Hierarchical Equivariant Policy via Frame Transfer

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent work in hierarchical policy learning shows the benefits of splitting control into high- and low-level agents, but current approaches overlook how these levels should interact and often ignore domain symmetries, leading to poor data efficiency. We introduce Hierarchical Equivariant Policy (HEP), which connects the two levels through a frame transfer interface and embeds symmetries directly into both agents. This design provides a strong inductive bias while preserving flexibility, yielding state-of-the-art results in both simulation and real-world manipulation.

1 Introduction

Learning-based visuomotor policies have achieved strong results in robotic manipulation, mapping high-dimensional observations to low-level actions [22, 2]. However, end-to-end policies typically require large datasets and struggle to generalize from limited demonstrations due to the complexity of modeling long-horizon reasoning and precise control in one shot.

A natural remedy is hierarchy: a high-level module proposes a subgoal while a low-level module executes fine-grained actions. Prior work validates this direction [13, 21], but commonly enforces a rigid interface (e.g., fixing the low-level goal to the high-level pose), which weakens flexibility and pushes precise geometric reasoning to the high level. Moreover, most hierarchical methods ignore ubiquitous geometric symmetries in manipulation, missing sample-efficiency gains available through equivariant design.

We propose **Hierarchical Equivariant Policy (HEP)**, which couples a flexible interface with symmetry-aware learning. The high-level predicts a *keypose*, a coarse 3D location, that induces a local frame for the low-level controller. Our *Frame Transfer* interface anchors low-level control to the subgoal while allowing local refinement. This decomposition cleanly separates long-horizon reasoning from fine control and provides a natural pathway to incorporate $T(3) \times \mathrm{SO}(2)$ equivariance: the subgoal transforms with the scene (global symmetry) and the low-level behaves consistently in the local keypose frame (local symmetry). To encode 3D perception efficiently, we use a stacked-voxel representation [23] and implement the low level with an equivariant diffusion model.

Contributions. (i) A hierarchical framework with a *Frame Transfer* interface that preserves flexibility while enforcing a strong geometric prior. (ii) A symmetry-integrated design that achieves full $T(3) \times \mathrm{SO}(2)$ equivariance across hierarchy levels. (iii) Extensive evaluation on 30 RLBench tasks and 3 real-world tasks, showing average gains of $\mathbf{10}$–$\mathbf{23}\%$ in simulation, with notable benefits on fine-control and long-horizon tasks.

2 Related Work

Learning from Demonstrations (LfD). Keyframe-based methods predict sparse gripper poses for efficiency [6, 17, 4], but fail on tasks like wiping [21, 13]. Trajectory-based methods imitate dense

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

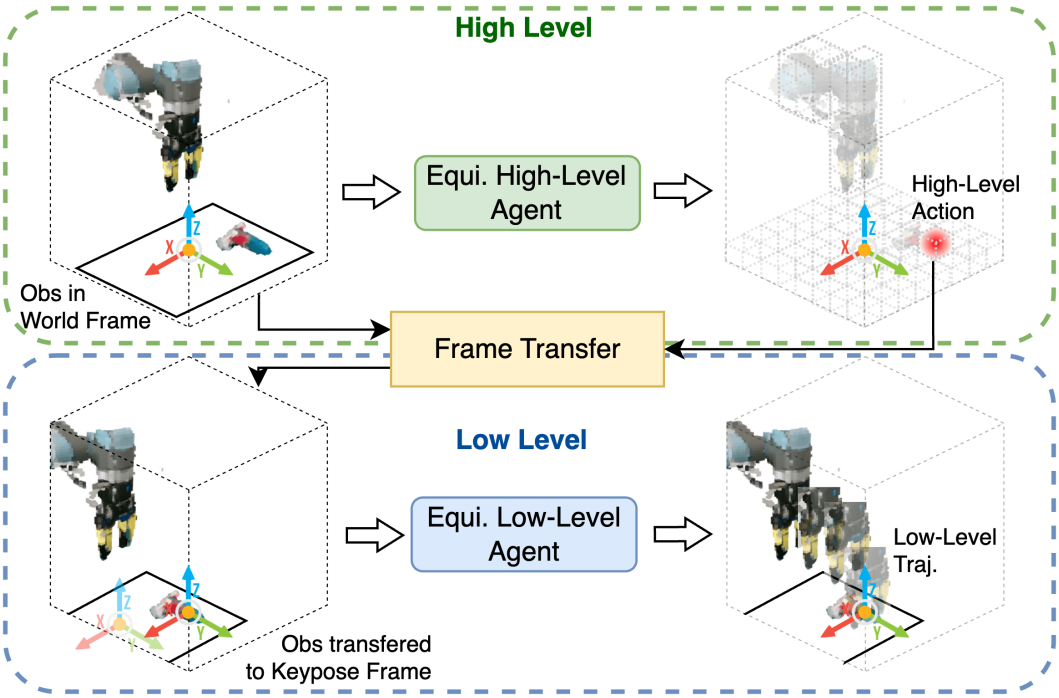

Figure 1: **HEP overview.** A high-level agent predicts a keypose (coarse 3D subgoal). *Frame Transfer* defines a local coordinate frame at this keypose, and a low-level agent predicts trajectories relative to it. This soft interface enables refinement and aligns naturally with symmetry.

actions [2, 20] but suffer from error accumulation. Our Frame Transfer interface bridges the two, anchoring trajectories to subgoals while keeping flexibility.

Hierarchical Policy. Hierarchical control decomposes actions either coarse-to-fine [11, 9] or by motion type [16, 19]. Recent designs enforce high-level outputs as rigid constraints [21, 13]. We instead use Frame Transfer as a soft anchor, simplifying high-level reasoning and enabling low-level correction.

Equivariant Robot Learning. Exploiting geometric symmetries improves sample efficiency and generalization [15, 20]. Prior work applies equivariance at one level only; we embed it into both high- and low-level modules, unifying symmetry across the hierarchy.

3 Background

3.1 Problem Definition

We study visuomotor Behavior Cloning (BC) for manipulation, learning a policy $\pi : O \to A$ that maps observations $o \in O$ to an action sequence $a = \{a_1, \ldots, a_m\} \in A$.

Each action is a gripper state $a_i = (x, y, z, q, c) \in \mathbb{R}^3 \times \mathrm{SO}(3) \times \mathbb{R}$, with position (x, y, z), orientation q, and aperture c. Observations combine geometry and proprioception: a point cloud $P = \{(x_i, y_i, z_i, f_i)\}_{i=1}^{n}$ with k-D features f_i (e.g., RGB), and optionally t steps of gripper state history, so $o \in \mathbb{R}^{n\times(3+k)} \times S^t$.

3.2 Equivariance

We design π to respect scene symmetries under $T(3) \times \mathrm{SO}(2)$ (3D translations and gravity-consistent planar rotations):

$$\pi(g \cdot o) = g \cdot \pi(o), \qquad g = (t, R_\theta).$$

Intuitively, translating/rotating the entire scene should translate/rotate the predicted actions accordingly. More details are provided in the Appendix N.

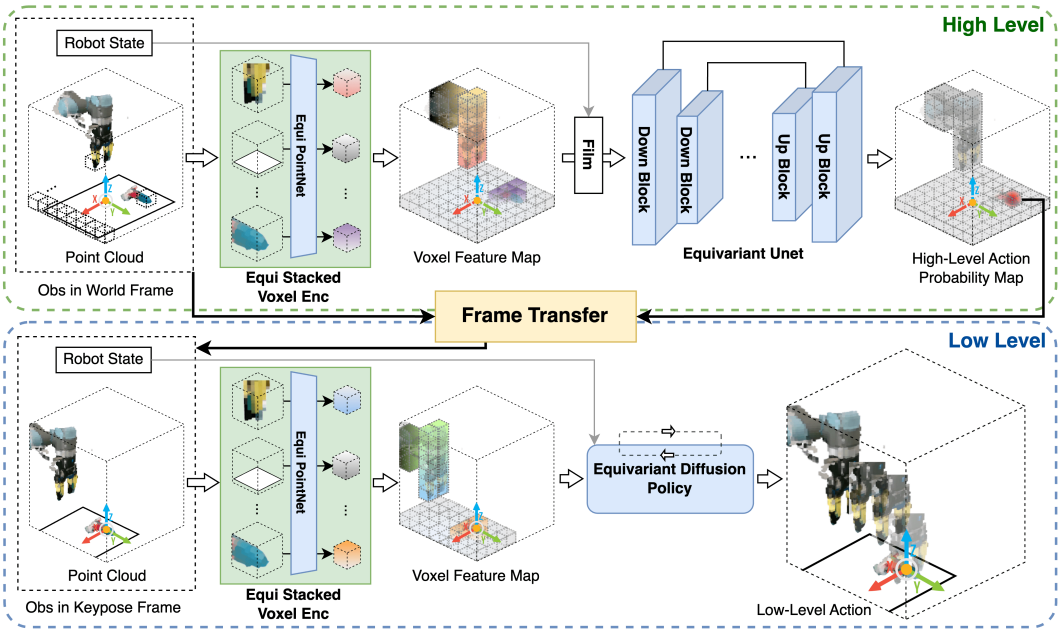

Figure 2: **Overview of Hierarchical Equivariant Policy (HEP).** A high-level agent predicts a coarse translation (keypose). The Frame Transfer interface shifts the input into this keypose frame, where a low-level equivariant diffusion policy generates the trajectory.

4 Method

Our Hierarchical Equivariant Policy (HEP) combines a high-level and low-level agent connected by a novel *Frame Transfer* interface (see Figure 2). The policy is defined as

$$\pi(o) = \pi_{\text{low}}(o, t_{\text{high}}), \quad t_{\text{high}} = \pi_{\text{high}}(o),$$

where the high-level agent predicts a translation $t_{\text{high}} \in T(3)$, and the low-level agent generates a trajectory conditioned on this prediction.

4.1 Frame Transfer Interface

Prior work often fixes the high-level output as a rigid $\mathrm{SE}(3)$ pose target for the low-level [21, 13], which reduces flexibility and burdens the high-level agent. We instead introduce a lightweight *Frame Transfer* interface: the high-level predicts a 3D translation t_{high}, and both observations and trajectories are expressed relative to this keypose frame. The low-level agent then outputs a relative trajectory, which is shifted back to the world frame. This design (see Appendix Q) yields three benefits: (i) efficient communication between levels, (ii) translation invariance in the low-level, and (iii) reduced complexity for the high-level agent, which only predicts a translation.

4.2 High-level Agent

The high-level agent outputs a voxel probability map over candidate translations. We implement it as an $\mathrm{SO}(2)$-equivariant 3D U-Net operating on stacked voxel features (Appendix R). The voxel with the highest probability is selected as t_{high}. Training uses a cross-entropy loss against one-hot expert targets.

4.3 Low-level Agent

Given t_{high} and the frame-transferred observation, the low-level agent generates an $\mathrm{SE}(3)$ trajectory. We build on the Equivariant Diffusion Policy [20], which denoises a noisy trajectory step-by-step while preserving $\mathrm{SO}(2)$ symmetry. This ensures that if the observation is rotated, the predicted trajectory rotates accordingly. Full loss details are given in Appendix S.

4.4 Symmetry of the Full Policy

HEP is designed to be equivariant under $T(3) \times \mathrm{SO}(2)$. Intuitively, shifting or rotating the entire scene should shift or rotate both the high-level keypose and the low-level trajectory in the same way. We formally prove this in Appendix T.

5 Simulation Experiment

We provide an overview here; full details are in Appendix U.

Setup. We use RLBench [8] with a 7-DoF Franka and four RGB-D cameras,10 to 30 tasks, 100 demos per task. Both *open-loop* and *closed-loop* are considered. HEP supports both via Frame Transfer.

Baselines. Open-loop: 3D Diffuser Actor, Chained Diffuser. Closed-loop: EquiDiff.

Results. Success is averaged over 100 eval episodes. HEP achieves SOTA in both *open-loop* and *closed-loop* (means below; full per-task table in Table 6).

Table 1: Mean success rates (%, last checkpoint).

Method	Open-loop	Closed-loop
HEP (ours)	**88**	**79**
Chained Diffuser	78	–
3D Diffuser Actor	56	–
EquiDiff	–	57

HEP improves means by +10 (open-loop) and +22 (closed-loop), leading on 28/30 open-loop tasks.

Ablations. Removing equivariance, Frame Transfer, or stacked voxels reduces performance; the complete model performs best.

6 Real-World Experiment

We deploy HEP on a UR5 robot with three RealSense cameras across three long-horizon tasks (Figure 6). Compared to Chained Diffuser and EquiDiff, our method achieves significantly higher success under both open- and closed-loop control (Table 8).

Beyond overall performance, HEP shows strong generalization: with only a single demonstration it succeeds in 80% of trials, far surpassing baselines (Figure 7, Table 9), and it remains robust under test-time variations such as color changes and distractor objects (Table 10).

Full experimental setup, task details, and analysis are in Appendix V.

7 Conclusion

In this work, we propose an Hierarchical Equivariant Policy for visuomotor policy learning. By utilizing Frame Transfer, our architecture naturally has both translational and rotational equivariance. Experimentally, HEP achieves significantly higher performance than previous methods on behavior cloning tasks that require fine motor control. One key limitation is that our experiments focus on tabletop manipulation. Extending HEP to more complex robotic tasks, such as humanoid motion, is a promising direction.Another limitation is the lack of memory mechanisms, which can be critical for tasks requiring history information. Future work could explore integrating Transformers [18] to enhance temporal reasoning.