EquiContact: A Hierarchical SE(3) Vision-to-Force Equivariant Policy for Spatially Generalizable Contact-rich Tasks

Joohwan Seo*, Arvind Kruthiventy*, Soomi Lee*, Megan Teng*, Seoyeon Choi*, Xiang Zhang*,
Jongeun Choi† and Roberto Horowitz *
*University of California, Berkeley, †Yonsei University
E-mails: {joohwan_seo, arvindkruthiventy, soomi_lee,
meganteng, xiang_zhang_98, seoyeon99, horowitz}@berkeley.edu
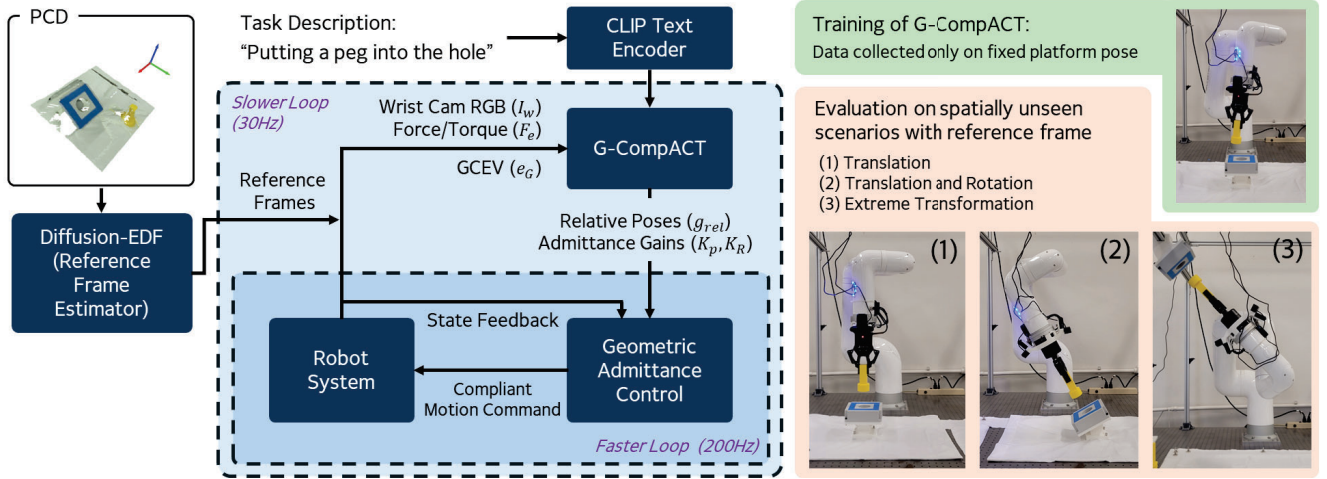, jongeunchoi@yonsei.ac.kr

Fig. 1: We propose an EquiContact, a hierarchical, provably $SE(3)$ vision-to-force equivariant policy for spatially generalizable contact-rich tasks. (**Left**) The proposed EquiContact consists of a Diffusion-Equivariant Descriptor Field (Diff-EDF) and a Geometric Compliant Action Chunking Transformer (G-CompACT). The Diff-EDF, the high-level planner, first processes the scene point cloud to produce reference frames for pick-and-place tasks for the G-CompACT to anchor on. With the provided reference frames, the G-CompACT outputs the relative pose and admittance gains from real-time wrist cameras and proprioceptive feedback. The output relative pose and admittance gains are then fed to geometric admittance control (GAC) that provides compliant motion command to the robot. (**Right**) The G-CompACT method is trained only on the fixed task configuration, but it can be generalized to task configurations that undergo arbitrary $SE(3)$ transformation, given the reference frames.

Abstract—This paper presents a framework for learning vision-based robotic policies for contact-rich manipulation tasks that generalize spatially across task configurations. We focus on achieving robust spatial generalization of the policy for the contact-rich tasks trained from a small number of demonstrations. We propose EquiContact, a hierarchical policy composed of a high-level vision planner (Diffusion Equivariant Descriptor Field, Diff-EDF) and a novel low-level compliant visuomotor policy (Geometric Compliant Action Chunking Transformers, G-CompACT). G-CompACT operates using only localized observations (geometrically consistent error vectors (GCEV), force-torque readings, and wrist-mounted RGB images) and produces actions defined in the end-effector frame. Through these design choices, we show that the entire EquiContact pipeline is $SE(3)$-equivariant, from perception to force control. We also outline three key components for spatially generalizable contact-rich policies: compliance, localized policies, and induced equivariance. Real-world experiments on peg-in-hole (PiH), screwing, and surface wiping tasks demonstrate a near-perfect success rate and robust generalization to unseen spatial configurations, validating the proposed framework and principles. The experimental videos will be attached as supplementary material, and the codes will be released.

I. INTRODUCTION

Imitation learning has recently shown significant success in expanding the capabilities of machine learning in real-world robotics applications [13, 1]. In the early stages of robot learning, many methods formulated manipulation as sequences of keyframe-based pick-and-place actions [34, 24]. More works have started to produce a continuous set of actions directly from vision inputs [4, 37, 23]. Similar to the trend seen in large language models (LLMs), there is a growing belief that large-scale data can unlock generalizable, vision-based policies for robotics [9]. This has led to massive efforts to build large datasets [13] for training policies with general knowledge.

However, such policies often lack spatial generalizability and therefore require a large amount of data to learn robust behaviors. As described in [30], both action chunking

transformers (ACT) [37] and diffusion policy (DP) [4] are evaluated only within the limited spatial variations. Furthermore, both methods exhibit near-linear performance growth as the demonstration dataset size increases, suggesting that the trained policies do not inherently generalize well to new spatial configurations, but rather tend to interpolate between seen demonstrations.

An alternative line of recent research focuses on leveraging symmetry—particularly equivariance—to enhance spatial generalizability, thereby improving sample efficiency during training [22, 19]. This approach requires less data but comes with its own challenges. Equivariant neural networks, being of a more specialized nature, are often not as well-developed and are more computationally intensive than their non-equivariant counterparts, making real-time and large-scale deployment more difficult. As a result, it becomes more attractive for users to use standard models trained with massive datasets in many instances.

In [19], a $SE(3)$-equivariant gain-scheduling policy using geometric impedance control (GIC) [20, 21] was proposed to solve peg-in-hole (PiH) problems. Inspired by the view that many manipulation tasks can be framed as pick-and-place problems [24], we modeled PiH as a *compliant* pick-and-place task, where final peg poses are provided by vision-based $SE(3)$-equivariant models such as Diffusion-EDF (Diff-EDF) [18]. Since both the high-level planner and low-level variable impedance controller are equivariant, they can be combined to form a vision-to-force equivariant policy. However, in practice, Diff-EDF's placement accuracy proved insufficient for precision tasks, which require sub-millimeter precision (details provided in Appendix AIII-A. This revealed a key limitation: high-level vision planners may capture global structures but struggle with precision and contact-sensitive execution. Henceforth, we introduce an intermediate layer between the planner and the low-level controller, which provides real-time visual feedback to correct the residual errors of the high-level planner.

In this paper, we propose EquiContact, a hierarchical $SE(3)$ vision-to-force equivariant for spatially generalizable, contact-rich tasks. It consists of two main components: a high-level planner using Diffusion Equivariant Descriptor Fields (Diff-EDF) [18], which estimates a local reference frame from point clouds, and a low-level compliant visuomotor policy based on Action Chunking Transformer (ACT) [37], which we refer to as Geometric Compliant ACT (G-CompACT). A key design feature of G-CompACT is that it only relies on local information: the force-torque signal in the end-effector frame, a geometrically consistent error vector (GCEV) [19], and wrist camera inputs. The output of G-CompACT is the relative desired pose and admittance gains, which are then sent to the geometric admittance controller (GAC) module to execute compliant control. Our contribution lies in the framework design, not in specific model choices; for example, Diff-EDF could be replaced by ET-SEED [26], or ACT by other visuomotor policies.

The main contributions of this paper are as follows:

1) We propose EquiContact, a hierarchical, provably $SE(3)$-equivariant policy from point clouds and RGB inputs to interaction forces for executing contact-rich tasks.
2) We identify **three key principles** for spatially generalizable contact-rich manipulation: **(1) left-invariant compliant control action** (via GAC [19]), **(2) localized policy (left invariance)**, and **(3) induced equivariance**. These enable $SE(3)$-equivariant behavior without requiring explicitly equivariant neural networks [22].
3) Under these principles, we present the necessary conditions for $SE(3)$ vision-to-force equivariant policy, and mathematically prove the equivariance property of EquiContact.
4) We demonstrate that EquiContact achieves near-perfect success rates and spatial generalizability when these conditions are met in real robot experiments involving peg-in-hole, screwing, and surface wiping tasks.

From these key principles, we propose a general framework to enhance the spatial generalization and interpretability of vision-based policies, namely, "anchoring localized policy on globally estimated reference frame." We emphasize that our work provides complementary insights to recent trends in robot learning [13, 9, 5, 1] that aim to build generalist policies from large-scale demonstration datasets. Our principles provide structural guidelines for improving spatial generalizability via $SE(3)$ equivariance.

II. RELATED WORKS

Visuomotor Servoing Methods Recently, generative modeling has become mainstream in realizing visuomotor servoing policies. Particularly, there are two dominant methods for visuomotor servoing: Action Chunking with Transformers (ACT) [37] and Diffusion Policy (DP) [4]. ACT uses a conditional variational autoencoder (CVAE) as its generative model, whereas DP uses denoising diffusion. ACT and DP have been extended to other approaches, including compliance and force-reactive behaviors [8, 7, 32], as well as structural improvements [14, 5, 29]. Our work is most closely related to CompliantACT (CompACT) [8], which integrates compliant control for visuomotor policies. We have significantly improved CompACT by incorporating a provable $SE(3)$ equivariant structure.

Equivariant Methods Earlier equivariant approaches attempted to handle manipulation tasks as an extension of pick-and-place tasks, by leveraging $SE(3)$ equivariance from point clouds [25, 26, 18] or $SO(2)$ equivariance [34] from top-down views. Equivariant approaches have been extended to visuomotor policies, such as DP or flow matching, [6, 33], using point clouds. [28, 27] proposed $SO(2)$ equivariant visuomotor policies using 2D images, not fully considering $SE(3)$. In contrast, our approach induces full $SE(3)$ equivariance from vision to control force without relying on explicitly equivariant neural networks, but using structured observations and actions via geometrical canonicalization. Furthermore, by integrating with $SE(3)$ equivariant control, we generalize beyond tabletop settings to contact-rich manipulation.
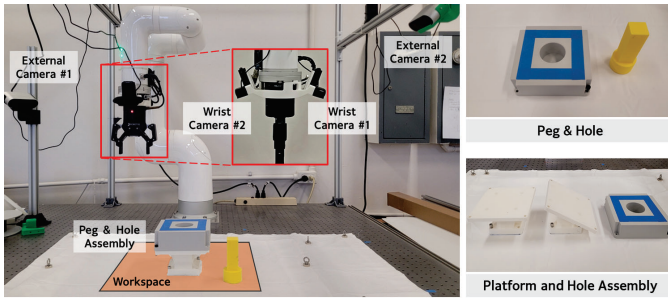
Fig. 2: (Left) Overview of the workspace for the peg-in-hole assembly task is presented. 2 external cameras with calibrated extrinsics and 2 wrists cameras are installed. The workspace shown is the Diff-EDF workspace. (Right-Top) Peg and hole assembly with 1mm of clearance. (Right-Bottom) Hole part with flat and tilted ($30°$) platforms.

Manipulation in Object Frame Our induced $SE(3)$ equivariant approach relies on representing the visuomotor policy in the end-effector frame. While recent works [3, 17, 36] define policies in the target (object) frame, policy representation in the end-effector frame offers improved fidelity and robustness. This is because the estimated object frame can be noisy, and the end-effector frame is reliably obtained via forward kinematics. Importantly, compared to [3, 17], we explicitly link the choice of reference frame to the equivariance property, and unlike [36], which only handles translational transformations, our method can deal with full $SE(3)$ transformations of the reference frame.

III. PROBLEM DEFINITION

In this paper, we aim to identify the key structural components required for learning policies that generalize spatially in contact-rich manipulation tasks. We will first focus on the peg-in-hole (PiH) problem as a representative force-based assembly task and validate the feasibility of the proposed approach to other contact-rich tasks later. Our proposed framework achieves $SE(3)$ vision-to-force equivariance through three essential design principles: (1) left-invariant compliant control, (2) localized policy, and (3) induced equivariance. These principles are validated through specific data collection and evaluation setups, as detailed below.

Unlike prior work [19] that assumes a known hole pose and a pre-grasped peg, we consider a more general setup: the robot must first grasp the peg and then perform insertion using vision, proprioception, and task description in text, as illustrated in Fig. 2. We assume the peg is upright and that the hole's yaw angle is known within $90°$ range. Given an initial estimate, we resolve the orientation by selecting the closest angle among the four symmetric candidates (e.g., $\psi, \psi + 90°, \psi + 180°, \psi + 270°$). As mentioned earlier, high-level vision planners often lack the precision needed for tight-tolerance tasks like PiH, which in our case requires $< 1mm$ accuracy.

To accommodate this, we propose a low-level compliant policy that:
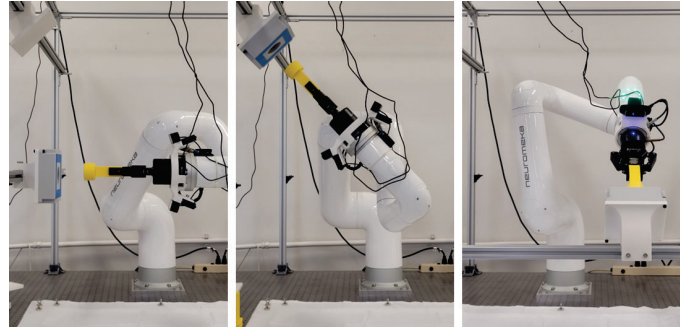- provides real-time visual feedback to refine the coarse high-level command,

Fig. 3: Extreme task transformations. (Left) $90°$ transformation in x axis. (Middle) $135°$ transformation in x axis. (Right) $45°$ transformation in y axis, facing towards the camera.

- handles fine force-based interactions through compliance,
- and achieves provable $SE(3)$-equivariance.

We assume that a high-level vision planner (e.g., Diff-EDF) can generate approximate pick-and-place poses. Our focus is on developing an equivariant compliant placing policy, i.e., insertion policy, using imitation learning. We further assume the peg is approximately aligned with the gripper during placement, since arbitrary peg poses introduce two challenges: (1) imprecise grasps can lead to slippage during contact, and (2) compensating for slippage requires continuous estimation of the gripper-to-peg transformation, which is difficult to achieve reliably in real time.

To train this policy, we collect expert demonstrations of insertions on a fixed-platform setting with a known hole location, where its objective is to train a policy that performs nearly perfectly in the trained scenario. We then evaluate benchmark and proposed methods, trained solely on these limited demonstrations, across arbitrarily translated and rotated test scenarios, thereby isolating and testing individual components of our spatially generalizable contact-rich policy. Further, we demonstrate that our proposed approach can adapt to extreme task transformations as shown in Fig. 3.

Finally, the validity of the proposed EquiContact method is shown in other contact-rich tasks, such as the screwing task and surface-wiping tasks.

IV. SOLUTION APPROACH

We introduce the EquiContact framework in this Section. The EquiContact framework integrates a high-level vision planner (Diffusion Equivariant Descriptor Field, Diff-EDF) with a low-level compliant visuomotor policy (Geometric Compliant ACT, G-CompACT) and geometric admittance control (GAC) at the lowest level. The Diff-EDF gets the point cloud inputs from external cameras to generate reference frames. Based on the estimated reference frames, the G-CompACT process the real-time proprioceptive and wrist camera feedback to output desired poses and admittance gains. In what follows, the GAC module outputs the geometrically consistent compliant motion from desired poses and admittance gains to enable equivariant force interaction. We first focus on the insertion task, with extension to picking addressed later in the paper.
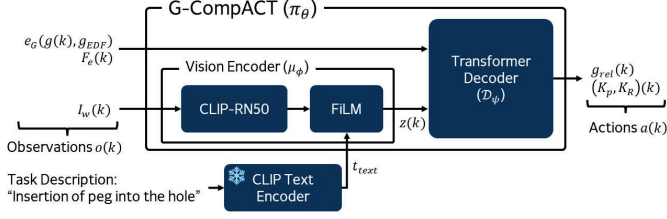
Fig. 4: G-CompACT architecture is presented. The G-CompACT π_θ receives GCEV e_G (2) and F/T sensor value F_e as proprioceptive inputs, along with two wrist camera images I_w. The wrist camera images are fed to the CLIP-ResNet50 visual backbone, followed by the Feature-wise Linear Modulation (FiLM) layer. The FiLM layer is modulated by the text tokens t_{text} from the (frozen) CLIP text encoder, which processes the task descriptions. The proprioceptive inputs and the latent features of modulated vision z are then fed to the transformer decoder D_ψ, which outputs the action signals a. Note that we omitted the style variable for the transformer decoder, and 0 values are used during inference.

Conceptually, EquiContact follows a simple yet powerful principle: **"anchoring a localized policy on a globally estimated reference frame."** In our framework, G-CompACT serves as a fully localized low-level policy, operating solely on observations defined in the current end-effector frame, and actions defined in the end-effector frame – See Fig. 4. The high-level planner, Diff-EDF, estimates the pose of the target (e.g., hole) in the world (global) frame. At the inference, the robot moves near the estimated reference frame, and G-CompACT is activated to perform compliant motion using only local feedback. Because the low-level policy does not depend on absolute global inputs, it can transfer robustly to unseen spatial configurations when the estimated reference frame is provided. This divide-and-conquer design provides a general framework for enhancing both spatial generalization and policy interpretability for contact-rich, and more broadly, general manipulation tasks. In the remainder of this section, we formalize this spatial generalization property as $SE(3)$ equivariance and show how EquiContact satisfies this by its design choices.

A. Geometric Compliant control Action Chunking with Transformers (G-CompACT)

G-CompACT is based on the Action Chunking with Transformer (ACT), which is a CVAE-based generative model designed for imitation learning in robotic manipulation tasks [37]. To make G-CompACT spatially equivariant, we follow the principles proposed in [19]:
- Left-invariant policy, and
- Policy representation in the end-effector body frame

We have designed the G-CompACT architecture to achieve this properties as described in this chapter. The overall structure of the G-CompACT is also summarized in Fig. 4. The observations are given by, (1) Geometrically Consistent Error Vector (GCEV) e_G proposed in [19], (2) FT sensor in the end-effector frame F_e to capture contact behaviors, and (3) RGB images $I_w = \{I_{w,1}, I_{w,2}\}$ from wrist cameras (see Fig. 5). For actions, we choose (1) relative pose from the current end-effector frame g_{rel}, and (2) admittance gains for Geometric

Admittance Control (GAC) (K_p, K_R). The details of GAC and the definition of gains will be provided later in the Section. Formally, the G-CompACT method π_θ can be written as:

$$a(k) = \pi_\theta(o(k)), \quad \text{where}$$
$$a(k) \triangleq (g_{rel}, K_p, K_R)(k), \quad o(k) \triangleq (e_G, F_e, I_w)(k), \quad (1)$$

where $a(k)$ denotes the actions, and $o(k)$ denotes the observation at time step k. Although the G-CompACT outputs the actions of chunk size N, we will only consider the single-step action after proper processing, such as a temporal ensemble, for notational compactness. The GCEV $e_G(g, g_{ref}) \in \mathbb{R}^6$ is defined as

$$e_G(g, g_{ref}) = \begin{bmatrix} R^T(p - p_{ref}) \\ (R_{ref}^T R - R^T R_{ref})^\vee \end{bmatrix}, \quad (2)$$

where $g = (p, R) \in SE(3)$ is a current end-effector pose, $g_{ref} = (p_{ref}, R_{ref}) \in SE(3)$ is the reference frame estimated by the global estimator, e.g., Diff-EDF, and $(\cdot)^\vee$ denotes the vee-map, a mapping from $so(3)$ (Lie algebra of $SO(3)$) to $\mathbb{R}^3$. The physical meaning of GCEV is an error vector between the current end-effector frame and the reference frame, defined on the current end-effector frame. As will be elaborated in Appendix AI, the proprioceptive signals e_G and F_e are left-invariant. For the details of GCEV e_G, we refer to [19, 20].

The images I_w are fed to the transformer decoder D_ψ after being processed by the vision encoder structure μ_ϕ; therefore, one can further represent G-CompACT as

$$a(k) = D_\psi(e_G, F_e, z)(k) = D_\psi(e_G, F_e, \mu_\phi(I_w))(k), \quad (3)$$

where $z = \mu_\phi(I_w)$ is a visual feature from the vision encoder. To satisfy the left-invariant condition of G-CompACT, the features from the vision encoder z need to be invariant to the left-transformation of the image. We formalize the left-invariant visual feature condition as the following assumption.

Assumption 1 (Approximately Left-invariant Visual Features). The visual encoder μ_ϕ produces features that are approximately left-invariant to task transformations, i.e.,

$$\mu_\phi(g_l \circ I_w) \approx \mu_\phi(I_w), \quad (4)$$

$\forall g_l \in SE(3)$ that preserves local task geometry.

Here, $\approx$ notation denotes invariance up to a bounded representation error that does not affect the policy's qualitative behavior. Note that we refer to $\circ$ as a group action [22]. The left-group action applied to the wrist-camera images $g_l \circ I_w$ is illustrated in Fig. 5. The meaning of the visual representation z being left-invariant is that the vision encoder μ_ϕ is trained to focus only on group action invariant features, such as the flat surface surrounding the hole on the platform. To satisfy this assumption, we use language grounding to extract vision features that are correlated with the language description. The core insight behind this is that language tokens encode object identity rather than pose, and thus provide a conditioning signal that is invariant to global $SE(3)$ transformations of the scene. For example, a "peg" is still "peg" no matter from which view it is seen.
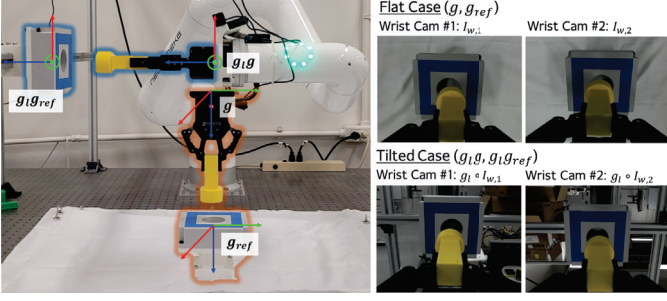
Specifically, the pretrained CLIP-ResNet50 (CLIP-RN50)

Fig. 5: Effects of the left group action g_l to the end-effector pose g and the reference frame g_{ref}, and to the wrists cameras $I_{w,1}$ and $I_{w,2}$. As the left group action is applied to the end-effector and the target object, the wrist cameras start to see the backgrounds of arbitrary lab objects.

[16] is utilized for the vision encoder, and the CLIP text encoder is also employed. Although the pretrained CLIP-RN50 was used, it was fully retrained (details provided in Appendix AII); thereby, it served as a good initialization. We provide the task descriptions to the text encoder to obtain text tokens t_{text}. The vision feature z_{raw} of CLIP-RN50 backbone is then modulated using feature-wise linear modulation (FiLM, [15]) layer from t_{text} via

$$z = \beta(t_{text})z_{raw} + \gamma(t_{text}), \tag{5}$$

where β and γ are trainable FiLM layer. Using FiLM, the vision features are suppressed or highlighted to align with task-relevant semantic concepts, empirically encouraging approximate left-invariance with respect to workspace transformations.

Under the satisfaction of Assumption 1, i.e., ideal left-invariant vision feature condition, the following proposition shows the left-invariance of the G-CompACT method in the end-effector frame.

Proposition 1 (Left-invariance of G-CompACT under *ideal* invariant visual features). Suppose Assumption 1 holds. Then the G-CompACT policy π_θ is left-invariant in the end-effector frame, i.e.,

$$\pi_\theta(g_l \circ o(k)) = \pi_\theta(o(k)), \quad \forall g_l \in SE(3). \tag{6}$$

The proof is presented in Appendix AI. A remark is provided.

Remark 1 (Implication of Proposition 1). Proposition 1 highlights a sufficient structural condition for spatial generalization: if the visual encoder and proprioceptive signals can be made (approximately) left-invariant to $SE(3)$ task transformations, then the resulting closed-loop policy inherits left-invariance by construction. This motivates learning or enforcing representations that satisfy Assumption 1.

In what follows, we present an equivariant property of the pose signal produced by G-CompACT when described in the spatial frame.

Corollary 1 ($SE(3)$ Equivariance of G-CompACT). The G-CompACT π_θ represented in the spatial frame satisfies the following equivariance property:

$$(g_l g_d, K_p, K_R)(k) = \pi_\theta(g_l \circ o(k)). \tag{7}$$

The proof is presented in Appendix AI.

B. Geometric Admittance Control (GAC)

We implement the geometric impedance control (GIC) proposed in [20, 21] in the geometric admittance control (GAC) setup [19]. Let the end-effector pose be denoted as $g \in SE(3)$ in a homogeneous matrix representation, or simply $g = (p, R)$, where $p \in \mathbb{R}^3$ is a position of the end-effector and $R \in SO(3)$ is a rotation matrix of the end-effector. The GAC operates with the (g_d, K_p, K_R) signal calculated from G-CompACT, where the desired end-effector pose is calculated via $g_d = g g_{rel}$. Given $g_d = (p_d, R_d)$, the desired end-effector dynamics for the GAC setup is written as follows:

$$M\dot{V}^b + K_d V^b + f_G = F_e, \tag{8}$$

where $M \in \mathbb{R}^{6 \times 6}$ is symmetric positive definite desired inertia matrix, $K_d \in \mathbb{R}^{6 \times 6}$ symmetric positive definite damping matrix, $F_e \in \mathbb{R}^6$ is external wrench applied to the end-effector in end-effector body frame and $V^b \in \mathbb{R}^6$ is a body-frame end-effector velocity. K_d matrix is selected to ensure overdamped system, as $K_d = 3 \cdot \text{blkdiag}(\sqrt{K_p}, \sqrt{K_R})$. Further, $f_G = f_G(g, g_d, K_p, K_R) \in \mathbb{R}^6$ is a elastic wrench given by:

$$f_G = \begin{bmatrix} f_p \\ f_R \end{bmatrix} = \begin{bmatrix} R^T R_d K_p R_d^T (p - p_d) \\ (K_R R_d^T R - R^T R_d K_R)^\vee \end{bmatrix}, \tag{9}$$

where $K_p, K_R \in \mathbb{R}^{3 \times 3}$ symmetric being positive stiffness matrices for the translational and rotational dynamics, respectively. The desired end-effector pose command is calculated using (8), which is then passed to the robot as the pose command signal. For details on GIC/GAC, we refer readers to [20, 19].

C. Diffusion-Equivariant Descriptor Field (Diff-EDF)

Diffusion-Equivariant Descriptor Field (Diff-EDF) [18] is an $SE(3)$-equivariant reference frame estimator for pick-and-place tasks. In EquiContact, Diff-EDF serves as a high-level vision module that provides a coarse target reference frame for the downstream localized policy.

Given a scene point cloud $\mathcal{O}^{scene}$ and a gripper point cloud $\mathcal{O}^{grasp}$ expressed in the end-effector frame, Diff-EDF outputs an estimated target pose $g_{EDF} \in SE(3)$:

$$g_{EDF} = f_\varphi(\mathcal{O}^{scene}, \mathcal{O}^{grasp}). \tag{10}$$

Diff-EDF is designed to be left-equivariant with respect to $SE(3)$ transformations of the target object [18]. Let $\mathcal{O}^{ref} \subset \mathcal{O}^{scene}$ denote the subset of points corresponding to the object of interest, e.g., hole assembly. Then, for any $g_l \in SE(3)$,

$$f_\varphi(g_l \circ \mathcal{O}^{ref}, \mathcal{O}^{grasp}) = g_l \cdot f_\varphi(\mathcal{O}^{ref}, \mathcal{O}^{grasp}). \tag{11}$$

EquiContact relies only on the equivariance property of the reference frame estimator; the specific architecture of Diff-EDF is otherwise not essential and may be replaced by any $SE(3)$-equivariant reference frame estimator. Importantly, the

localized policy G-CompACT is left-invariant by construction and does not require an $SE(3)$-equivariant estimator. In the absence of an equivariant reference frame estimator, the overall pipeline no longer guarantees end-to-end $SE(3)$ equivariance; still, the local equivariance of G-CompACT is preserved.

D. EquiContact

The proposed EquiContact method comprises the high-level Diff-EDF and the low-level G-CompACT. In Proposition 2, we demonstrate that if an $SE(3)$ equivariant reference frame estimator, such as Diff-EDF, is used, then the resulting EquiContact possesses the equivariance property. Let EquiContact be written as h_Θ so that $h_\Theta(g, g_{ref}, F_e) \mapsto f_G$, i.e., $h_\Theta : SE(3) \times SE(3) \times \mathbb{R}^6 \to \mathbb{R}^6$.

Proposition 2. Suppose that the Assumption 1 holds. The EquiContact policy h_Θ is equivariant if it is described relative to the spatial frame.

The proof is shown in the Appendix AI.

E. Extensions to Pick Tasks

So far, we have described our method in terms of the insertion (placement) task. The proposed method can be extended to pick tasks in the same manner. The Diff-EDF can be utilized to obtain the pick reference frame, which is used for e_G for the picking G-CompACT. The picking G-CompACT is trained in such a way that the manipulator grasps a peg in a fixed, aligned pose, which helps EquiContact bypass the right-equivariance issue. For G-CompACT, the FT sensor values are not utilized as one of its observations, and it does not output the admittance gains; instead, it uses fixed gains.

V. EXPERIMENTS AND DISCUSSIONS

We have conducted sets of experiments to validate the proposed $SE(3)$ vision-to-force equivariance property of the EquiContact. In particular, we aim to answer the following research questions:

RQ1 What are the key principles for spatially generalizable contact-rich manipulation tasks?
RQ2 Can the EquiContact framework be extended to general contact-rich tasks other than the PiH task?
RQ3 Do our design choices of EquiContact really lead to spatial generalization?

First, to answer **RQ1**, we compare the proposed EquiContact against three baselines in the PiH task: ACT with world frame observations and actions, executed with and without GAC, and CompACT [8]. To answer **RQ2**, we have trained and tested EquiContact with known reference frames on the screwing and surface wiping tasks – See Fig. 6. Finally, to answer **RQ3**, we have tested EquiContact with known reference frames on the extreme transformation scenarios for all tasks as shown in Fig. 3.

Before diving into the experimental results, we introduce the implementation details for training and inference of EquiContact.
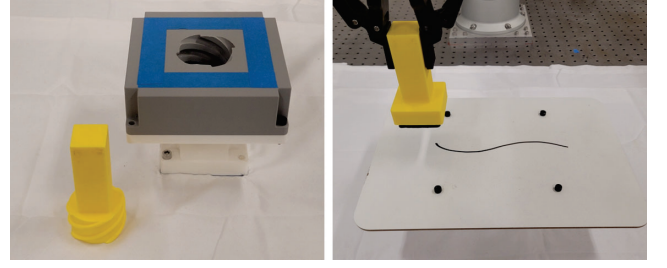

Fig. 6: (Left) Screwing task is first to align the peg to the hole and screw-insert the peg. (Right) Surface wiping (erasing) task is to erase the black marker lines with the eraser. The same platform structure of the PiH is used.

A. Implementation details

Training: First, we note that the G-CompACT and Diff-EDF are separately trained but executed in a single pipeline.

To train a G-CompACT, we collect expert demonstrations via teleoperation at a fixed platform pose. We have collected a dataset not only with a pure white background but also with arbitrary visual distractors, so that the policy can learn to reject background perturbations. We have provided 13-18 prompts for each task. The details regarding the gain modes, language prompts, and scene randomization are provided in Appendix AII.

Since we know the platform's fixed pose a priori during training, e.g., a ground-truth reference frame, the GCEV vector can be computed. Nevertheless, the reference frame needs to be estimated via Diff-EDF (as g_{EDF}) during the inference stage, which may have non-negligible errors. To handle this issue, we have added noise to the reference frame g_{ref} to calculate e_G during dataset preprocessing. This provides the model with an inductive bias to primarily rely on e_G values for rough alignment and rely on vision feedback for fine-grained motion. The rest of the training follows the standard imitation learning pipeline.

To train Diff-EDF, the scene and grasp point clouds are collected together with the target reference frames, which represent the desired poses of the end-effector for pick-and-place operations. 20 demonstrations were collected for the Diff-EDF: 10 samples of the flat platform and 10 samples of the tilted platform, both translationally and rotationally randomized. The training process of Diff-EDF follows the procedure in [18].

Inference: We have implemented the EquiContact pipeline using the ROS2 framework. First, the scene point clouds are obtained and processed by Diff-EDF, producing reference frames for pick-and-place. Using these reference frames, the robot first moves near them, and the G-CompACT is activated near the target objects. During inference, we first obtain the task tokens from the previously used $13-18$ task prompts and feed the mean value of these tokens to the policy. The overall pipeline of the EquiContact is presented in Fig. 1, and also summarized in Algorithm. 1 in the Appendix. Please refer to the Appendix AII for in-depth implementation details.
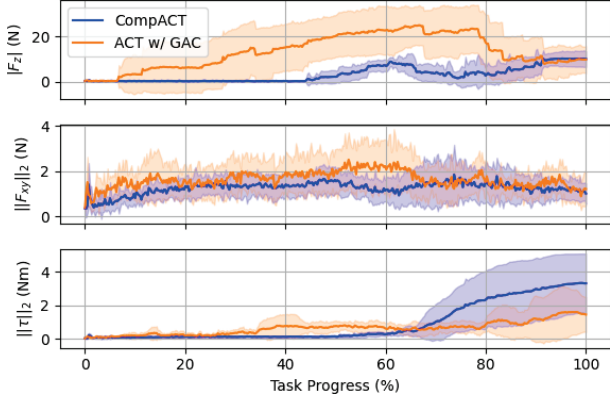
Fig. 7: Force profiles of CompACT and ACT with GAC (fixed gains) during insertion tasks are presented. The CompACT with force-torque sensor inputs and output gains shows lower interaction force in all directions.

B. Peg-in-Hole Benchmark Results

Table. I summarizes the observation/action representations used in each method and reports the benchmark results across all setups.

1) Demonstration of Compliance: As the importance of left-invariant compliant control action has already been verified in [19], we focus on verifying the necessity of compliant control action. We begin by evaluating the role of compliance using the same ACT model architecture, executed with and without the Geometric Admittance Control (GAC). Results for this comparison are shown in the 1st and 2nd rows of Table I. Without GAC, the ACT model shows significantly lower success rates. The failure mode of the ACT w/o GAC involves collision: as the robot approaches the platform, excessive contact forces trigger safety shutdowns, preventing the task from completing. Due to the safety issue, we limited the number of trials without GAC to 5. This result demonstrates that compliant control is nearly a deciding factor between success and failure in contact-rich tasks.

The advantage of variable-compliant gain is that one can achieve the desired force interaction behavior through admittance gains. To show this, we compare ACT with GAC and CompACT under in-distribution (In-Dist.) flat platform settings. Although both methods achieve near-perfect success rates, their force profiles during insertion differ substantially. As shown in Fig. 7, CompACT, which outputs task-adaptive admittance gains based on force-torque feedback, consistently produces lower interaction forces, especially in the z-direction. Note that during the data collection, we modulate the gains to reduce the force interaction in the z-direction but do not consider the magnitudes of torques. As a result, CompACT showed higher interaction torque throughout the task. The effectiveness of the CompACT compared to the baseline ACT was already presented in [8].

2) Demonstration of Equivariance: Although the CompACT succeeds in insertion tasks in trained scenarios without excessive force exertion, it fails to generalize to spatially unseen configurations. This is expected, as its observation and action representations are defined in the global spatial

frame, which neither guarantees nor encourages equivariance. The result of applying CompACT to the translationally unseen cases is shown in the Table I 3rd row - Flat Platform (OOD). Note that only a flat platform is used, meaning it is randomized only translationally. We tested for 10 cases and did not conduct more tests because it resulted in a 0% success rate.

In contrast, the proposed method (EquiContact) achieves perfect success rates on the translationally unseen flat platform, as can be seen in the 4th row of Table I. As shown in Section IV, EquiContact has $SE(3)$ vision-to-force equivariance, achieving a near-perfect success rate, even on the tilted platform, which undergoes a full $SE(3)$ transformation. We attribute the single failure case to a large error from Diff-EDFs that exceeded the noise level applied during training.

The result of EquiContact for the full pick-and-place task is summarized in Table II. The EquiContact also demonstrates a near-perfect success rate in the full pick-and-place pipeline for peg-in-hole tasks.

C. Validating Feasibility to Other Contact-rich Tasks

To further validate the EquiContact framework, we test it on two additional contact-rich tasks: screwing and surface wiping (see Fig. 6). In the screwing task, the robot aligns and screws a peg; in the wiping task, it erases a line from a whiteboard. In this experiment, we assume that the reference frames are known. The reference frame for screwing is the end-effector pose at full insertion; for wiping, it is the center of the board. As in the PiH setup, demonstrations are collected on a fixed platform and evaluated on out-of-distribution configurations, including tilted platforms. Results in Table III show consistent success rates across all conditions, confirming EquiContact's feasibility for various contact-rich tasks.

D. Results on Extreme Transformation Cases

We test the EquiContact method on the extreme transformed configuration for each task. The results are summarized in Table IV. As described in Fig. 5, extreme task transformation cases have different camera inputs, e.g., unseen background objects and light & shadow variations. Therefore, the requirement for invariant visual features becomes more stringent, and a robust vision encoder is needed.

The G-CompACT for the PiH task showed almost perfect success rates on all extreme task transformations, validating that a well-trained G-CompACT policy can handle arbitrary task transformations.

For the surface wiping tasks, the trained G-CompACT policy can handle the smallest angle perturbation $45°$, but failed completely on the other cases. The failure mode is that the robot is trying to track the slots of aluminum extrusion or black cables, not the black lines on the whiteboard. This may be because the prompts we provided for surface wiping include the phrase "the black line." In addition, the lack of a prominent blue square mark on the target object might be the issue. In order to overcome this, one might need to provide more diverse visual distractors that are similar to the black lines, so that the transformer can learn meaningful cross-attention between the

TABLE I: Success rates of the insertion policies in real-world experiments for the proposed and benchmark approaches. "In-Dist." denotes in-distribution data and "OOD" denotes out-of-distribution data. For the In-Dist. (in distribution) scenario, the initial pose of the end-effector is randomized around the flat platform.

Methods	Observation	Action	Test Scenario	Success Rate
ACT w/o GAC	`[World Pose]`	`[World Pose]`	Flat Platform (In-Dist.)	1 / 5
ACT w/ GAC	`[World Pose]`	`[World Pose]`	Flat Platform (In-Dist.)	18 / 20
CompACT	`[World Pose, FT]`	`[World Pose, Gains]`	Flat Platform (In-Dist.) Flat Platform (OOD)	19 / 20 0 / 10
EquiContact (Place, Ours)	`[GCEV, FT]`	`[Relative Pose, Gains]`	Flat Platform (OOD) Tilted Platform (30°, OOD)	20 / 20 19 / 20

TABLE II: Success Rates of the proposed EquiContact for a full pipeline of pick-and-place.

Test Scenario	Success Rate	Failure Cases
Flat Platform (OOD)	20 / 20	N/A
Tilted Platform (30°, OOD)	19 / 20	1 Place

TABLE III: Success Rates of the G-CompACT for screwing and surface wiping tasks. The evaluation is conducted with the ground-truth reference frames.

Test Scenario	Screwing	Wiping
Flat Platform (OOD)	10 / 10	10 / 10
Tilted Platform (30°, OOD)	9 / 10	10 / 10

TABLE IV: Success Rates of the G-CompACT for PiH, screwing, and surface wiping tasks to extreme task transformations (See Fig. 3). The evaluation is conducted with the ground-truth reference frames.

Testing Scenarios	PiH	Screwing	Wiping
45° in y	10 / 10	4 / 10	10 / 10
90° in x	9 / 10	0 / 10	0 / 5
135° in x	10 / 10	2 / 10	0 / 5

vision and GCEV signals. We have not tried more than 5 trials, as it consistently fails.

For the screwing task, we have relaxed the success condition to insert and rotate by at least 20% due to vibration. Unlike the flat platform and 30° platform cases, where the platforms are tightly assembled on the optical table, the platforms for extreme transformation cases are attached to the aluminum extrusion cages, as in Fig. 3. However, as the aluminum extrusion cages are cantilever beams fixed to the optical tables, exerting forces in x and y directions on the end-effector frame leads to high vibration, resulting in complete failures. Despite the relaxed success criterion, the success rates of screwing to these scenarios are significantly lower than those of mild transformations. This is because the screwing task is much more complicated than the PiH task, since it requires perfect alignment to be inserted and progressed. Therefore, we might need a more diverse and carefully curated dataset to finish the task on these cases successfully.

E. Limitations and Future Work

Symmetry Braking: The most prominent failure case of EquiContact is the symmetry braking, specifically, manipulator singularities. When the robot is near singular, pose tracking accuracy degrades because controllers sacrifice tracking to avoid singularity, leading to poorly executed policy commands, resulting in a distributional shift issue. Therefore, the testing scenarios for extreme transformations are carefully selected so that the singularities are not encountered during execution.

Lack of Dataset: Especially for the surface wiping and screwing tasks, the overall quality of the policy could be increased with a larger dataset of better quality. For the surface wiping, the variations of the backgrounds need to be increased, and for the screwing, the demonstration examples with less force interaction are required. In the latter case, a teleoperation device that provides force feedback, i.e., a bilateral teleoperation [10] device, would be beneficial.

Generalization to Other Visuomotor Policies: We have utilized an ACT-based visuomotor policy in our current work, but our EquiContact framework can be generalized to diffusion policy (DP) or flow-matching style.

Vision Encoders: We employed a language-guided visual feature to realize a left-invariance. In fact, the left-invariant vision encoder is closely related to the object-centric representation, such as slot attentions [11]. Moreover, although we have used CLIP-RN50 as our vision backbone, newer versions of vision-language models (VLMs) are available, such as SigLIP [35]. Notably, recent works [12, 31] explored inducing 3D equivariance from 2D images. We will investigate these works of vision encoders to improve left-invariance for future work.

VI. CONCLUSION

In this work, we introduced EquiContact, a vision-to-force equivariant policy for spatially generalizable contact-rich tasks. By integrating a global reference frame estimator (Diff-EDF) with a fully localized visuomotor servoing policy module (G-CompACT), we demonstrate how compliance, localized policy, and induced equivariance can be unified to enable the peg-in-hole (PiH) task, a representative contact-rich precision task, under spatial perturbations. We proved the $SE(3)$ equivariance of the policy under assumptions on point cloud and image observations, validated its effectiveness through real-world experiments on PiH benchmarks, and its feasibility towards screwing and surface wiping tasks. Compared to benchmark methods, our approach generalizes to unseen platform positions and orientations while maintaining low contact force and near-perfect success rates. Through extensive benchmark studies, we highlighted the effectiveness of the three principles – compliance, localized policy, and induced equivariance – for achieving spatial generalizability in contact-rich manipulation. We conclude that these principles offer a

simple yet powerful design guideline for developing spatially generalizable and interpretable robotic policies complementing recent trends in end-to-end visuomotor learning and enabling a structured divide-and-conquer approach.

REFERENCES

[1] Kevin Black et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[2] Francesco Bullo and Richard M Murray. Tracking for fully actuated mechanical systems: a geometric framework. *Automatica*, 35(1):17–34, 1999.

[3] Haonan Chen et al. Tool-as-interface: Learning robot policies from observing human tool use. In *3rd RSS Workshop on Dexterous Manipulation: Learning and Control with Diverse Data*, 2025.

[4] Cheng Chi et al. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.

[5] Sudeep Dasari et al. The ingredients for robotic diffusion transformers. *arXiv preprint arXiv:2410.10088*, 2024.

[6] Nicholas Funk et al. Actionflow: Equivariant, accurate, and efficient policies with spatially symmetric flow matching. *arXiv preprint arXiv:2409.04576*, 2024.

[7] Zihao He et al. Foar: Force-aware reactive policy for contact-rich robotic manipulation. *IEEE Robotics and Automation Letters*, 2025.

[8] Tatsuya Kamijo et al. Learning variable compliance control from a few demonstrations for bimanual robot with haptic feedback teleoperation system. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12663–12670. IEEE, 2024.

[9] Moo Jin Kim et al. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[10] Dongjun Lee and Mark W Spong. Passive bilateral teleoperation with constant time delay. *IEEE transactions on robotics*, 22(2):269–281, 2006.

[11] Francesco Locatello et al. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020.

[12] Thomas Mitchel et al. Neural isometries: Taming transformations for equivariant ml. *Advances in Neural Information Processing Systems*, 37:7311–7338, 2024.

[13] Abby O'Neill et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.

[14] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[15] Ethan Perez et al. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[16] Alec Radford et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[17] Krishan Rana et al. Learning from 10 demos: Generalisable and sample-efficient policy learning with oriented affordance frames. In *Conference on Robot Learning*, pages 5464–5482. PMLR, 2025.

[18] Hyunwoo Ryu et al. Diffusion-edfs: Bi-equivariant denoising generative modeling on SE(3) for visual robotic manipulation. In *2024 IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2024.

[19] Joohwan Seo et al. Contact-rich SE(3)-equivariant robot manipulation task learning via geometric impedance control. *IEEE Robotics and Automation Letters*, 2023.

[20] Joohwan Seo et al. Geometric impedance control on SE(3) for robotic manipulators. *IFAC World Congress 2023, Yokohama, Japan*, 2023.

[21] Joohwan Seo et al. A comparison between lie group- and lie algebra-based potential functions for geometric impedance control. In *2024 American Control Conference (ACC)*, pages 1335–1342. IEEE, 2024.

[22] Joohwan Seo et al. SE(3)-equivariant robot learning and control: A tutorial survey. *International Journal of Control, Automation and Systems*, 23(5):1271–1306, 2025.

[23] Nur Muhammad Shafiullah et al. Behavior transformers: Cloning k modes with one stone. *Advances in neural information processing systems*, 35:22955–22968, 2022.

[24] Mohit Shridhar et al. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning*, pages 894–906. PMLR, 2022.

[25] Anthony Simeonov et al. Neural descriptor fields: SE(3)-equivariant object representations for manipulation. *arXiv preprint arXiv:2112.05124*, 2021.

[26] Chenrui Tie et al. Et-seed: Efficient trajectory-level SE(3) equivariant diffusion policy. *arXiv preprint arXiv:2411.03990*, 2024.

[27] Dian Wang et al. Equivariant diffusion policy. *arXiv preprint arXiv:2407.01812*, 2024.

[28] Dian Wang et al. A practical guide for incorporating symmetry in diffusion policy. *arXiv preprint arXiv:2505.13431*, 2025.

[29] Zhendong Wang et al. One-step diffusion policy: Fast visuomotor policies via diffusion distillation. *arXiv preprint arXiv:2410.21257*, 2024.

[30] Ziniu Wu et al. Fast-umi: A scalable and hardware-independent universal manipulation interface. *arXiv preprint arXiv:2409.19499*, 2024.

[31] Yinshuang Xu et al. SE(3) equivariant ray embeddings for implicit multi-view depth estimation. *Advances in Neural Information Processing Systems*, 37:13627–13659, 2024.

[32] Han Xue et al. Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation. *arXiv preprint arXiv:2503.02881*, 2025.

[33] Jingyun Yang et al. Equibot: Sim(3)-equivariant diffusion policy for generalizable and data efficient learning. *arXiv preprint arXiv:2407.01479*, 2024.

[34] Andy Zeng et al. Transporter networks: Rearranging the

visual world for robotic manipulation. *Conference on Robot Learning (CoRL)*, 2020.

[35] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[36] Haibo Zhao et al. Hierarchical equivariant policy via frame transfer. *arXiv preprint arXiv:2502.05728*, 2025.

[37] Tony Z Zhao et al. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.