, which is trained to predict the noise ε^k added to an action a. The training loss is $\mathcal{L} = ||\varepsilon_\theta(o, a+\varepsilon^k, k) - \varepsilon^k||^2$, where ε^k is a random noise conditioned on a randomly sampled denoising step k. At inference, starting from $a^k \sim \mathcal{N}(0, 1)$, the model iteratively denoises

$$a^{k-1} = \alpha(a^k - \gamma\varepsilon_\theta(o, a^k, k) + \epsilon), \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. α, γ, σ are functions of the denoising step k (also known as the noise schedule). The action a^0 is the final executed action trajectory.

Equivariance: A function $f\colon X \to Y$ is *equivariant* to a group G if, for all $g \in G$,

$$f\big(\rho_x(g)\, x\big) = \rho_y(g)\, f(x),$$

where ρ_x and ρ_y are representations of G on X and Y. Equivariance ensures that applying g before or after f yields the same result, thus f is G-symmetric. When ρ is clear, we often write $g \cdot x$ or gx.

For 2D images, the group $\mathrm{SO}(2) = \{\mathrm{Rot}_\theta : 0 \leq \theta < 2\pi\}$ of planar rotations and its subgroup $C_u = \{\mathrm{Rot}_\theta : \theta \in \{\frac{2\pi i}{u} | 0 \leq i < u\}\}$ of rotations by multiples of $\frac{2\pi}{u}$ are often used to construct equivariant neural networks [62, 4] that can capture rotated features. The *regular representation* $\rho_{\mathrm{reg}} : G \to \mathbb{R}^{u\times u}$ of C_u is of particular interest of this paper, which defines how C_u acts on a vector $x \in \mathbb{R}^u$ by $u \times u$ permutation matrices. Intuitively, the vector x can be viewed as containing information for each rotation in C_u. Let $r^v \in C_u = \{1, r^1, \dots r^{u-1}\}$ and $x = (x_1, \dots x_u) \in \mathbb{R}^u$, then $\rho_{\mathrm{reg}}(r^v)x = (x_{u-v+1}, \dots, x_u, x_1, x_2, \dots, x_{u-m})$ cyclically permutes the coordinates of x.

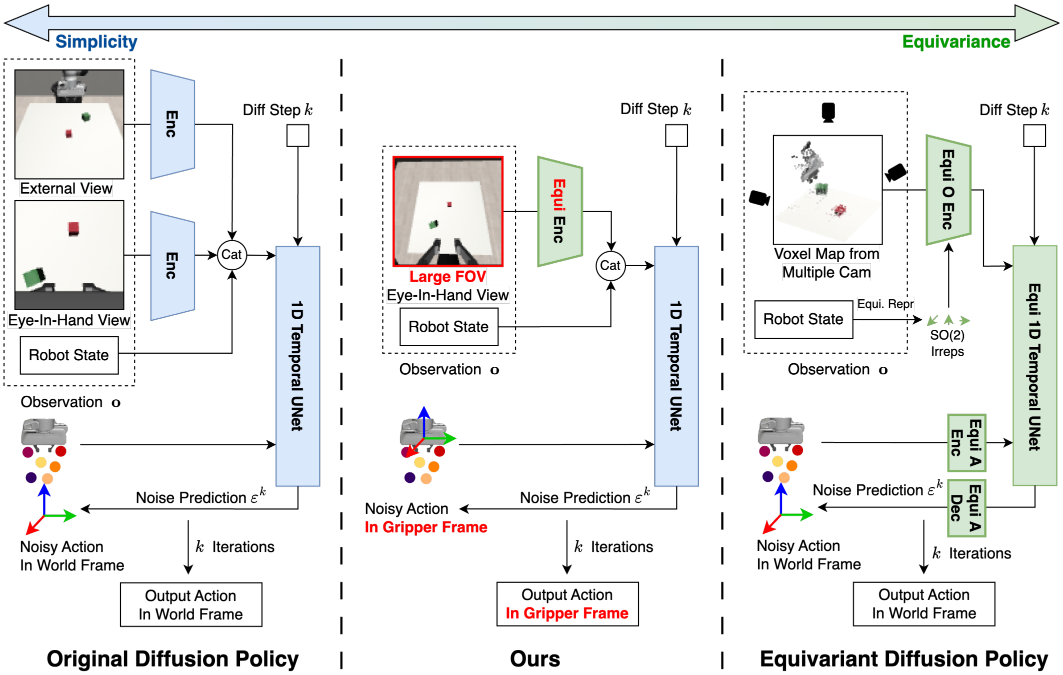

Figure 2: The difference comparing our method (middle) with the original Diffusion Policy (left) and the Equivariant Diffusion Policy (right). We use only an eye-in-hand image as the input, and use an equivariant encoder to acquire symmetry-aware features from the input image. In policy denoising, the noisy action and the noise-free action output are both in the gripper frame. Other components remain identical to the original Diffusion Policy. Compared with Equivariant Diffusion Policy (right), our approach is significantly simpler while maintaining a comparable experimental performance.

4 Approaches for Incorporating Symmetry in Diffusion Policy

In this section, we introduce three practical approaches for incorporating symmetry into diffusion policies without requiring complex end-to-end equivariant architecture design. First, we examine how invariant action and perception representations naturally induce symmetric properties. Second, we explore integrating equivariant vision encoders that extract symmetry-aware features while maintaining standard diffusion heads. Finally, we present how to leverage pre-trained vision encoders in an equivariant way through Frame Averaging [48]. Together, these approaches offer a spectrum of options for balancing symmetry benefits with implementation simplicity. As shown in Figure 2, our proposed approach (middle) requires minimal architectural change compared with the original Diffusion Policy [5] (left), and is much simpler than the fully equivariant model [60] (right).

4.1 Representing Actions as Absolute, Relative, and Delta Trajectory

Choosing the right representation for actions and observations is crucial for sample-efficient policy learning in robotic manipulation. When a robotic task and environment exhibit rigid-body symmetries, incorporating an equivariant or invariant representation can significantly enhance generalization to unseen object configurations and poses. In this section, we explore three natural trajectory action representations: absolute, relative, and delta trajectories, highlighting their symmetry properties under global $\mathrm{SE}(3)$ transformations. Notice that although relative trajectories were introduced by Chi et al. [7], their symmetric advantages have not yet been explored.

Definition 1 (Absolute Trajectory Action)**.** *An absolute trajectory action specifies future gripper poses directly in the world frame as*

$$a = \{A_t, A_{t+1}, \dots, A_{t+n-1}\},$$

where each $A_{t+i} \in \mathrm{SE}(3)$ *is the desired pose at time* $t+i$*.*

Definition 2 (Relative Trajectory Action)**.** *Let $T_t \in \mathrm{SE}(3)$ be the current gripper pose in the world frame. A* relative trajectory action *is a sequence*

$$a^r = \{A^r_t, A^r_{t+1}, \dots, A^r_{t+n-1}\},$$

where each $A^r_{t+i} \in \mathrm{SE}(3)$ specifies the gripper's pose relative to its initial frame at time t. The corresponding absolute poses are recovered via

$$A_{t+i} = T_t\, A^r_{t+i}, \qquad i = 0, \dots, n-1. \tag{2}$$

Definition 3 (Delta Trajectory Action)**.** *A* delta trajectory action *is a sequence of incremental transforms expressed in a moving local frame:*

$$a^d = \{A^d_t, A^d_{t+1}, \dots, A^d_{t+n-1}\},$$

where $A^d_{t+i} \in SE(3)$ represents the incremental motion at time step $t+i$ expressed relative to the gripper's frame at the previous time step $t+i-1$. The absolute poses are reconstructed as:

$$A_{t+i} = \; T_t \left(\prod_{j=0}^{i-1} A^d_{t+j}\right), \qquad i = 0, \dots, n-1. \tag{3}$$

We now formalize the transformation properties of these action representations (see Appendix A for the proof):

Proposition 1 (Equivariance and Invariance under $\mathrm{SE}(3)$)**.** *Consider a global transformation $g \in \mathrm{SE}(3)$ applied to the world coordinate frame, which transforms the current gripper pose as $T_t \mapsto gT_t$. Under this transformation:*

1. *The absolute trajectory action transforms equivariantly, i.e., $g \cdot a \;=\; \{gA_t,\, gA_{t+1}, \dots\}$.*
2. *The relative trajectory action is invariant, i.e., $g \cdot a^r = a^r$.*
3. *The delta trajectory action is invariant, i.e., $g \cdot a^d = a^d$.*

4.2 SE(3)-Invariant Policy Learning

A key advantage of using relative or delta trajectory action representations is that, if the policy being modeled is equivariant, i.e., $\pi(go) = g\pi(o)$, the learned function $\bar{\pi} : o \mapsto a^r$ or $\bar{\pi} : o \mapsto a^d$ becomes invariant, i.e., $\bar{\pi}(go) = \bar{\pi}(o)$. This makes the underlying denoising network ϵ_θ also invariant, significantly reducing the function space and potentially easing the training.

Moreover, when the relative or delta trajectory is combined with eye-in-hand perception, it naturally yields an $\mathrm{SE}(3)$-invariant canonicalization. Specifically, consider an agent equipped with a gripper-mounted camera, which captures an eye-in-hand image I_t. When an arbitrary transformation $g \in \mathrm{SE}(3)$ is applied to the world, the visual input from the eye-in-hand camera remains unchanged since the relative position and orientation between the camera and the world do not vary. For the input observation $o_t = (I_t, T_t, w_t)$ consisting of the invariant image I_t, a gripper pose T_t, and gripper open-width w_t, applying the transformation g to the world frame affects only the gripper pose T_t,

$$g \cdot o_t = (I_t, gT_t, w_t). \tag{4}$$

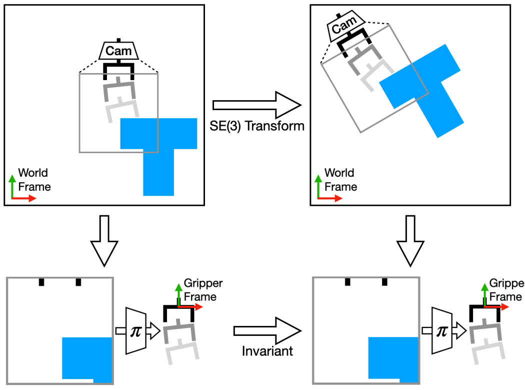

Figure 3: The invariant property of an eye-in-hand perception and action representation in policy learning. Top: an $\mathrm{SE}(3)$ transform applied to the world, the action should transform accordingly. Bottom: for the policy, both the input eye-in-hand image and the output relative trajectory (in the gripper frame) remain invariant.

Let us first assume that the policy does not depend on the gripper pose T_t, then the policy $\pi\colon o \mapsto a$ is $\mathrm{SE}(3)$-equivariant when using eye-in-hand perception and relative or delta action:

Proposition 2. *Let $\bar{\pi} : o \mapsto a^r$ or $\bar{\pi} : o \mapsto a^d$ be a function mapping from the observation o to the relative trajectory a^r or the delta trajectory a^d. Assume an eye-in-hand observation is used where Equation 4 is satisfied and $\bar{\pi}$ does not depend on T_t. If the policy $\pi : o \mapsto a$ reconstructs the absolute trajectory a using Equation 2 or 3, then π is $\mathrm{SE}(3)$-equivariant, i.e., $\pi(go) = g\pi(o)$.*

See Appendix B for the proof. This equivariance property implies that once a policy is learned, it automatically generalizes across different poses in space without additional training data, thus enhancing sample efficiency and robustness. As shown in Figure 3, when an $\mathrm{SE}(3)$ transformation is applied to the world, both the perception and action remain invariant.

In practice, the assumption that $\bar{\pi}$ does not depend on T_t will not perfectly hold because the changing T_t will affect the network prediction, thus we will only have an approximate invariance property. Since T_t provides important information to the policy despite breaking the symmetry, we choose not to explicitly constrain the policy to be invariant to T_t. Still, our experiments demonstrate significant performance improvements when employing this approximate invariant property.

4.3 Equivariant Vision Encoders

While the invariant representations described in Section 4.2 theoretically achieve $\mathrm{SE}(3)$-equivariant policy learning, we experimentally found that they do not fully match the performance of end-to-end equivariant policies [60]. This is because equivariant neural networks not only guarantee global symmetric transformations, but more importantly, they extract richer local features that can capture the underlying symmetries of the problem domain. Instead of falling back to a network architecture that is end-to-end equivariant, we propose a novel approach that incorporates an equivariant vision encoder to extract symmetry-aware features while preserving a standard non-equivariant diffusion backbone. This approach would preserve the benefits of symmetric feature extraction without the complexity of a full equivariant model.

Specifically, we can replace the standard CNN vision encoder in a Diffusion Policy with an equivariant CNN that operates on the group $C_u \subset \mathrm{SO}(2)$. This encoder maps the input eye-in-hand image to a feature vector that transforms according to the regular representation of C_u, providing a richer representation to the diffusion head with explicit information about how features transform under rotations, significantly enhancing learning.

4.3.1 Incorporating Pretrained Vision Encoders with Frame Averaging

Although there exists a wide variety of equivariant neural network architectures [8, 62, 4], they typically require defining each layer to be equivariant by constraining the weights with specialized kernels [10]. This usually implies that the network is built specifically for a task and is trained from scratch. However, modern computer vision has seen tremendous progress through large-scale pretrained models, which provide powerful general-purpose representations. To bridge the gap between equivariance and pre-training, we employ *Frame Averaging* [48] for turning an arbitrary function $\Phi : X \to Y$ into a G-equivariant network by averaging over a *Frame* $\mathcal{F} : X \to 2^G \setminus \{\emptyset\}$ that satisfies equivariance as a set $\mathcal{F}(gx) = \mathcal{F}(x)$:

$$\Psi(x) = \frac{1}{|\mathcal{F}(x)|} \sum_{g \in \mathcal{F}(x)} \rho_y(g)\Phi(\rho_x(g)^{-1}x), \tag{5}$$

where $\Psi : X \to Y$ will have the equivariant property $\Psi(\rho_x(g)x) = \rho_y(g)\Psi(x)$. For a finite group G, one can set the frame to be the whole group $\mathcal{F}(x) = G$, and Equation 5 becomes *symmetrization*:

$$\Psi(x) = \frac{1}{|G|} \sum_{g \in G} \rho_y(g)\Phi(\rho_x(g)^{-1}x). \tag{6}$$

When using a pretrained encoder with Frame Averaging, we obtain the benefits of both powerful pretrained representations and explicit rotational equivariance, allowing us to leverage state-of-the-art vision backbones without sacrificing symmetry properties.

5 Experiments

In this section, we conduct a systematic experimental study comparing different approaches for incorporating symmetry in diffusion policies. We investigate the following key research questions:

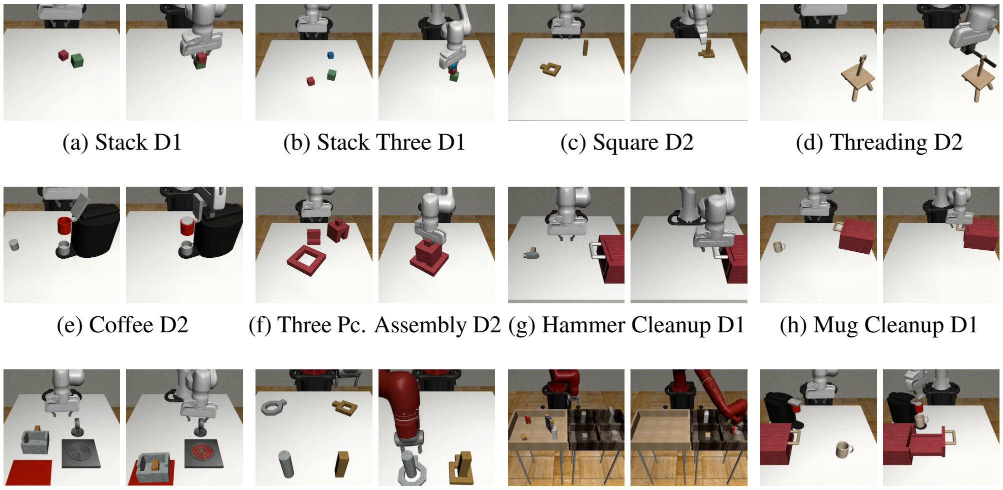
(a) Stack D1
(b) Stack Three D1
(c) Square D2
(d) Threading D2
(e) Coffee D2
(f) Three Pc. Assembly D2
(g) Hammer Cleanup D1
(h) Mug Cleanup D1
(i) Kitchen D1
(j) Nut Assembly D0
(k) Pick Place D0
(l) Coffee Preparation D1

Figure 4: The experimental environments from MimicGen [42]. The left image in each subfigure shows the initial state of the environment; the right image shows the goal state.

1. **Invariant representations:** How do the invariant action and perception representations analyzed in Section 4.1 impact diffusion policy learning performance?
2. **Equivariant vision encoders:** Can diffusion policies benefit from incorporating equivariant vision encoders?
3. **Pre-trained encoders:** How effectively can we leverage pre-trained encoders with Frame Averaging (Equation 5)?
4. **Comparison to end-to-end equivariant diffusion:** How do these approaches compare with fully equivariant diffusion policies [60]?

We evaluate our approaches on 12 robotic manipulation tasks in the MimicGen [42] benchmark, as illustrated in Figure 4. We perform an additional Robomimic [41] experiment in Appendix D.

5.1 Action and Observation Representation

We first evaluate the three action representations (absolute trajectory, relative trajectory, and delta trajectory) discussed in Section 4.1 across two different observation settings, *Large FOV In-Hand* and *In-hand + External*. Figure 2 illustrates the differences between these configurations.

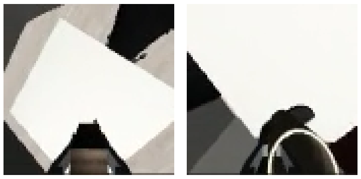
(a) Threading
(b) Coffee Prep.

Figure 5: The failures of the in-hand observation due to occlusion and insufficient information.

The results, presented in Table 1, demonstrate several key findings. First, relative trajectory consistently outperforms absolute trajectory in 11 (for Large FOV In-Hand) or 10 (for In-Hand + External) out of 12 tasks. On average, relative trajectory provides a 5.9% improvement over absolute trajectory with Large FOV In-Hand observations and a 7.4% improvement with In-Hand + External observations. These results align with our theoretical analysis in Section 4.1, confirming that the symmetry properties of relative trajectory representations contribute to better performance. However, despite having the same theoretical guarantee, delta trajectory empirically performs poorly, underperforming absolute trajectory by 2.9% on average, where it only performs well in relatively simple tasks. We hypothesize that this is because delta trajectory can be interpreted as a sequence of velocity vectors, containing less temporal and structural information for the denoising process. Notice that similar observations of the underperformance of velocity control in diffusion policy learning were also reported in prior works [5, 60].

Table 1: The performance comparison between different action representations and different observation setups. All policies are trained using 100 expert demonstrations. We perform 60 evaluations (each with 50 policy rollouts) throughout the training, and report the best average success rate. Results averaged over three seeds, ± indicates standard error. Blue box indicates relative trajectory or delta trajectory outperforming absolute trajectory in the corresponding observation setting; Red box indicates underperforming. **Bold** indicates best performing method across different settings. Performance of Abs Traj in In-Hand + External is reported by [60].

Obs	Action	Mean	Stack D1	Stack Three D1	Square D2	Threading D2	Coffee D2	Three Pc. D2
Large FOV In-Hand	Rel Traj	46.7	**98.0±0.0**	**72.0±1.2**	16.0±1.2	16.0±1.2	56.7±1.8	**4.0±1.2**
	Delta Traj	37.9	96.7±0.7	54.7±3.5	10.0±1.2	11.3±1.3	38.7±0.7	2.0±1.2
	Abs Traj	40.8	94.0±1.2	52.7±2.4	7.3±0.7	10.0±0.0	44.0±4.6	2.0±0.0
In-Hand + External	Rel Traj	49.4	89.3±2.7	52.7±3.7	**20.7±1.3**	**18.7±3.5**	**63.3±1.3**	3.3±0.7
	Abs Traj	42.0	76.0±4.0	38.0±0.0	8.0±1.2	17.3±1.8	44.0±1.2	**4.0±0.0**

Obs	Method		Hammer Cl. D1	Mug Cl. D1	Kitchen D1	Nut Asse. D0	Pick Place D0	Coffee Prep. D1
Large FOV In-Hand	Rel Traj		60.7±0.7	**50.7±1.8**	**68.7±3.7**	44.7±3.3	**42.7±1.6**	30.7±1.8
	Delta Traj		**62.7±2.4**	47.3±1.8	57.3±4.8	17.0±1.5	31.7±0.4	26.0±1.2
	Abs Traj		62.0±2.0	44.7±1.8	66.7±0.7	43.0±2.5	32.0±0.6	30.7±2.9
In-Hand + External	Rel Traj		58.7±0.7	46.7±0.7	62.0±2.0	**57.3±0.9**	39.8±1.2	**80.0±2.0**
	Abs Traj		52.0±1.2	42.7±0.7	66.7±2.4	54.7±2.3	35.3±2.2	65.3±0.7

When comparing across observation settings using relative trajectory, we find that Large FOV In-Hand generally performs better or on par with In-Hand + External. However, if averaged across all tasks, the Large FOV In-Hand setup underperforms by 2.7%. This performance gap is primarily due to a significant drop in the Coffee Preparation task, where the eye-in-hand view alone provides insufficient information for completing this long-horizon task. Moreover, we found that tasks like Threading sometimes encounter occlusion challenges. As shown in Figure 5, those limitations of a single eye-in-hand image constitute the majority of the failure modes. Despite these drawbacks, leveraging the invariant observation and action representations provides a 4.7% improvement compared with the original Diffusion Policy, which uses In-Hand + External views and absolute trajectory.

5.2 Integrating Symmetry into the Vision Encoder

Having established the advantages of invariant action representations, we now investigate different approaches for incorporating symmetry into the vision encoder component of diffusion policies. We compare four methods: *CNN Encoder (CNN Enc):* A standard ResNet-18 [17] without any symmetry constraints, trained from scratch; *Equivariant Encoder (Equi Enc):* An equivariant ResNet-18 architecture implemented with equivariant layers using the escnn [4] library, enforcing C_8-equivariance with outputs in the regular representation of C_8; *Pretrained Encoder (Pretrain):* A standard ResNet-18 pretrained on ImageNet-1k [12], without explicit symmetry constraints; *Pretrained Encoder with Frame Averaging (Pretrain + FA):* A pretrained ResNet-18 enhanced with Frame Averaging (Equation 5) to achieve C_8-equivariance without modifying the underlying network architecture.

Table 2 presents our findings across all 12 manipulation tasks. The results reveal several important insights: First, comparing non-pretrained encoders (Equi Enc vs. CNN Enc), we observe that incorporating equivariance improves performance in 11 out of 12 tasks, yielding a substantial 9.1% average improvement. This confirms that explicit symmetry constraints significantly benefit diffusion policy learning. Second, in the pretrained encoder setting, adding Frame Averaging (Pretrain + FA vs. Pretrain) leads to a 4.1% average performance improvement, with superior results in 7 out of 12 tasks. This demonstrates that symmetry benefits can be obtained even when leveraging powerful pretrained representations. Third, comparing our approaches to Equivariant Diffusion Policy [60] (EquiDiff), we find that both Equi Enc and Pretrain + FA achieve competitive performance.

Specifically, our Equi Enc approach outperforms image-based EquiDiff (Im) on average, while Pretrain + FA achieves results only 2.5% below voxel-based EquiDiff (Vo). This is particularly impressive considering that EquiDiff (Vo) utilizes RGBD inputs from four cameras and employs a substantially more complex architecture. In contrast, our Pretrain + FA approach requires only a single eye-in-hand RGB image and minimal equivariant reasoning, making it considerably more practical for real-world deployment.

Table 2: The performance comparison between symmetric encoder and standard encoder. All policies are trained using 100 expert demonstrations. We perform 60 evaluations (each with 50 policy rollouts) throughout the training, and report the best average success rate. Results averaged over three seeds, ± indicates standard error. Dark green box indicates outperforming EquiDiff with voxel inputs; Light green box indicates outperforming EquiDiff with image inputs; Yellow box indicates underperforming both. **Bold** indicates whether the symmetric encoder outperforms the non-symmetric version in the corresponding training setting. Performance of EquiDiff is reported by [60].

Obs	Action	Method	Mean	Stack D1	Stack Three D1	Square D2	Threading D2	Coffee D2	Three Pc. D2
Large FOV In-Hand	Rel Traj	Pretrain + FA	61.4	**100.0±0.0**	**86.7±1.8**	**43.3±0.7**	16.7±1.8	63.3±0.7	**28.7±1.8**
		Pretrain	57.3	**100.0±0.0**	78.0±2.0	33.3±1.8	**18.0±1.2**	**65.3±2.7**	14.0±0.0
		Equi Enc	55.8	**99.3±0.7**	**75.3±2.4**	**32.0±1.2**	14.0±1.2	**63.3±1.8**	**26.0±2.0**
		CNN Enc	46.7	98.0±0.0	72.0±1.2	16.0±1.2	**16.0±1.2**	56.7±1.8	4.0±1.2
In-Hand + Ext. Voxel	Abs Traj	EquiDiff (Im)	53.7	93.3±0.7	54.7±5.2	25.3±8.7	22.0±1.2	60.0±2.0	15.3±1.8
		EquiDiff (Vo)	63.9	98.7±0.7	74.7±4.4	38.7±1.3	38.7±0.7	64.7±0.7	37.3±2.7

Obs	Action	Method	Hammer Cl. D1	Mug Cl. D1	Kitchen D1	Nut Asse. D0	Pick Place D0	Coffee Prep. D1
Large FOV In-Hand	Rel Traj	Pretrain + FA	**76.0±2.0**	60.0±2.3	78.7±1.8	**74.7±1.5**	**50.8±0.7**	**58.0±2.0**
		Pretrain	**76.0±1.2**	**62.7±1.8**	**80.7±2.4**	61.0±2.1	48.2±1.2	50.0±3.1
		Equi Enc	**67.3±2.9**	**58.7±1.8**	**72.0±3.1**	**69.3±0.9**	**47.2±2.8**	**44.7±2.7**
		CNN Enc	60.7±0.7	50.7±1.8	68.7±3.7	44.7±3.3	42.7±1.6	30.7±1.8
In-Hand + Ext. Voxel	Abs Traj	EquiDiff (Im)	65.3±0.7	49.3±0.7	67.3±0.7	74.0±1.2	41.7±3.2	76.7±0.7
		EquiDiff (Vo)	70.0±2.0	52.7±1.3	85.3±0.7	67.3±0.9	57.7±1.8	80.0±1.2

Overall, these results suggest that integrating symmetry through equivariant encoders provides significant performance benefits for diffusion policies, with Frame Averaging offering an elegant way to leverage powerful pretrained representations while maintaining equivariance properties.

6 Discussion

In this paper, we present a practical guide for incorporating symmetry in diffusion policies, achieving performance competitive with or exceeding fully equivariant architectures while requiring significantly less implementation complexity. Notably, our method performs only 2.5% below voxel-based EquiDiff, despite using only a single eye-in-hand RGB image compared to EquiDiff's four RGBD cameras. Our approach not only defines a new state-of-the-art performance for RGB eye-in-hand diffusion policy, but more importantly, it addresses the trade-off between architectural complexity and sample efficiency when introducing symmetries into policy learning.

Concretely, we investigate three straightforward approaches for incorporating symmetry: invariant representations through relative trajectory and eye-in-hand perception, integrating equivariant vision encoders, and using Frame Averaging with pretrained encoders. Our extensive experimental evaluation across 12 manipulation tasks in MimicGen yields several important findings. First, we demonstrate that relative trajectory actions consistently outperform absolute trajectory, confirming our theoretical analysis that relative trajectory induces $\mathrm{SE}(3)$-invariance. This finding is particularly valuable because a simple coordinate frame change in action representation can bring a 5-7% improvement. Second, we found that incorporating symmetry through equivariant vision encoders significantly enhances performance by 9.1%, highlighting the value of symmetry-aware features while avoiding complex end-to-end reasoning. Lastly, we show that Frame Averaging provides an elegant solution for leveraging the power of pre-trained vision encoders while maintaining equivariance.

6.1 Limitations

There are several limitations of this work that suggest directions for future research. First, only leveraging an eye-in-hand image assumes a good coverage of the entire workspace (thus we use an enlarged FOV in our experiments); however, as shown in Figure 5, the limited view still constitutes the most significant failure mode of our system. In future works, this could be addressed by using a fish-eye camera [7], or a memory mechanism to maintain context across timesteps. Second, while our approaches are theoretically applicable to other policy learning frameworks beyond diffusion models, such as ACT [72], we limited our investigation to diffusion policies and only experimented in the MimicGen [42] and Robomimic [41] benchmarks. Third, leveraging an equivariant encoder, especially with Frame Averaging, could be computationally expensive. Our method roughly takes twice the GPU hours to train compared with the original Diffusion Policy, but is twice as fast as EquiDiff (Im). Finally, although our method is well-suited for real-world deployment on systems like UMI [7] , we have not yet demonstrated this transfer to physical robots.

Acknowledgment

This work was supported in part by NSF grants 2107256, 2134178, 2314182, 2409351, 2442658, and NASA grant 80NSSC19K1474.

References

[1] Jorge Aldaco, Travis Armstrong, Robert Baruch, Jeff Bingham, Sanky Chan, Kenneth Draper, Debidatta Dwibedi, Chelsea Finn, Pete Florence, Spencer Goodrich, et al. Aloha 2: An enhanced low-cost hardware for bimanual teleoperation. *arXiv preprint arXiv:2405.02292*, 2024.

[2] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

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

[4] Gabriele Cesa, Leon Lang, and Maurice Weiler. A Program to Build $E(n)$-Equivariant Steerable CNNs. In *International Conference on Learning Representations*, 2021.

[5] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[6] 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.

[7] 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. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[8] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016.

[9] Taco S. Cohen and Max Welling. Steerable CNNs. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=rJQKYt5ll.

[10] Taco S Cohen, Mario Geiger, and Maurice Weiler. A general theory of equivariant cnns on homogeneous spaces. *Advances in neural information processing systems*, 32, 2019.

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

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[13] Ben Eisner, Yi Yang, Todor Davchev, Mel Vecerik, Jonathan Scholz, and David Held. Deep SE(3)-equivariant geometric reasoning for precise placement tasks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2inBuwTyL2.

[14] 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. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=eJHy0AF5TO.

[15] Mario Geiger, Tess Smidt, Alby M., Benjamin Kurt Miller, Wouter Boomsma, Bradley Dice, Kostiantyn Lapchevskyi, Maurice Weiler, Michał Tyszkiewicz, Simon Batzner, Dylan Madisetti, Martin Uhrin, Jes Frellsen, Nuri Jung, Sophia Sanborn, Mingjian Wen, Josh Rackers, Marcel Rød, and Michael Bailey. Euclidean neural networks: e3nn, April 2022. URL https://doi.org/10.5281/zenodo.6459381.

[16] 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.

[17] 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*, pages 770–778, 2016.

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[19] Tai Hoang, Huy Le, Philipp Becker, Vien Anh Ngo, and Gerhard Neumann. Geometry-aware RL for manipulation of varying shapes and deformable objects. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7BLXhmWvwF.

[20] Kyle Hsu, Moo Jin Kim, Rafael Rafailov, Jiajun Wu, and Chelsea Finn. Vision-based manipulators need to also see from their hands. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RJkAHKp7kNZ.

[21] Boce Hu, Xupeng Zhu, Dian Wang, Zihao Dong, Haojie Huang, Chenghao Wang, Robin Walters, and Robert Platt. Orbitgrasp: Se (3)-equivariant grasp learning. In *8th Annual Conference on Robot Learning*, 2024.

[22] Haojie Huang, Dian Wang, Robin Walters, and Robert Platt. Equivariant Transporter Network. In *Robotics: Science and Systems*, 2022.

[23] Haojie Huang, Dian Wang, Arsh Tangri, Robin Walters, and Robert Platt. Leveraging Symmetries in Pick and Place. *The International Journal of Robotics Research*, 2023.

[24] Haojie Huang, Dian Wang, Xupeng Zhu, Robin Walters, and Robert Platt. Edge Grasp Network: A Graph-Based $\mathrm{SE}(3)$-invariant Approach to Grasp Detection. In *International Conference on Robotics and Automation (ICRA)*, 2023.

[25] Haojie Huang, Owen Lewis Howell, Dian Wang, Xupeng Zhu, Robert Platt, and Robin Walters. Fourier transporter: Bi-equivariant robotic manipulation in 3d. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=UulwvAU1W0.

[26] Haojie Huang, Haotian Liu, Dian Wang, Robin Walters, and Robert Platt. Match policy: A simple pipeline from point cloud registration to manipulation policies. *arXiv preprint arXiv:2409.15517*, 2024.

[27] Haojie Huang, Karl Schmeckpeper, Dian Wang, Ondrej Biza, Yaoyao Qian, Haotian Liu, Mingxi Jia, Robert Platt, and Robin Walters. IMAGINATION POLICY: Using generative point cloud models for learning manipulation policies. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=56IzghzjfZ.

[28] 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.

[29] 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.

[30] Mingxi Jia, Dian Wang, Guanang Su, David Klee, Xupeng Zhu, Robin Walters, and Robert Platt. SEIL: Simulation-augmented Equivariant Imitation Learning. In *International Conference on Robotics and Automation (ICRA)*, 2023.

[31] Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. In *8th Annual Conference on Robot Learning*, 2024. URL `https://openreview.net/forum?id=gqCQxObVz2`.

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

[33] Zhixuan Liang, Yao Mu, Mingyu Ding, Fei Ni, Masayoshi Tomizuka, and Ping Luo. AdaptDiffuser: Diffusion Models as Adaptive Self-evolving Planners. In *International Conference on Machine Learning*, 2023.

[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] Yi-Lun Liao, Brandon M Wood, Abhishek Das, and Tess Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=mCOBKZmrzD`.

[36] Byeongdo Lim, Jongmin Kim, Jihwan Kim, Yonghyeon Lee, and Frank C. Park. Equigraspflow: SE(3)-equivariant 6-dof grasp pose generative flows. In *8th Annual Conference on Robot Learning*, 2024. URL `https://openreview.net/forum?id=5lSkn5v4LK`.

[37] Shiqi Liu, Mengdi Xu, Peide Huang, Xilun Zhang, Yongkang Liu, Kentaro Oguchi, and Ding Zhao. Continual Vision-based Reinforcement Learning with Group Symmetries. In *Conference on Robot Learning*, pages 222–240. PMLR, 2023.

[38] 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. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=yAzN4tz7oI`.

[39] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2018.

[40] Xiao Ma, Sumit Patidar, Iain Haughton, and Stephen James. Hierarchical diffusion policy for kinematics-aware multi-task robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18081–18090, 2024.

[41] 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. In Aleksandra Faust, David Hsu, and Gerhard Neumann, editors, *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 1678–1690. PMLR, 08–11 Nov 2022.

[42] 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. In *7th Annual Conference on Robot Learning*, 2023.

[43] Hai Nguyen, Tadashi Kozuno, Cristian C Beltran-Hernandez, and Masashi Hamaya. Symmetry-aware reinforcement learning for robotic assembly under partial observability with a soft wrist. *arXiv preprint arXiv:2402.18002*, 2024.

[44] Hai Huu Nguyen, Andrea Baisero, David Klee, Dian Wang, Robert Platt, and Christopher Amato. Equivariant reinforcement learning under partial observability. In *Conference on Robot Learning*, pages 3309–3320. PMLR, 2023.

[45] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Charles Xu, Jianlan Luo, Tobias Kreiman, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

[46] Chuer Pan, Brian Okorn, Harry Zhang, Ben Eisner, and David Held. TAX-Pose: Task-Specific Cross-Pose Estimation for Robot Manipulation. In *Conference on Robot Learning*, pages 1783–1792. PMLR, 2023.

[47] Jyothish Pari, Nur Muhammad Shafiullah, Sridhar Pandian Arunachalam, and Lerrel Pinto. The surprising effectiveness of representation learning for visual imitation. *arXiv preprint arXiv:2112.01511*, 2021.

[48] Omri Puny, Matan Atzmon, Edward J. Smith, Ishan Misra, Aditya Grover, Heli Ben-Hamu, and Yaron Lipman. Frame averaging for invariant and equivariant network design. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=zIUyj55nXR.

[49] Allen Z. Ren, Justin Lidard, Lars Lien Ankile, Anthony Simeonov, Pulkit Agrawal, Anirudha Majumdar, Benjamin Burchfiel, Hongkai Dai, and Max Simchowitz. Diffusion policy policy optimization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=mEpqHvbD2h.

[50] Hyunwoo Ryu, Hong in Lee, Jeong-Hoon Lee, and Jongeun Choi. Equivariant Descriptor Fields: SE(3)-Equivariant Energy-Based Models for End-to-End Visual Robotic Manipulation Learning. In *The Eleventh International Conference on Learning Representations*, 2023.

[51] Hyunwoo Ryu, Jiwoo Kim, Junwoo Chang, Hyun Seok Ahn, Joohwan Seo, Taehan Kim, Jongeun Choi, and Roberto Horowitz. Diffusion-EDFs: Bi-equivariant Denoising Generative Modeling on SE(3) for Visual Robotic Manipulation. *arXiv preprint arXiv:2309.02685*, 2023.

[52] 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.

[53] Anthony Simeonov, Yilun Du, Yen-Chen Lin, Alberto Rodriguez Garcia, Leslie Pack Kaelbling, Tomás Lozano-Pérez, and Pulkit Agrawal. SE(3)-Equivariant Relational Rearrangement with Neural Descriptor Fields. In *Conference on Robot Learning*, pages 835–846. PMLR, 2023.

[54] Shuran Song, Andy Zeng, Johnny Lee, and Thomas Funkhouser. Grasping in the wild: Learning 6dof closed-loop grasping from low-cost demonstrations. *IEEE Robotics and Automation Letters*, 5(3):4978–4985, 2020.

[55] Yang Song and Stefano Ermon. Generative Modeling by Estimating Gradients of the Data Distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

[56] 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.

[57] Dian Wang, Robin Walters, Xupeng Zhu, and Robert Platt. Equivariant Q Learning in Spatial Action Spaces. In *5th Annual Conference on Robot Learning*, 2021.

[58] Dian Wang, Mingxi Jia, Xupeng Zhu, Robin Walters, and Robert Platt. On-Robot Learning With Equivariant Models. In *6th Annual Conference on Robot Learning*, 2022.

[59] Dian Wang, Robin Walters, and Robert Platt. $\mathrm{SO}(2)$-Equivariant Reinforcement Learning. In *International Conference on Learning Representations*, 2022.

[60] Dian Wang, Stephen Hart, David Surovik, Tarik Kelestemur, Haojie Huang, Haibo Zhao, Mark Yeatman, Jiuguang Wang, Robin Walters, and Robert Platt. Equivariant diffusion policy. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=wD2kUVLT1g.

[61] Zhendong Wang, Jonathan J Hunt, and Mingyuan Zhou. Diffusion Policies as an Expressive Policy Class for Offline Reinforcement Learning. In *The Eleventh International Conference on Learning Representations*, 2022.

[62] Maurice Weiler and Gabriele Cesa. General e (2)-equivariant steerable cnns. *Advances in Neural Information Processing Systems*, 32, 2019.

[63] Ziniu Wu, Tianyu Wang, Chuyue Guan, Zhongjie Jia, Shuai Liang, Haoming Song, Delin Qu, Dong Wang, Zhigang Wang, Nieqing Cao, et al. Fast-umi: A scalable and hardware-independent universal manipulation interface. *arXiv preprint arXiv:2409.19499*, 2024.

[64] Zhou Xian, Nikolaos Gkanatsios, Theophile Gervet, Tsung-Wei Ke, and Katerina Fragkiadaki. ChainedDiffuser: Unifying Trajectory Diffusion and Keypose Prediction for Robotic Manipulation. In *7th Annual Conference on Robot Learning*, 2023.

[65] Jingyun Yang, Ziang Cao, Congyue Deng, Rika Antonova, Shuran Song, and Jeannette Bohg. Equibot: SIM(3)-equivariant diffusion policy for generalizable and data efficient learning. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=ueBmGhLOXP.

[66] Jingyun Yang, Congyue Deng, Jimmy Wu, Rika Antonova, Leonidas Guibas, and Jeannette Bohg. Equivact: Sim (3)-equivariant visuomotor policies beyond rigid object manipulation. In *2024 IEEE international conference on robotics and automation (ICRA)*, pages 9249–9255. IEEE, 2024.

[67] Sarah Young, Dhiraj Gandhi, Shubham Tulsiani, Abhinav Gupta, Pieter Abbeel, and Lerrel Pinto. Visual imitation made easy. In *Conference on Robot learning*, pages 1992–2005. PMLR, 2021.

[68] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In *Robotics: Science and Systems*, 2024.

[69] Andy Zeng, Shuran Song, Stefan Welker, Johnny Lee, Alberto Rodriguez, and Thomas Funkhouser. Learning synergies between pushing and grasping with self-supervised deep reinforcement learning. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4238–4245. IEEE, 2018.

[70] Andy Zeng, Pete Florence, Jonathan Tompson, Stefan Welker, Jonathan Chien, Maria Attarian, Travis Armstrong, Ivan Krasin, Dan Duong, Vikas Sindhwani, et al. Transporter networks: Rearranging the visual world for robotic manipulation. *arXiv preprint arXiv:2010.14406*, 2020.

[71] 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.

[72] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[73] Tony Z. Zhao, Jonathan Tompson, Danny Driess, Pete Florence, Seyed Kamyar Seyed Ghasemipour, Chelsea Finn, and Ayzaan Wahid. ALOHA unleashed: A simple recipe for robot dexterity. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=gvdXE7ikHI.

[74] Xupeng Zhu, Dian Wang, Ondrej Biza, Guanang Su, Robin Walters, and Robert Platt. Sample Efficient Grasp Learning Using Equivariant Models. In *Robotics: Science and Systems*, 2022.

[75] Xupeng Zhu, Dian Wang, Guanang Su, Ondrej Biza, Robin Walters, and Robert Platt. On robot grasp learning using equivariant models. *Autonomous Robots*, 47(8):1175–1193, 2023.

A Proof of Proposition 1

Proof. 1. Absolute trajectory equivariance: Given an absolute trajectory action $a = \{A_{t+i}\}_{i=0}^{n-1}$, each pose A_{t+i} is defined in the world frame. Under the transformation g, each pose transforms as $gA_{t+i}, i = 0, \ldots, n-1$. Thus, by definition, the absolute trajectory transforms equivariantly:

$$g \cdot a \ = \ \{gA_t,\, gA_{t+1}, \ldots\}$$

2. Relative trajectory invariance: Consider a relative trajectory action $a^r = \{A^r_{t+i}\}_{i=0}^{n-1}$ defined in the local gripper frame at the initial time step t. The corresponding absolute poses are obtained as $A_{t+i} = T_t A^r_{t+i}$. Under the global transform g, the absolute pose becomes $g \cdot A_{t+i} = gT_t A^r_{t+i}$. Since the relative pose A^r_{t+i} appears as a right multiplication factor, it remains unchanged under the global transform. Hence, we have invariance:

$$g \cdot a^r = a^r$$

3. Delta trajectory invariance: For a delta trajectory action $a^d = \{A^d_{t+i}\}_{i=0}^{n-1}$, each incremental transform A^d_{t+i} is expressed in the gripper's local frame at time $t + i - 1$. The absolute pose reconstruction is given by $A_{t+i} = T_t \prod_{j=0}^{i-1} A^d_{t+j}$. Under the global transform g, we have $g \cdot A_{t+i} = gT_t \prod_{j=0}^{i-1} A^d_{t+j}$, where each incremental transform A^d_{t+j} is multiplied on the right and thus remains unaffected by the global transformation g. Therefore, the delta trajectory action is invariant:

$$g \cdot a^d = a^d$$

□

B Proof of Proposition 2

Proof. We treat the two cases in parallel. In both, the policy

$$\pi(o_t) \ = \ \big\{A_{t+i}\big\}_{i=0}^{n-1}$$

outputs a sequence of absolute poses $A_{t+i} \in \mathrm{SE}(3)$. Internally it first predicts a "local" sequence $\bar{\pi}(o_t) = \{A^r_{t+i}\}_{i=0}^{n-1}$ or $\bar{\pi}(o_t) = \{A^d_{t+i}\}_{i=0}^{n-1}$ (either relative or delta) and then reconstructs absolute poses by anchoring to the current gripper pose T_t.

Case 1: Relative trajectories. By Definition 2,

$$A_{t+i} \ = \ T_t\, A^r_{t+i}, \quad i = 0, \ldots, n-1,$$

where A^r_{t+i} is the ith pose in the relative sequence $\bar{\pi}(o_t) = \{A^r_{t+i}\}$. Thus

$$\pi(o_t) \ = \ \big\{\, T_t\, A^r_{t+i}\big\}_{i=0}^{n-1}.$$

By assumption, $\bar{\pi}$ does not depend on the gripper pose T_t explicitly. Thus, applying the transformation g to the observation has no effect on the relative trajectory prediction:

$$\bar{\pi}(g \cdot o_t) = \bar{\pi}(I_t, gT_t, w_t) = \bar{\pi}(I_t, T_t, w_t) = \bar{\pi}(o_t). \tag{7}$$

Reconstructing absolute poses from the transformed observation gives, for each i,

$$\big[\pi(g \cdot o_t)\big]_i = (gT_t)\, A^r_{t+i} = g\big(T_t\, A^r_{t+i}\big) = g \cdot A_{t+i} = g \cdot \big[\pi(o_t)\big]_i$$

Because this holds for all $i = 0, \ldots, n-1$, we conclude

$$\pi\big(g \cdot o_t\big) = g \cdot \pi(o_t).$$

Case 2: Delta trajectories. By Definition, the delta-reconstruction is

$$A_{t+i} \ = \ T_t\, \Big(\prod_{j=0}^{i-1} A^d_{t+j}\Big), \quad i = 0, \ldots, n-1,$$

where $\{A^d_{t+j}\}$ is the delta-sequence from $\bar{\pi}(o_t)$. Again invariance of $\bar{\pi}$ under g gives $\bar{\pi}(g \cdot o_t) = \bar{\pi}(o_t)$, so

$$\big[\pi(g \cdot o_t)\big]_i = (gT_t)\, \Big(\prod_{j=0}^{i-1} A^d_{t+j}\Big) = g\big[T_t\, (\prod_{j=0}^{i-1} A^d_{t+j})\big] = g \cdot A_{t+i} = g \cdot \big[\pi(o_t)\big]_i$$

Hence once again

$$\pi\big(g \cdot o_t\big) = g \cdot \pi(o_t).$$

In both cases we have shown that the entire trajectory satisfies $\pi(g \cdot o_t) = g \cdot \pi(o_t)$, i.e. π is SE(3)-equivariant. □

				Lift		Can		Square		tool hang
Obs	Action	Method	Mean	100 PH	100 MH	100 PH	100 MH	100 PH	100 MH	100 PH
Large FOV In-Hand	Rel Traj	Ours	**93.4**	**100.0±0.0**	**100.0±0.0**	99.3±0.7	**100.0±0.0**	**88.0±2.3**	**92.0±0.0**	74.7±2.4
Voxel	Abs Traj	EquiDiff	90.4	**100.0±0.0**	**100.0±0.0**	99.3±0.7	96.7±0.7	84.0±1.2	76.7±1.3	**76.0±0.0**
In-Hand + External	Abs Traj	DiffPo	87.9	**100.0±0.0**	**100.0±0.0**	**100.0±0.0**	95.3±0.7	85.3±0.7	70.7±0.7	64.0±5.8

Table 3: The performance of our method compared with the baselines in Robomimic. We experiment with 100 Proficient-Human (PH) or Multi-Human (MH) demos in each environment. Results averaged over three seeds. ± indicates standard error.

C Training Detail

We follow the training setup and hyper-parameters of the prior works [5, 60, 7]. Specifically, our RGB observation has a size of $3 \times 84 \times 84$ (which will be random cropped to $3 \times 76 \times 76$ during training), and all tasks have a full 6 DoF SE(3) action space. The observation contains two steps of history observation, and the output of the denoising process is a sequence of 16 action steps. We use all 16 steps for training but only execute eight steps in evaluation. In all pretraining encoder variations, we use two steps of proprioceptive observation but only one step of visual observation, following Chi et al. [7]. The vision encoder's output dimension is 64 for for CNN Enc (following [5]), 128×8 for Equi Enc (128 channel regular representation of C_8, following [60]), and 512 for Pretrain (following [7]). The diffusion UNet has [512, 1024, 2048] hidden channels for end-to-end training variations (following Chi et al. [5]), and [256, 512, 1024] hidden channels for pretraining encoder variations (following Chi et al. [7]). We train our models with the AdamW [39] optimizer (with a learning rate of 10^{-4} and weight decay of 10^{-6}) and Exponential Moving Average (EMA). We use a cosine learning rate scheduler with 500 warm-up steps. We use DDPM [18] with 100 denoising steps for both training and evaluation. We perform training for 600 epochs, and evaluate the method every 10 episodes (60 evaluations in total). All trainings are performed on a single GPU, where we perform training on internal clusters and desktops with different GPU models. Each training of the Pretrain + FA method takes from 3 hours (Stack D1) to 24 hours (Pick Place D0), due to the different sizes of the dataset. The total amount of compute used in this project is roughly 3000 GPU hours.

D Robomimic Experiment

In this section, we perform an experiment in the Robomimic [41] environments. We compare our Pretrain + FA method against EquiDiff [60] and the vanilla Diffusion Policy [5]. As shown in Table 3, our method generally outperforms the baselines, achieving an average improvement of 3% compared with EquiDiff.

E Pretraining and Frame Averaging with External View

In this experiment, we extend our analysis from Section 5.2 to the external view setting to verify whether our findings generalize across different observation configurations. Specifically, we perform an additional experiment on using pretrained encoders (with and without Frame Averaging) with In-Hand + External view. Similar to Section 5.2, we consider three methods in each view setting: 1) *No Pretrain*: A standard ResNet-18 [17] trained from scratch; 2) *Pretrain*: A standard ResNet-18 pretrained on ImageNet-1k [12], without explicit symmetry constraints; 3) *Pretrain + FA*: A pretrained ResNet-18 enhanced with Frame Averaging (Equation 5) to achieve C_8-equivariance without modifying the underlying network architecture.

As shown in Table 4, the benefits of Frame Averaging remain consistent across both observation settings. In the In-Hand + External view setting, Pretrain + FA yields a 6.7% improvement compared with not using Frame Averaging (Pretrain), and a 15.5% improvement compared with training from scratch (No Pretrain). Notably, Pretrain + FA in In-Hand + External view outperforms EquiDiff (Im) in all tasks, and even achieves a 1% higher average performance compared with EquiDiff (Vo). This is particularly impressive considering the additional complexity of EquiDiff, as discussed in Section 5.2. Comparing Pretrain + FA across different views, despite using an additional camera, In-Hand + External only outperforms Large FOV In-Hand by 2.5%. This finding verifies our analysis in Section 4.2, and suggests Large FOV In-Hand is preferable in many applications due to its simplicity

Table 4: The performance comparison of pretrained encoder with Frame Averaging, pretrained encoder, and no pretraining. All policies are trained using 100 expert demonstrations. We perform 60 evaluations (each with 50 policy rollouts) throughout the training, and report the best average success rate. Results averaged over three seeds, ± indicates standard error. Dark green box indicates outperforming EquiDiff with voxel inputs; Light green box indicates outperforming EquiDiff with image inputs; Yellow box indicates underperforming both. **Bold** indicates whether the symmetric encoder outperforms the non-symmetric version in the corresponding training setting. Performance of EquiDiff is reported by [60].

Obs	Action	Method	Mean	Stack D1	Stack Three D1	Square D2	Threading D2	Coffee D2	Three Pc. D2
In-Hand + Ext.	Rel Traj	Pretrain + FA	64.9	**100.0±0.0**	**80.7±0.7**	**42.7±1.3**	**28.0±1.2**	**72.7±3.3**	**32.0±2.3**
		Pretrain	58.2	**100.0±0.0**	72.7±2.4	33.3±1.3	22.0±2.0	63.3±0.7	18.0±2.0
		No Pretrain	49.4	89.3±2.7	52.7±3.7	20.7±1.3	18.7±3.5	63.3±1.3	3.3±0.7
Large FOV In-Hand	Rel Traj	Pretrain + FA	61.4	**100.0±0.0**	**86.7±1.8**	**43.3±0.7**	16.7±1.8	63.3±0.7	**28.7±1.8**
		Pretrain	57.3	**100.0±0.0**	78.0±2.0	33.3±1.8	**18.0±1.2**	**65.3±2.7**	14.0±0.0
		No Pretrain	46.7	98.0±0.0	72.0±1.2	16.0±1.2	16.0±1.2	56.7±1.8	4.0±1.2
In-Hand + Ext. Voxel	Abs Traj	EquiDiff (Im)	53.7	93.3±0.7	54.7±5.2	25.3±8.7	22.0±1.2	60.0±2.0	15.3±1.8
		EquiDiff (Vo)	63.9	98.7±0.7	74.7±4.4	38.7±1.3	38.7±0.7	64.7±0.7	37.3±2.7

Obs	Action	Method	Hammer Cl. D1	Mug Cl. D1	Kitchen D1	Nut Asse. D0	Pick Place D0	Coffee Prep. D1
In-Hand + Ext.	Rel Traj	Pretrain + FA	**67.3±1.8**	**55.3±1.3**	**82.7±0.7**	**78.0±1.5**	**61.8±0.6**	78.0±1.2
		Pretrain	61.3±0.7	54.0±1.2	76.0±1.2	72.3±1.7	55.5±3.8	69.3±1.3
		No Pretrain	58.7±0.7	46.7±0.7	62.0±2.0	57.3±0.9	39.8±1.2	**80.0±2.0**
Large FOV In-Hand	Rel Traj	Pretrain + FA	**76.0±2.0**	60.0±2.3	78.7±1.8	**74.7±1.5**	**50.8±0.7**	**58.0±2.0**
		Pretrain	**76.0±1.2**	**62.7±1.8**	**80.7±2.4**	61.0±2.1	48.2±1.2	50.0±3.1
		No Pretrain	60.7±0.7	50.7±1.8	68.7±3.7	44.7±3.3	42.7±1.6	30.7±1.8
In-Hand + Ext. Voxel	Abs Traj	EquiDiff (Im)	65.3±0.7	49.3±0.7	67.3±0.7	74.0±1.2	41.7±3.2	76.7±0.7
		EquiDiff (Vo)	70.0±2.0	52.7±1.3	85.3±0.7	67.3±0.9	57.7±1.8	80.0±1.2

Table 5: Validating the Pretrain+FA encoder in absolute trajectory and multi-camera observations.

Obs	Action	Method	Mean	Stack Three D1	Square D2	Coffee D2	Nut Assembly D0
In-Hand + Ext.	Abs Traj	DP	36.2	38.0±0.0	8.0±1.2	44.0±1.2	54.7±2.3
		DP + Pretrain + FA	57.7	58.7±1.2	38.0±9.2	56.0±2.0	78.0±2.6
		EquiDiff	53.5	54.7±5.2	25.3±8.7	60.0±2.0	74.0±1.2
		EquiDiff + Pretrain + FA	63.2	68.7±2.3	29.3±3.0	78.0±2.0	76.7±3.2

(requiring only a single eye-in-hand camera) and easier transferability to diverse robotic platforms beyond tabletop manipulation scenarios.

F Pretraining and Frame Averaging with Absolute Action and Multi-Camera Observation

To demonstrate generality beyond the in-hand setup, we evaluate the proposed Pretrain+FA encoder under the same observation/action setting as Diffusion Policy and EquiDiff (in-hand + external views and absolute actions) in four different environments. As shown in Table 5, employing the Pretrain+FA encoder yields a significant 21.5% and 9.7% improvement for Diffusion Policy and EquiDiff, respectively. These results confirm the encoder's plug-and-play nature across observation and action parameterizations.

G Ablation Study

G.1 Isolating Relative Trajectory and Symmetric Feature Extraction

We explicitly isolate each component in our design. As shown in Table 6, starting from the full model, replacing the relative trajectory with absolute trajectory reduces performance by 11.1% across four tasks, while replacing the encoder with a non-pretrained CNN reduces performance more significantly by 19.6%. This result highlights the complementary benefits from symmetry-aware features and invariant action parameterization.

Table 6: Ablation isolating the contributions of (i) relative trajectory and (ii) symmetric feature extraction (Pretrain+FA).

Obs	Rel Traj	Pretrain + FA	Mean	Stack Three D1	Square D2	Coffee D2	Nut Asse. D0
Large FOV In-Hand	✓	✓	67.0	86.7±1.8	43.3±0.7	63.3±0.7	74.7±1.5
	✗	✓	55.9	68.7±2.3	26.7±11.5	60.0±0.0	68.0±3.0
	✓	✗	47.4	72.0±1.2	16.0±1.2	56.7±1.8	44.7±3.3

G.2 Ablating Proprioception

We test Proposition 2's assumption by removing the gripper pose from the policy input in Table 7. As expected, this tighter symmetry assumption reduces performance—indicating that proprioception, while symmetry-breaking, provides valuable context.

Table 7: Ablating proprioception input in the policy

	Stack Three D1	Square D2
With proprioception	72.0	16.0
No proprioception	64.7	14.7

G.3 Ablating Symmetry Group

We vary the SO(2) discretization used by the equivariant encoder. As shown in Table 8, reducing the cyclic group order degrades performance (C8 > C4 > C2), aligning with prior observations [62, 58] that C8 is a practical sweet spot for 2D rotational symmetry.

Table 8: Effect of SO(2) discretization.

Group	Stack Three D1	Square D2
C8	75.3	32.0
C4	68.0	30.7
C2	63.3	20.7

H EquiDiff with Large FOV In-Hand Only

In this experiment, we evaluate EquiDiff with only a large-FOV eye-in-hand camera setting. The comparison is shown in Table 9. Compared to the original external-camera configuration, EquiDiff's mean success drops by 12.8%, underscoring the benefit EquiDiff derives from external viewpoint signals. This result also justifies our choice of using the original EquiDiff observation setting in our main results.

Table 9: Performance of EquiDiff under a large FOV in-hand-only camera setting vs. the original in-hand + external and voxel (multi-RGBD) settings. Removing external cameras substantially reduces EquiDiff performance on several tasks.

Obs	Action	Method	Mean	Stack D1	Stack Three D1	Square D2	Threading D2	Coffee D2	Three Pc. D2
Large FOV In-Hand		EquiDiff (Im)	40.9	96.0±0.0	61.3±5.0	8.7±1.2	13.3±2.3	47.3±3.1	3.3±1.2
In-Hand + Ext.	Abs Traj	EquiDiff (Im)	53.7	93.3±0.7	54.7±5.2	25.3±8.7	22.0±1.2	60.0±2.0	15.3±1.8
Voxel		EquiDiff (Vo)	63.9	98.7±0.7	74.7±4.4	38.7±1.3	38.7±0.7	64.7±0.7	37.3±2.7

Obs	Action	Method	Hammer Cl. D1	Mug Cl. D1	Kitchen D1	Nut Asse. D0	Pick Place D0	Coffee Pre. D1
Large FOV In-Hand		EquiDiff (Im)	59.3±4.2	50.7±2.3	55.3±1.2	40.0±4.0	27.7±2.9	27.3±1.2
In-Hand + Ext.	Abs Traj	EquiDiff (Im)	65.3±0.7	49.3±0.7	67.3±0.7	74.0±1.2	41.7±3.2	76.7±0.7
Voxel		EquiDiff (Vo)	70.0±2.0	52.7±1.3	85.3±0.7	67.3±0.9	57.7±1.8	80.0±1.2

I Broader Impact

This work has a couple of positive and negative social impacts. First, we provide a simple policy learning framework from eye-in-hand images, which could benefit the development of household robots or assistive robots that could be beneficial for society. However, since it is a behavior cloning algorithm and the robots' behavior completely depends on the training data, it could also potentially be used to train robots with harmful behavior. Therefore, it is important for future works to highlight safety monitoring, especially when deploying in real-world environments with human presence.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Our abstract and introduction clearly state the claims of this paper, including the contributions and the assumption that we are using an eye-in-hand camera. The claims made match our theoretical and experimental results.

   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]

   Justification: We include a detailed limitation section in the paper, including the assumption and potential negative impact of using eye-in-hand camera, the limitation of the testing dataset, and the computational efficiency.

   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 the definitions and propositions in the paper are numbered and cross-referenced. All the assumptions are clearly stated.

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 release the code of this paper, including the code for all models and experiments. We also discuss the training detail 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 code for all experiments in the supplementary material, including data generation and all the models. All the results will be reproducible using the code. In the final version of the paper, we will provide a GitHub repository for the code of the project.

   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.
   - 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: We provide the experimental setting in both the paper and the code.

   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 include the standard error of all our experiments, which is calculated from three 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 discuss 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 with the NeurIPS Code of Ethics in every respect.

   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 include a broader impact statement in the appendix, discussing the potential positive and negative societal impacts.

    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**

    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: Our work does not have 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: All the prior works used are properly cited, we will also cite the codebases used in the paper in the GitHub repo.

    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: We will include those in the supplementary material as well as the GitHub repo for the final submission.

    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 paper 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 paper 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. We only use LLM for editing, which is denoted in Openreview.

    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.