Guaranteed $SE(3)$-Equivariant Control via Hand-Centric Behavior Cloning

Julius Jankowski, Pascal Klink, Ingmar Posner, Erhan Gundogdu, Kiru Park, Can Erdogan
Amazon Robotics
jankowju@amazon.com

Abstract: Behavior Cloning is a powerful tool for acquiring skill policies from task demonstrations. However, exploiting spatial symmetries and invariances is important to enable efficient learning and generalization. In this paper, we show that a hand-centric representation of action chunks and proprioceptive observations yield denoising vector fields that are invariant to global rigid transformations for a Flow Matching objective. With this $SE(3)$-invariance of the underlying representation to be learned, policies can be trained on equivariant skills with significantly fewer demonstrations while achieving better generalization without the need for specialized model architectures. We demonstrate the advantage of a hand-centric action representation on a low-tolerance visuomotor manipulation task using a simple multi-layer perceptron for the policy.

Keywords: Equivariant Control, Visuomotor Policy, Behavior Cloning

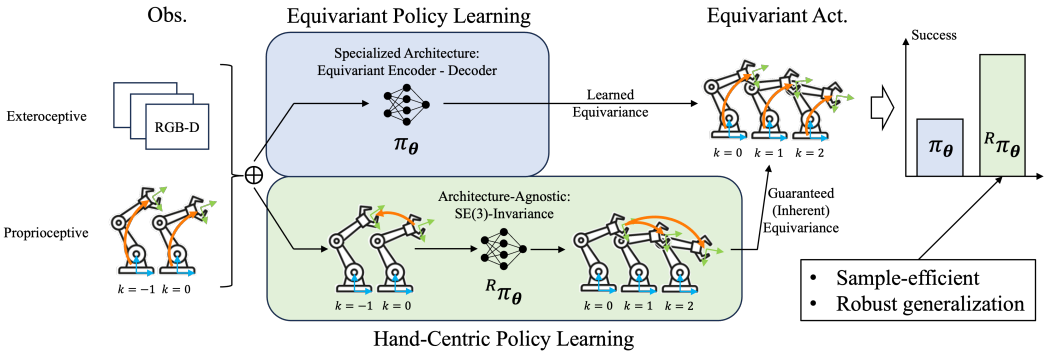

Figure 1: In this paper we show that hand-centric action representations (bottom branch) can be used to train guaranteed equivariant control policies agnostic of the neural network architecture used. This leads to significantly more sample efficient and robust learning when compared to methods that rely on network architectures to encourage equivariant control policies (top branch).

1 Introduction

Many manipulation tasks such as grasping and peg insertion exhibit symmetries that should allow for sample efficient learning and robust generalization [1]. Consequently, a number of recent works in robot learning have focused on learning policies in a way that leverages these symmetries. Typically, these works focus on the use of architectural inductive biases which leverage equivariance in the underlying task but do not guarantee equivariance [2, 3, 4, 5, 6, 7, 8]. At the same time, robotics practitioners have long observed the key influence different action representations play in policy learning [9, 10, 11, 12] with hand-centric representations found to be more amenable for learning in terms of sample efficiency and robustness [13, 14, 15] than, for example, absolute actions.

In this paper, we provide the theoretical underpinning explaining the influence of action representations. In particular, we prove mathematically that actions expressed in a global reference frame exhibit no symmetry with regards to global transformations and that the common relative action representation exhibits translation-invariance but not rotation-invariance. In contrast, we show that hand-centric representations more recently adopted by some works exhibit invariance to global transformations to the scene. Based on these findings we propose a framework leveraging hand-centric action and observation representations to arrive at policies that are provably inherently equivariant to global transformations irrespective of the architecture used. In this framework, absolute actions and observations are first transformed into a hand-centric reference frame, in which they are invariant to global transformations. The policy is then trained in this hand-centric frame. Finally, actions are transformed back to the original reference frame for execution (see Figure 1). We demonstrate the efficacy of our equivariant policy learning framework in both simulated and real-world robot manipulation experiments and show that our approach comprehensively outperforms more traditional action representations as well as established approaches which use architectural biases to leverage task symmetries. In particular, we present the following two contributions:

1. A rigorous analysis of how the selection of action representation, i.e. *absolute*, *relative*, and *hand-centric*, for Cartesian control affects the denoising flow to be learned with respect to rigid global transformations.
2. We introduce Hand-Centric Flow Matching, which builds upon the analysis in contribution 1 and results in an $SE(3)$-invariant denoising flow that inherently generalizes to out-of-distribution global transformations.

2 Background

In this section, we introduce definitions of action representations, $SE(3)$-equivariance, and the Flow Matching objective used for the analysis in Sec. 3.

2.1 Action Representations for Cartesian Control

Policies are typically modeled to directly output absolute or relative robot poses with a fixed world frame as a reference, e.g. the robot's base. This is due to the fact that proprioceptive observations and action data are oftentimes obtained this way. Accordingly, an *absolute action* at future time step k is defined as the pose of the robot's hand R with respect to a fixed world frame W, i.e. $\boldsymbol{a}_k \coloneqq {}^{W}\boldsymbol{T}_{R_k} = [{}^{W}\boldsymbol{R}_{R_k}, {}^{W}\boldsymbol{p}_{R_k}] \in SE(3)$, where time step $k = 0$ denotes the time of the current control step. We furthermore consider *relative actions* that are defined as the delta between two absolute actions, i.e. $\Delta\boldsymbol{a}_k = (\Delta\boldsymbol{p}_k, \Delta\boldsymbol{r}_k) \coloneqq ({}^{W}\boldsymbol{p}_{R_k} - {}^{W}\boldsymbol{p}_{R_0}, \boldsymbol{r}({}^{W}\boldsymbol{R}_{R_0}^{\top}\,{}^{W}\boldsymbol{R}_{R_k}))$. Here, $\boldsymbol{r}({}^{A}\boldsymbol{R}_{B}) \in \mathbb{R}^3$ denotes an axis-angle representation of a rotation vector that rotates frame A into B through the tangent space of SO(3).

Hand-centric actions are computed by changing the frame of reference from a fixed world frame to a frame fixed to the robot's hand. For this, we use the frame of the robot's hand at the time of control, i.e. R_0, such that a *hand-centric action* at time step k is defined by

$$ {}^{R}\boldsymbol{a}_k \coloneqq {}^{R_0}\boldsymbol{T}_{R_k} = {}^{W}\boldsymbol{T}_{R_0}^{-1}\,{}^{W}\boldsymbol{T}_{R_k} = {}^{W}\boldsymbol{T}_{R_0}^{-1}\boldsymbol{a}_k. \tag{1} $$

As a result, the translation ${}^{R_0}\boldsymbol{p}_{R_k}$ and rotation ${}^{R_0}\boldsymbol{R}_{R_k}$ components of a hand-centric action, respectively, can be expressed as functions of translations and rotations in the fixed world frame:

$$ {}^{R_0}\boldsymbol{p}_{R_k} = {}^{W}\boldsymbol{R}_{R_0}^{\top}({}^{W}\boldsymbol{p}_{R_k} - {}^{W}\boldsymbol{p}_{R_0}), \quad {}^{R_0}\boldsymbol{R}_{R_k} = {}^{W}\boldsymbol{R}_{R_0}^{\top}\,{}^{W}\boldsymbol{R}_{R_k}. \tag{2} $$

Figure 2 illustrates the hand-centric representation for an example insertion task. We furthermore define proprioceptive observations analogue to the actions in the corresponding representation. Thus, we define world-centric proprioceptive observations by $\boldsymbol{o}_{-h}^{\mathrm{p}} \coloneqq {}^{W}\boldsymbol{T}_{R_{-h}}$, which are used in combination with absolute and relative actions. Hand-centric proprioceptive observations are given by

$$ {}^{R}\boldsymbol{o}_{-h}^{\mathrm{p}} \coloneqq {}^{R_0}\boldsymbol{T}_{R_{-h}} = {}^{W}\boldsymbol{T}_{R_0}^{-1}\,{}^{W}\boldsymbol{T}_{R_{-h}} = {}^{W}\boldsymbol{T}_{R_0}^{-1}\boldsymbol{o}_{-h}^{\mathrm{p}}. \tag{3} $$

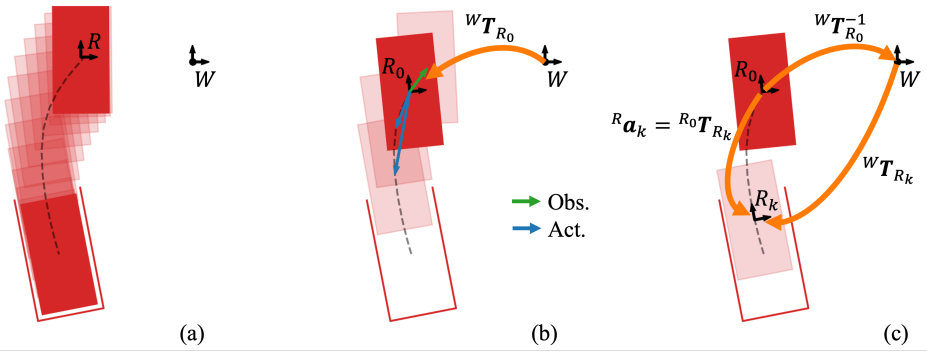

Figure 2: (a) A demonstrated insertion trajectory recorded in a fixed world frame W. (b) Each demonstration is structured in multiple chunks with ${}^{W}\boldsymbol{T}_{R_0}$ corresponding to the pose of the robot (or hand) at the time of control. (c) We propose to learn hand-centric policies, i.e. representing proprioceptive observations and actions with respect to the pose of the robot at the time of control.

2.2 $SE(3)$-Equivariant Policy

Suppose that $\tilde{\boldsymbol{T}} = [\tilde{\boldsymbol{R}}|\tilde{\boldsymbol{p}}] \in SE(3)$ denotes a rigid transformation applied to the pose of the robot's hand and its environment. The operator $\circ$ denotes a composition, such that $\tilde{\boldsymbol{T}} \circ \cdot$ denotes how the corresponding observation or action is transformed when the rigid transformation $\tilde{\boldsymbol{T}}$ is applied to the scene. A Cartesian policy is $SE(3)$-equivariant *iff* the action generated from transformed observations is equivalent to the transformed action generated from the original observations, i.e.

$$\boldsymbol{a}(\tilde{\boldsymbol{T}} \circ \boldsymbol{o}^{\mathrm{p}}, \tilde{\boldsymbol{T}} \circ \boldsymbol{o}^{\mathrm{v}}) \equiv \tilde{\boldsymbol{T}} \circ \boldsymbol{a}(\boldsymbol{o}^{\mathrm{p}}, \boldsymbol{o}^{\mathrm{v}}). \tag{4}$$

We separately denote proprioceptive observations $\boldsymbol{o}^{\mathrm{p}}$ and exteroceptive observations $\boldsymbol{o}^{\mathrm{v}}$, respectively. Proprioceptive observations contain information about the current and past poses of the robot's hand, while exteroceptive observations such as RGB-D images capture information about the robot's environment.

2.3 Flow Matching for Policies in Euclidean Space

The objective of Flow Matching is to learn a probability distribution based on samples provided by modeling the flow of probabilities. Effectively, a Flow Matching model learns to transform a simple source distribution p_0 into a target distribution p_1. Using Rectified Linear Flow [16, 17], a sample from p_0 is transformed into a sample from p_1 on a straight line. In Robotics, this technique is used to learn policies generating action $\boldsymbol{a}$ based on observation $\boldsymbol{o}$, which in effect models the conditional distribution $p_1 = p(\boldsymbol{a}|\boldsymbol{o})$ from data $\mathcal{D}$, i.e. task demonstrations. If the action is defined in a Euclidean space[1], the probability path is defined as $\boldsymbol{a}(t) = t\boldsymbol{a} + (1-t)\boldsymbol{\varepsilon}$, with $\boldsymbol{a} \sim p_1$ and $\boldsymbol{\varepsilon} \sim p_0$. The vector field transforming samples from p_0 towards p_1 is then defined as $\nabla_t \boldsymbol{a}(t) = \boldsymbol{a} - \boldsymbol{\varepsilon}$. Using a deep neural network $\boldsymbol{f}_{\boldsymbol{\theta}}$, we are interested in accurately predicting the flow vector for a given noisy action $\boldsymbol{a}(t)$, the denoising time $t \in [0, 1]$, and the observation $\boldsymbol{o}$. Thus, the training objective is to minimize

$$L = \sum_{\mathcal{D}} |\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{a}_{\mathcal{D}}(t), t, \boldsymbol{o}_{\mathcal{D}}) - \nabla_t \boldsymbol{a}_{\mathcal{D}}(t)|_2^2. \tag{5}$$

In this paper, we use Flow Matching to learn a policy that generates an action chunk, i.e. $\boldsymbol{a}_{1:K} = \pi_{\boldsymbol{\theta}}(\boldsymbol{o}^{\mathrm{p}}, \boldsymbol{o}^{\mathrm{v}})$, which is a sequence of actions with horizon K:

$$\pi_{\boldsymbol{\theta}}(\boldsymbol{o}^{\mathrm{p}}, \boldsymbol{o}^{\mathrm{v}}) = \arg\max_{\boldsymbol{a}_{1:K}} p(\boldsymbol{a}_{1:K}|\boldsymbol{o}^{\mathrm{p}}, \boldsymbol{o}^{\mathrm{v}}), \tag{6}$$

by modeling the conditional data distribution using Rectified Linear Flow.

[1]The Flow Matching objective requires random variables to be defined in Euclidean space as the denoising path is defined to be on a straight line.

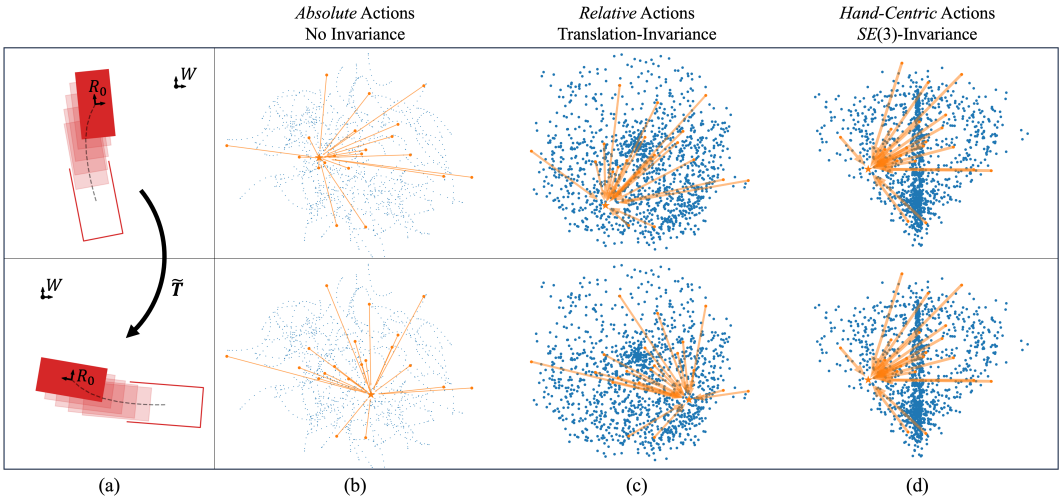

Figure 3: (a) An originally demonstrated action chunk (top) and the same action chunk after applying a rigid transformation (bottom). (b-d) Denoising flows for the 2D position of the first action $\boldsymbol{a}_1$ of the action chunk (orange), and all training samples for $\boldsymbol{a}_1$ in the corresponding control representation (blue). (b): Absolute actions – Denoising flow is not invariant; (c): Relative actions – Denoising flow is translation-invariant only; (d): Hand-centric actions – Denoising flow is $SE(3)$-invariant.

3 Guaranteed Equivariant Control via $SE(3)$-Invariant Flow

Learning skills that are equivariant with respect to global transformations is substantially simplified if those symmetries are exploited. We argue that equivariance of absolute actions is the result of an underlying invariance of the observation-action mapping when changing the frame of reference to a hand-centric representation. In this section, we aim to exploit this by using the hand-centric action representation introduced in Sec. 2.1 to train an invariant policy ${}^R\pi_{\boldsymbol{\theta}}$ to construct an outer policy $\boldsymbol{a}_{1:K}(\boldsymbol{o}^{\mathrm{p}}, \boldsymbol{o}^{\mathrm{v}})$ that is guaranteed to generate equivariant absolute actions $\boldsymbol{a}_k$ based on absolute observations $\boldsymbol{o}^{\mathrm{p}}_{-h}$ (cf. Fig. 1). In particular, this is done by

1. $\boldsymbol{o}^{\mathrm{p}} \to {}^R\boldsymbol{o}^{\mathrm{p}}$: Transforming absolute observations into a hand-centric representation via (3),
2. ${}^R\pi_{\boldsymbol{\theta}}({}^R\boldsymbol{o}^{\mathrm{p}}, \boldsymbol{o}^{\mathrm{v}})$: Inferring hand-centric actions from the learned policy,
3. ${}^R\boldsymbol{a}_{1:K} \to \boldsymbol{a}_{1:K}$: Transforming inferred actions into absolute actions via the inverse of (1).

Theorem 1. *Let* ${}^R\boldsymbol{a}_{1:K} = {}^R\pi_{\boldsymbol{\theta}}({}^R\boldsymbol{o}^{\mathrm{p}}, \boldsymbol{o}^{\mathrm{v}})$ *be a sequence of hand-centric actions and let* $\boldsymbol{o}^{\mathrm{v}}$ *be composed of eye-in-hand observations, then* $\boldsymbol{a}_{1:K}(\boldsymbol{o}^{\mathrm{p}}, \boldsymbol{o}^{\mathrm{v}})$ *is guaranteed to be* $SE(3)$*-equivariant as defined in* (4) *if* $\boldsymbol{a}_k$ *is computed via the hand-centric action with* $\boldsymbol{a}_k(\boldsymbol{o}^{\mathrm{p}}, \boldsymbol{o}^{\mathrm{v}}) = \boldsymbol{o}^{\mathrm{p}}_0 \, {}^R\boldsymbol{a}_k(\boldsymbol{o}^{\mathrm{p}}, \boldsymbol{o}^{\mathrm{v}})$.

In the following, we show in two steps that Theorem 1 is true. First, in Sec. 3.1 we show that the output of the hand-centric policy, i.e. the hand-centric action chunk, has to be invariant to global transformations for Theorem 1 to hold. Second, we prove in Sec. 3.2 that the input to the hand-centric policy is invariant to global transformations, inducing that the output of the hand-centric policy is invariant as well and therefore guaranteeing that the resulting absolute actions are equivariant.

3.1 Hand-Centric Actions Yield $SE(3)$-Invariant Flow

In order to prove that Theorem 1 is true, we compute how the resulting absolute action transforms as a result of a global transformation $\tilde{\boldsymbol{T}}$:

$$\boldsymbol{a}_k(\tilde{\boldsymbol{T}} \circ \boldsymbol{o}^{\mathrm{p}}, \tilde{\boldsymbol{T}} \circ \boldsymbol{o}^{\mathrm{v}}) = \tilde{\boldsymbol{T}} \boldsymbol{o}^{\mathrm{p}}_0 \, {}^R\boldsymbol{a}_k(\tilde{\boldsymbol{T}} \circ \boldsymbol{o}^{\mathrm{p}}, \tilde{\boldsymbol{T}} \circ \boldsymbol{o}^{\mathrm{v}}). \tag{7}$$

Lemma 1. *Theorem 1 holds if the hand-centric policy is invariant to global transformations with* ${}^R\pi_{\boldsymbol{\theta}}(\tilde{\boldsymbol{T}} \circ {}^R\boldsymbol{o}^{\mathrm{p}}, \tilde{\boldsymbol{T}} \circ \boldsymbol{o}^{\mathrm{v}}) \equiv {}^R\pi_{\boldsymbol{\theta}}({}^R\boldsymbol{o}^{\mathrm{p}}, \boldsymbol{o}^{\mathrm{v}})$.

Proof. Assuming invariance of the hand-centric policy, (7) can be reformulated to

$$\begin{aligned} &\boldsymbol{a}_k(\tilde{\boldsymbol{T}} \circ \boldsymbol{o}^{\mathrm{p}}, \tilde{\boldsymbol{T}} \circ \boldsymbol{o}^{\mathrm{v}}) = \tilde{\boldsymbol{T}} \boldsymbol{o}_0^{\mathrm{p}}\, {}^R\boldsymbol{a}_k(\boldsymbol{o}^{\mathrm{p}}, \boldsymbol{o}^{\mathrm{v}}) = \tilde{\boldsymbol{T}} \boldsymbol{a}_k(\boldsymbol{o}^{\mathrm{p}}, \boldsymbol{o}^{\mathrm{v}}) \\ \Rightarrow\ &\boldsymbol{a}_{1:K}(\tilde{\boldsymbol{T}} \circ \boldsymbol{o}^{\mathrm{p}}, \tilde{\boldsymbol{T}} \circ \boldsymbol{o}^{\mathrm{v}}) = \tilde{\boldsymbol{T}} \circ \boldsymbol{a}_{1:K}(\boldsymbol{o}^{\mathrm{p}}, \boldsymbol{o}^{\mathrm{v}}). \end{aligned} \tag{8}$$

□

This is an important result as it shows that the underlying denoising flow $\nabla_t {}^R\boldsymbol{a}(t)$ that is required to transform samples from a source distribution p_0 to the target distribution p_1 has to be invariant to global transformations itself. We hypothesize that the invariance of the denoising flow that is to be predicted with a neural network significantly simplifies learning and lets the policy inherently generalize skills to global transformations that have not been seen during training. This is in contrast to directly learning absolute policies using equivariant encoder-decoder architectures as in related work [6, 7], which facilitates learning equivariant mappings, but does not guarantee the policy to be equivariant. In addition to this analysis, we provide a closed-form derivation of the target denoising flow for *absolute*, *relative*, and *hand-centric* actions in App. A based on the proposition that the training data is generated for an equivariant skill. We find that using absolute actions results in a denoising flow that does not have any symmetry properties. However, using relative actions makes the denoising flow invariant to global translations. The derivation also proves that using hand-centric actions in fact results in a denoising flow that is invariant to global transformations. Fig. 3 illustrates the symmetry properties of the denoising flow for the corresponding action representation. It is worth noting that the invariance in the model output to be learned is independent of the model input, which renders a hand-centric action representation useful even without guaranteeing invariance of the model input.

3.2 Hand-Centric Observations Yield $SE(3)$-Invariant Model Input

The next step of the proof is to show that the hand-centric policy is $SE(3)$-invariant. For this, the input to the policy is required to be invariant to global transformations.

Lemma 2. *Let the proprioceptive observation ${}^R\boldsymbol{o}^{\mathrm{p}}$ and the exteroceptive eye-in-hand observation $\boldsymbol{o}^{\mathrm{v}}$ be $SE(3)$-invariant with $\tilde{\boldsymbol{T}} \circ {}^R\boldsymbol{o}^{\mathrm{p}} = {}^R\boldsymbol{o}^{\mathrm{p}}, \tilde{\boldsymbol{T}} \circ \boldsymbol{o}^{\mathrm{v}} = \boldsymbol{o}^{\mathrm{v}}$, then the hand-centric policy ${}^R\pi_{\boldsymbol{\theta}}({}^R\boldsymbol{o}^{\mathrm{p}}, \boldsymbol{o}^{\mathrm{v}})$ is $SE(3)$-invariant.*

Proof. Applying the invariances of the inputs yields ${}^R\pi_{\boldsymbol{\theta}}(\tilde{\boldsymbol{T}} \circ {}^R\boldsymbol{o}^{\mathrm{p}}, \tilde{\boldsymbol{T}} \circ \boldsymbol{o}^{\mathrm{v}}) = {}^R\pi_{\boldsymbol{\theta}}({}^R\boldsymbol{o}^{\mathrm{p}}, \boldsymbol{o}^{\mathrm{v}})$. □

Therefore, we can conclude that the resulting absolute actions are $SE(3)$-equivariant if the hand-centric observations are $SE(3)$-invariant. A single hand-centric proprioceptive observation ${}^R\boldsymbol{o}^{\mathrm{p}}_{-h}$ at time step $-h$ is defined by (3). Applying a global transformation to that observation results in

$$\tilde{\boldsymbol{T}} \circ {}^R\boldsymbol{o}^{\mathrm{p}}_{-h} = (\tilde{\boldsymbol{T}}\, {}^W\boldsymbol{T}_{R_0})^{-1}\, \tilde{\boldsymbol{T}}\, {}^W\boldsymbol{T}_{R_{-h}} = {}^W\boldsymbol{T}_{R_0}^{-1}\, {}^W\boldsymbol{T}_{R_{-h}} = {}^R\boldsymbol{o}^{\mathrm{p}}_{-h}. \tag{9}$$

Hence, hand-centric proprioceptive pose observations are by-design invariant to global transformations. As noted in Theorem 1, we furthermore assume exteroceptive observations to be gathered from an eye-in-hand perspective, e.g. from a camera that is rigidly attached to the robot's hand. A central proposition to this paper is that eye-in-hand observations $\boldsymbol{o}^{\mathrm{v}}$ are inherently invariant to rigid transformations, i.e. $\tilde{\boldsymbol{T}} \circ \boldsymbol{o}^{\mathrm{v}} \equiv \boldsymbol{o}^{\mathrm{v}}$. This is due to the fact that an eye-in-hand camera, the robot's end-effector, and objects of interest for the task do not move relative to each other under global rigid transformations [13].

4 Experimental Validation

We validate the practical implications of training a hand-centric policy against other action representations in two experiments. For this, we are particularly interested in comparing *sample-efficiency* and *generalization* capabilities on closed-loop control problems with eye-in-hand observations of the scene.

Number of Demos:	5	10	20	40	60	80	100	100 – O.O.D.
Hand-Centric-Flow	**0.01**	**0.25**	**0.56**	**0.79**	**0.90**	**0.94**	**0.98**	**0.96 (-2 %)**
Relative-Action-Flow	0.00	0.00	0.05	0.16	0.42	0.48	0.58	0.01 (-98 %)
Absolute-Action-Flow	0.00	0.00	0.01	0.06	0.08	0.12	0.14	0.02 (-86 %)
Equivariant-DiffPo [7]	0.00	0.01	0.03	0.07	0.12	0.18	0.27	0.12 (-56 %)

Table 1: 2D Peg-in-Hole Insertion: Comparison of success rates over number of demonstrations used for training of four different policy architectures. Success on out-of-distribution (O.O.D.) scenes is reported as a measure of generalization capability. All success rates result from 100 trials.

4.1 Implementation

We note that the introduced hand-centric behavior cloning framework does not impose a particular network architecture to guarantee equivariance of the resulting robot skill as the representation already guarantees equivariance with respect to a transformation in the fixed world frame irrespective of the model architecture. Because of this, we choose to regress the hand-centric denoising flow with a simple multi-layer perceptron (MLP). In both experiments, the demonstrated absolute actions and observations are first pre-processed into chunks $\boldsymbol{a}_{1:K}$ and $\boldsymbol{o}^{\mathrm{P}}_{-H:0}$, respectively. H denotes the history of observations. Next, each pair of chunks is converted into a hand-centric representation by using (1) and (3). For vision-based observations as used in Sec. 4.3, we train end-to-end a ResNet-18 architecture [18] to encode image observations and feed the image features to the MLP.

4.2 2D Peg-in-Hole Insertion

To better understand the impact of the data representation used for behavior cloning, we use a minimal virtual setup of a 2D peg that is supposed to be inserted into a hole as illustrated in Figure 2. Rollouts are simulated only kinematically in this toy example with a collision checker to assess the success of the rollout. In order to emulate eye-in-hand observations of the scene in a simplified way, we explicitly provide two key points on the hole expressed in the robot frame, i.e. $\boldsymbol{o}_{\mathrm{v}} = ({}^{R_0}\boldsymbol{p}_{G_1}, {}^{R_0}\boldsymbol{p}_{G_2})$. This can be interpreted as circumventing a vision encoder by hard-coding spatial image features. Demonstrations are collected from a trajectory optimization pipeline [19] by discarding invalid solutions and gathering successful trajectories in a dataset. A successful rollout fully inserts the 2D peg into the hole without colliding with the edges and corners of the hole.

Sample-Efficiency. Table 1 reports the success rates of trained policies for an increasing number of demonstrations used for training. We observe that the hand-centric representation (Hand-Centric-Flow) leads to higher success rates with few demonstrations and reaches success rates of up to 0.98 with 100 demonstrations. On the other hand, we observe that directly learning an absolute policy (Absolute-Action-Flow) does not capitalize on the equivariance in the underlying task. We furthermore observe a better sample-efficiency for relative actions (Relative-Action-Flow) compared to absolute actions, which indicates that the translation invariance of the denoising flow (cf. Figure 3) simplifies learning, though it still falls significantly short of the sample-efficiency of our approach. Last, we evaluate Equivariant Diffusion Policy (Equivariant-DiffPo) [7] on the same data, observing an improved sample-efficiency compared to the absolute policy without an equivariant network architecture. However, the sample-efficiency of Equivariant-DiffPo is significantly lower compared to Hand-Centric-Flow.

Generalization. We test the generalization capability of the learned policies by sampling peg and hole poses from distributions that are different from the distributions used for generating the training data (out-of-distribution). More specifically, the training data contains scenes where the peg is inserted top-down and sideways, while the out-of-distribution set consists of scenes where the peg has to be inserted bottom-up. The right-most column of Table 1 reports the success rate of the respective policy on the out-of-distribution test set. First, we observe that naive approaches using absolute and relative actions almost drop to zero as they do not leverage the equivariance inherent to the task.

Figure 4: Experimental setup for the real-world visuomotor manipulation task. With an eye-in-hand camera providing RGB-D observations, the task is to insert the manipulation blade into tight gaps and to use sideway-suction to pick thin items from clutter.

Second, Equivariant-DiffPo drops by 56 % compared to the in-distribution performance, which indicates that it partially learned to capture the equivariance of the data. Last, the inherent equivariance of the hand-centric-flow policy that has been derived theoretically in Sec. 3 is confirmed by the performance on the out-of-distribution test set, which is in the same range as the performance on the in-distribution set.

4.3 Real-World Visuomotor Blade Insertion

We validate the practical advantage of using a hand-centric representation for policy learning by applying the presented approach to a real-world visual servoing task in a warehouse setting. Figure 4 illustrates the task setup consisting of a sideway-suction blade attached to a robot arm. The sideway-suction blade is used to pick thin items by engaging with the side of such items instead of the front surface. For this, the blade has to be placed next to the target item such that the suction area makes contact with it. In densely packed environments as faced in warehouse environments, this requires an accurate and precise insertion of the sideway-suction blade between the target item and its neighboring item by feeding back RGB-D observations from an eye-in-hand camera.

We train three policies with different action and observation representations on the same dataset of $N = 6000$ demonstrations collected in simulation. Each trained policy is tested on the same set of 30 real-world, in-distribution (I.D.) scenes. In addition, we validate our hypothesis of gaining inherent generalization by training a hand-centric policy by testing all trained policies on real-world, out-of-distribution (O.O.D.) scenes. While the training data contains demonstrations of the blade picking upright items by inserting the blade into vertical gaps, the O.O.D. scenes require picking thin items that are stacked horizontally. Table 2 reports the success rates, showing that the number of demonstrations was high enough to enable all policies to perform in-distribution. However, only the hand-centric policy is able to pick thin items from horizontal stacks.

	I.D.	O.O.D.
Hand-Centric-Flow	**26/30**	**5/5**
Relative-Action-Flow	19/30	0/5
Absolute-Action-Flow	23/30	0/5

Table 2: 3D Visuomotor Blade Insertion

5 Related Work

SE(3)-Equivariance in Policy Learning. Equivariance of skills has been exploited in many different settings of policy learning. $SE(3)$-equivariant architectures are used as encoder-decoder modules in network architectures in reinforcement learning settings [20, 21] and in behavior cloning settings [6, 7]. Such approaches facilitate learning equivariant policies from data, but they do not constrain the policy to be equivariant. As a result, generalization capabilities are improved but not guaranteed. We compare the performance of our proposed approach with the Equivariant Diffusion Policy proposed by Wang et al. [7] in our experimental validation in Sec. 4.

SE(3)-Invariance in Policy Learning. Exploiting invariances for efficient robot learning has been explored in a few settings. Huang et al. [22] exploit relative features between point clouds and grasp poses to define an $SE(3)$-invariant grasp quality prediction model. Similarly to our approach, Funk et al. [8] propose to define a denoising flow that is invariant to $SE(3)$ transformations to facilitate learning. However, they modify the denoising flow corresponding to absolute actions by rotating the vector field into the predicted hand pose, in fact resulting in a denoising flow that is rotation-invariant, but not translation-invariant (cf. App. A.4). In this paper, we show that the right choice of action representation inherently leads to an $SE(3)$-invariant flow without additional modification of the flow formulation and without specialized network architectures.

Hand-Centric Learning. While most works rely on absolute or relative action representations [2, 3, 4, 5, 6, 7, 8, 23], using hand-centric representations has been explored by few works in robot control. Hsu et al. [13] show that hand-centric (or eye-in-hand) observations are invariant to global transformations and thus improve generalization capabilities compared to third-person perspectives. Seo et al. [14] learn gains of an impedance controller in a hand-centric frame to generalize impedance behavior to global transformations. Chi et al. [15] use hand-centric representations for training a diffusion policy as a result of using hand-held grippers for data collection. While they introduce hand-centric control for practical reasons, we establish connections between the choice of the action representation and the effect on symmetries to be learned by the underlying model regardless of the demonstration interface. Most recently, Wang et al. [24] also establish the connection between hand-centric action representations and the resulting control equivariance for Diffusion policies. Our work adds more insights into the generalizability of resulting Flow policies and furthermore shows real-world applicability.

6 Conclusion

This work introduces a theoretically grounded and empirically validated approach to robot policy learning that achieves guaranteed $SE(3)$-equivariance through hand-centric action and observation representations. By framing the learning task in a coordinate system relative to the robot's hand, the proposed Hand-Centric Flow Matching exhibits denoising flows that are invariant to global transformations, thus enabling more robust generalization and higher sample-efficiency compared to absolute and relative representations. Crucially, this invariance holds independently of the neural network architecture, allowing simple models to outperform more complex equivariant architectures. The experimental results underline the importance of representation choice in robot learning and establish hand-centric control as a powerful tool for scalable and generalizable manipulation policies.

However, several directions remain open for exploration. First, the current theory and validation is partially limited to eye-in-hand observations; extending the approach to third-person or mixed-view visual inputs would broaden its applicability. Second, although the method was validated on Cartesian control tasks, many real-world applications require control in joint space or hybrid action spaces – integrating hand-centric invariance in these contexts poses an interesting challenge. Finally, combining the hand-centric framework with reinforcement learning or other policy optimization strategies beyond behavior cloning may offer improved adaptability and performance in more complex, interactive environments.

References

[1] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges, 2021. URL https://arxiv.org/abs/2104.13478.

[2] H. Huang, D. Wang, A. Tangri, R. Walters, and R. Platt. Leveraging symmetries in pick and place. *The International Journal of Robotics Research*, 43(4):550–571, 2024. doi:10.1177/02783649231225775. URL https://doi.org/10.1177/02783649231225775.

[3] H. Ryu, J. Kim, H. An, J. Chang, J. Seo, T. Kim, Y. Kim, C. Hwang, J. Choi, and R. Horowitz. Diffusion-edfs: Bi-equivariant denoising generative modeling on se(3) for visual robotic manipulation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18007–18018, 2024.

[4] T. Gervet, Z. Xian, N. Gkanatsios, and K. Fragkiadaki. Act3d: 3d feature field transformers for multi-task robotic manipulation. In *7th Annual Conference on Robot Learning*, 2023. URL https://openreview.net/forum?id=-HFJuX1uqs.

[5] A. Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, V. Sindhwani, and J. Lee. Transporter networks: Rearranging the visual world for robotic manipulation. *Conference on Robot Learning (CoRL)*, 2020.

[6] J. Yang, Z. Cao, C. Deng, R. Antonova, S. Song, and J. 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.

[7] D. Wang, S. Hart, D. Surovik, T. Kelestemur, H. Huang, H. Zhao, M. Yeatman, J. Wang, R. Walters, and R. Platt. Equivariant diffusion policy. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=wD2kUVLT1g.

[8] N. Funk, J. Urain, J. Carvalho, V. Prasad, G. Chalvatzaki, and J. Peters. Actionflow: Equivariant, accurate, and efficient policies with spatially symmetric flow matching. *arXiv preprint arXiv:2409.04576*, 2024.

[9] M. Shridhar, L. Manuelli, and D. Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Proceedings of the 6th Conference on Robot Learning (CoRL)*, 2022.

[10] V. Vosylius, Y. Seo, J. Uruç, and S. James. Render and diffuse: Aligning image and action spaces for diffusion-based behaviour cloning, 2024. URL https://arxiv.org/abs/2405.18196.

[11] A. Goyal, V. Blukis, J. Xu, Y. Guo, Y.-W. Chao, and D. Fox. RVT-2: Learning precise manipulation from few demonstrations. In *Robotics: Science and Systems (RSS)*, July 2024.

[12] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[13] K. Hsu, M. J. Kim, R. Rafailov, J. Wu, and C. 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.

[14] J. Seo, N. P. S. Prakash, X. Zhang, C. Wang, J. Choi, M. Tomizuka, and R. Horowitz. Contact-rich se(3)-equivariant robot manipulation task learning via geometric impedance control. *IEEE Robotics and Automation Letters*, 9(2):1508–1515, 2024. doi:10.1109/LRA.2023.3346748.

[15] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[16] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

[17] X. Liu, C. Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi:10.1109/CVPR.2016.90.

[19] J. Jankowski, L. Brudermüller, N. Hawes, and S. Calinon. Vp-sto: Via-point-based stochastic trajectory optimization for reactive robot behavior. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10125–10131, 2023. doi:10.1109/ICRA48891.2023.10160214.

[20] E. van der Pol, D. E. Worrall, H. van Hoof, F. A. Oliehoek, and M. Welling. MDP homomorphic networks: Group symmetries in reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

[21] D. Wang, R. Walters, and R. Platt. SO(2)-equivariant reinforcement learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=7F9cOhdvfk_.

[22] H. Huang, D. Wang, X. Zhu, R. Walters, and R. Platt. Edge grasp network: A graph-based se(3)-invariant approach to grasp detection. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3882–3888, 2023. doi:10.1109/ICRA48891.2023.10160728.

[23] E. Aljalbout, F. Frank, M. Karl, and P. van der Smagt. On the role of the action space in robot manipulation learning and sim-to-real transfer. *IEEE Robotics and Automation Letters*, 9(6): 5895–5902, 2024. doi:10.1109/LRA.2024.3398428.

[24] D. Wang, B. Hu, S. Song, R. Walters, and R. Platt. A practical guide for incorporating symmetry in diffusion policy, 2025. URL https://arxiv.org/abs/2505.13431.

[25] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 0(0):02783649241273668, 0. doi:10.1177/02783649241273668. URL https://doi.org/10.1177/02783649241273668.

A Action Representations and Symmetries in the Resulting Denoising Flow

A.1 Absolute Actions yield No Symmetries in the Denoising Flow

An absolute action at future time step k is defined as the pose of the robot's end-effector with respect to a fixed world frame (typically the base of the robot is used), i.e. $\boldsymbol{a}_k = {}^W\boldsymbol{T}_{R_k}$. We can separate this action into a translation component ${}^W\boldsymbol{p}_{R_k} \in \mathbb{R}^3$ and an axis-angle rotation component ${}^W\boldsymbol{r}_{R_k} \in \mathbb{R}^3$ to numerically represent actions as vectors in the tangent space of $SE(3)$. As a result, the denoising flow for the translation component is defined as

$$\nabla_t {}^W\boldsymbol{p}_{R_k}({}^W\boldsymbol{p}_{R_k}) = {}^W\boldsymbol{p}_{R_k} - \boldsymbol{\varepsilon}, \tag{10}$$

with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Applying a rigid 3D transformation $\tilde{\boldsymbol{T}} = [\tilde{\boldsymbol{R}}|\tilde{\boldsymbol{p}}]$ to the scene yields the denoising vector field transformed with

$$\nabla_t {}^W\boldsymbol{p}_{R_k}(\tilde{\boldsymbol{T}} \circ {}^W\boldsymbol{p}_{R_k}) = \tilde{\boldsymbol{R}}\,{}^W\boldsymbol{p}_{R_k} + \tilde{\boldsymbol{p}} - \boldsymbol{\varepsilon} = \tilde{\boldsymbol{T}} \circ \nabla_t {}^W\boldsymbol{p}_{R_k} + \tilde{\boldsymbol{R}}\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}. \tag{11}$$

The denoising flow depends on the translation $\tilde{\boldsymbol{p}}$ and rotation $\tilde{\boldsymbol{R}}$ applied to the scene. Because $\boldsymbol{\varepsilon}$ is not subject to global transformations, the denoising flow is neither equivariant nor invariant with respect to global transformations. We furthermore analyze the rotation component of absolute actions, which results in the following denoising flow

$$\nabla_t {}^W\boldsymbol{r}_{R_k}({}^W\boldsymbol{r}_{R_k}) = {}^W\boldsymbol{r}_{R_k} - \boldsymbol{\varepsilon}. \tag{12}$$

The denoising flow for the rotation transforms in the same way subject to a global transformation:

$$\nabla_t {}^W\boldsymbol{r}_{R_k}(\tilde{\boldsymbol{T}} \circ {}^W\boldsymbol{r}_{R_k}) = \boldsymbol{r}(\tilde{\boldsymbol{R}}\,{}^W\boldsymbol{R}_{R_k}) - \boldsymbol{\varepsilon} = \tilde{\boldsymbol{T}} \circ \nabla_t {}^W\boldsymbol{r}_{R_k} + \tilde{\boldsymbol{R}}\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}. \tag{13}$$

Consequently, symmetries that are present in the data are not leveraged when using absolute observation and action representations. This is also what we observe in our experimental validation (cf. Sec. 4).

A.2 Relative Actions yield Translation-Invariant Flow

A relative action is defined as the delta between two absolute actions, i.e.

$$\Delta\boldsymbol{a}_k = (\Delta\boldsymbol{p}_k, \Delta\boldsymbol{r}_k) = ({}^W\boldsymbol{p}_{R_k} - {}^W\boldsymbol{p}_{R_0}, \boldsymbol{r}({}^W\boldsymbol{R}_{R_k}) - \boldsymbol{r}({}^W\boldsymbol{R}_{R_0})). \tag{14}$$

As for absolute actions, we analyze how the representation of actions affects symmetries in the resulting denoising flow. For this, we denote the denoising flow of the translation component of a relative action with

$$\nabla_t \Delta\boldsymbol{p}_k(\Delta\boldsymbol{p}_k) = {}^W\boldsymbol{p}_{R_k} - {}^W\boldsymbol{p}_{R_0} - \boldsymbol{\varepsilon}. \tag{15}$$

Applying the global transformation $\tilde{\boldsymbol{T}}$ to the delta action yields

$$\nabla_t \Delta\boldsymbol{p}_k(\tilde{\boldsymbol{T}} \circ \Delta\boldsymbol{p}_k) = \tilde{\boldsymbol{R}}\,{}^W\boldsymbol{p}_{R_k} + \tilde{\boldsymbol{p}} - \tilde{\boldsymbol{R}}\,{}^W\boldsymbol{p}_{R_0} - \tilde{\boldsymbol{p}} - \boldsymbol{\varepsilon} = \tilde{\boldsymbol{R}}\,{}^W\boldsymbol{p}_{R_k} - \tilde{\boldsymbol{R}}\,{}^W\boldsymbol{p}_{R_0} - \boldsymbol{\varepsilon}. \tag{16}$$

We see that the denoising flow for the translation component does not depend on global translations anymore. To complete the analysis, we denote the denoising flow for the rotation component of relative actions with

$$\nabla_t \Delta\boldsymbol{r}_k(\Delta\boldsymbol{r}_k) = \boldsymbol{r}({}^W\boldsymbol{R}_{R_k}) - \boldsymbol{r}({}^W\boldsymbol{R}_{R_0}) - \boldsymbol{\varepsilon}. \tag{17}$$

Applying the global transformation $\tilde{\boldsymbol{T}}$ to the delta action yields

$$\nabla_t \Delta\boldsymbol{r}_k(\tilde{\boldsymbol{T}} \circ \Delta\boldsymbol{r}_k) = \boldsymbol{r}(\tilde{\boldsymbol{R}}^W\boldsymbol{R}_{R_k}) - \boldsymbol{r}(\tilde{\boldsymbol{R}}^W\boldsymbol{R}_{R_0}) - \boldsymbol{\varepsilon}. \tag{18}$$

While the denoising flow still depends on global rotations, relative actions yield a translation-invariant flow. Our experimental results in Sec. 4.2 indicate that the translation-invariance improves sample-efficiency.

A.3 Hand-Centric Actions yield $SE(3)$-Invariant Flow

Suppose that the robot is currently in absolute pose ${}^W\boldsymbol{T}_{R_0}$. We then define a hand-centric action at future step k as

$${}^R\boldsymbol{a}_k = {}^{R_0}\boldsymbol{T}_{R_k} = {}^W\boldsymbol{T}_{R_0}^{-1}\,{}^W\boldsymbol{T}_{R_k}. \tag{19}$$

This describes the pose of the robot at future step k from the perspective of the current robot pose. As a result, the translation ${}^{R_0}\boldsymbol{p}_{R_k}$ and rotation ${}^{R_0}\boldsymbol{R}_{R_k}$ components of a hand-centric action, respectively, are computed with

$${}^{R_0}\boldsymbol{p}_{R_k} = {}^W\boldsymbol{R}_{R_0}^\top({}^W\boldsymbol{p}_{R_k} - {}^W\boldsymbol{p}_{R_0}), \quad {}^{R_0}\boldsymbol{R}_{R_k} = {}^W\boldsymbol{R}_{R_0}^\top\,{}^W\boldsymbol{R}_{R_k} \tag{20}$$

Figure 2 illustrates the hand-centric representation for an example insertion task. The full action chunk is then defined as a sequence of hand-centric actions with ${}^R\boldsymbol{a}_{1:K} = ({}^R\boldsymbol{a}_1, ..., {}^R\boldsymbol{a}_K)$. In order to prove that the $SE(3)$-invariance of the target flow is satisfied, we analyze separately how the rigid transformation $\tilde{\boldsymbol{T}}$ affects the denoising flow for the translation and rotation component for a single prediction step, respectively.

Denoising Flow of the Translation Component. The denoising path for the translation is given by

$${}^{R_0}\boldsymbol{p}_{R_k}(t) = t\,{}^{R_0}\boldsymbol{p}_{R_k} + (1-t)\boldsymbol{\varepsilon}, \tag{21}$$

with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Consequently, the denoising vector field that we want to model is given by

$$\nabla_t {}^{R_0}\boldsymbol{p}_{R_k}(t) = {}^{R_0}\boldsymbol{p}_{R_k} - \boldsymbol{\varepsilon} = {}^{W}\boldsymbol{R}_{R_0}^{\top}({}^{W}\boldsymbol{p}_{R_k} - {}^{W}\boldsymbol{p}_{R_0}) - \boldsymbol{\varepsilon}. \tag{22}$$

Now, applying a rigid 3D transformation $\tilde{\boldsymbol{T}} = [\tilde{\boldsymbol{R}}|\tilde{\boldsymbol{p}}]$ to the scene yields the denoising vector field transformed with

$$\begin{aligned} \tilde{\boldsymbol{T}} \circ \nabla_t {}^{R_0}\boldsymbol{p}_{R_k}(t) &= (\tilde{\boldsymbol{R}}\,{}^{W}\boldsymbol{R}_{R_0})^{\top}(\tilde{\boldsymbol{R}}\,{}^{W}\boldsymbol{p}_{R_k} + \tilde{\boldsymbol{p}} - \tilde{\boldsymbol{R}}\,{}^{W}\boldsymbol{p}_{R_0} - \tilde{\boldsymbol{p}}) - \boldsymbol{\varepsilon} \\ &= {}^{W}\boldsymbol{R}_{R_0}^{\top}({}^{W}\boldsymbol{p}_{R_k} - {}^{W}\boldsymbol{p}_{R_0}) - \boldsymbol{\varepsilon} = \nabla_t {}^{R_0}\boldsymbol{p}_{R_k}(t). \end{aligned} \tag{23}$$

This shows that the denoising vector field for the translation at any step k is invariant to a global rigid transformation.

Denoising Flow of the Rotation Component. In order to apply Euclidean Flow Matching for rotations, we choose to represent 3D rotations by using rotation vectors that are constructed by scaling the rotation axis $\boldsymbol{v}$ with the rotation angle ϕ of an axis-angle representation, i.e. $\boldsymbol{r}(\boldsymbol{R}) = \phi\boldsymbol{v} \in \mathbb{R}^3$. According to Euler's Rotation Theorem, there is a unique mapping from a rotation matrix $\boldsymbol{R} \in \mathbb{R}^{3\times3}$ to an axis-angle representation that lets us construct $\boldsymbol{r}(\boldsymbol{R})$. In the following, we interchangeably use ${}^{R_0}\boldsymbol{r}_{R_k} = \boldsymbol{r}({}^{R_0}\boldsymbol{R}_{R_k})$ to describe the rotation from the robot's current pose to the predicted pose at step k. As a result, the denoising path for the rotation component is given by

$${}^{R_0}\boldsymbol{r}_{R_k}(t) = t\,{}^{R_0}\boldsymbol{r}_{R_k} + (1-t)\boldsymbol{\varepsilon}, \tag{24}$$

with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Consequently, the denoising vector field that we want to model is given by

$$\nabla_t {}^{R_0}\boldsymbol{r}_{R_k}(t) = {}^{R_0}\boldsymbol{r}_{R_k} - \boldsymbol{\varepsilon} = \boldsymbol{r}({}^{W}\boldsymbol{R}_{R_0}^{\top}\,{}^{W}\boldsymbol{R}_{R_k}) - \boldsymbol{\varepsilon}. \tag{25}$$

Again applying the rigid 3D transformation $\tilde{\boldsymbol{T}} = [\tilde{\boldsymbol{R}}|\tilde{\boldsymbol{p}}]$ to the scene yields the denoising vector field transformed with

$$\begin{aligned} \tilde{\boldsymbol{T}} \circ \nabla_t {}^{R_0}\boldsymbol{r}_{R_k}(t) &= \boldsymbol{r}((\tilde{\boldsymbol{R}}\,{}^{W}\boldsymbol{R}_{R_0})^{\top}\tilde{\boldsymbol{R}}\,{}^{W}\boldsymbol{R}_{R_k}) - \boldsymbol{\varepsilon} \\ &= \boldsymbol{r}({}^{W}\boldsymbol{R}_{R_0}^{\top}\,{}^{W}\boldsymbol{R}_{R_k}) - \boldsymbol{\varepsilon} = \nabla_t {}^{R_0}\boldsymbol{r}_{R_k}(t). \end{aligned} \tag{26}$$

With the translation and rotation component of the hand-centric action chunk being $SE(3)$-invariant, we conclude that the underlying vector field that we want to capture with our model $\boldsymbol{f_\theta}$ is in fact $SE(3)$-invariant. In figure 3, we illustrate how the choice of the action representation affects the denoising flow for transformed action chunks. *Absolute* actions result in $SE(3)$-equivariant flows as shown in App. A.1. In contrast, *relative* actions make the underlying flow translation-invariant. However, the flow is not rotation-invariant such that the underlying flow transforms upon a global rigid transformation as shown in App. A.2. Only *hand-centric* actions yield an $SE(3)$-invariant flow that is unaffected by global transformations.

A.4 Denoising Flow in ActionFlow is not $SE(3)$-Invariant

Next, we analyze the denoising flow of ActionFlow as presented in Funk et al. [8]. It is claimed that the denoising flow resulting from their formulation is invariant with respect to global $SE(3)$ transformations. We find that their derivation assumes that the source distributin p_0 is affected by global transformations, which cannot be true if the global transformation itself is unknown. In fact, the source distribution for the translation component in [8] is defined as a normal distribution $p_0 = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. In the following, we derive how their denoising flow actually transforms under global $SE(3)$ transformations. For this, we only show that the translation component of the denoising flow is not $SE(3)$-invariant, which suffices to show that the overall denoising process is not $SE(3)$-invariant. In equation (2) in [8], the denoising path for the translation component is defined with

$${}^{W}\boldsymbol{p}_{R_k}(t) = t\,{}^{W}\boldsymbol{r}_{R_k} + (1-t)\boldsymbol{\varepsilon}, \tag{27}$$

with $\boldsymbol{\varepsilon} \sim p_0$. While using an absolute representation for robot actions as in App. A.1, they modify the resulting flow by rotating the flow vectors into the frame of the currently predicted robot pose. In equation (3), they define the denoising flow with

$$\nabla_t {}^{W}\boldsymbol{p}_{R_k}({}^{W}\boldsymbol{p}_{R_k}, {}^{W}\boldsymbol{R}_{R_k}) = {}^{W}\boldsymbol{R}_{R_k}^{\top}\frac{{}^{W}\boldsymbol{p}_{R_k} - {}^{W}\boldsymbol{p}_{R_k}(t)}{1-t}. \tag{28}$$

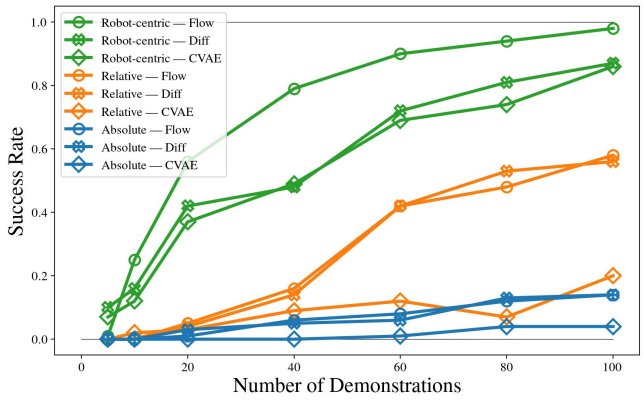

Figure 5: Success rates on the 2D insertion task presented in Sec. 4.2 across different learning algorithms and action representations.

We can rewrite this equation by inserting (27) into (28) to obtain

$$\nabla_t {}^W\boldsymbol{p}_{R_k}({}^W\boldsymbol{p}_{R_k}, {}^W\boldsymbol{R}_{R_k}) = {}^W\boldsymbol{R}_{R_k}^\top({}^W\boldsymbol{p}_{R_k} - \boldsymbol{\varepsilon}). \tag{29}$$

We can now analyze how the denoising flow transforms when a global transformation is applied to the inputs of the flow, i.e.

$$\begin{aligned}\nabla_t {}^W\boldsymbol{p}_{R_k}(\tilde{\boldsymbol{T}} \circ {}^W\boldsymbol{p}_{R_k}, \tilde{\boldsymbol{T}} \circ {}^W\boldsymbol{R}_{R_k}) = (\tilde{\boldsymbol{R}}{}^W\boldsymbol{R}_{R_k})^\top(\tilde{\boldsymbol{R}}{}^W\boldsymbol{p}_{R_k} + \tilde{\boldsymbol{p}} - \boldsymbol{\varepsilon}) = {}^W\boldsymbol{R}_{R_k}^\top({}^W\boldsymbol{p}_{R_k} + \tilde{\boldsymbol{R}}\tilde{\boldsymbol{p}} - \tilde{\boldsymbol{R}}^\top\boldsymbol{\varepsilon}) \\ \neq \nabla_t {}^W\boldsymbol{p}_{R_k}({}^W\boldsymbol{p}_{R_k}, {}^W\boldsymbol{R}_{R_k}).\end{aligned} \tag{30}$$

It can be seen that the denoising flow does depend on global translations and rotations and is thus not $SE(3)$-invariant.

B Consistency of Success Rates Across Learning Algorithms

While we use Flow Matching in this paper to model the generative process of sampling from the data distribution, we conduct an ablation study on the impact of the learning approach. Fig. 5 illustrates the success rates of various policies trained with Flow Matching, Diffusion Modeling [25], and conditional variational autoencoder (CVAE). We observe that the action representation dominates the performance over the choice of the learning algorithm. Results reported for Flow Matching are consistent.