EQUIVARIANCE BEGINS AT THE SOURCE: CHARACTERIZING COMPATIBLE SOURCES FOR GENERATIVE ROBOT POLICIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Equivariant transport alone does not guarantee an equivariant manipulation policy. A compatible source distribution—one that transforms consistently with the observation—is required. We show that, under an invertible equivariant flow, the policy inherits the source's equivariance error exactly and therefore cannot correct an incompatible source. In particular, the commonly used isotropic Gaussian centered at the world-frame origin is not translation-equivariant—the KL divergence between the source and its translate grows quadratically with translation distance. We further characterize all compatible conditional sources through a unique frame-relative normal form. For robot actions, this characterization shows that translation requires observation dependence, whereas normalized Haar measure is the unique observation-independent rotational source. Building on these results, we propose GARLIC, an $SE(3)$-equivariant flow-matching policy that anchors its translational source at the end-effector, samples rotations from the normalized Haar measure, and uses an equivariant geometric-algebra transformer for transport. On RLBench, GARLIC improves task success and out-of-distribution generalization over existing baselines, demonstrates greater parameter efficiency, and transfers to real-robot manipulation.

1 INTRODUCTION

When the mug, gripper, and table move together, a generative robot policy should transform its predicted action accordingly; this transformation law is formalized by equivariance. For generative policies, combining an equivariant transport with a compatible source distribution yields an equivariant output distribution (Köhler et al., 2020). The translational component makes source design nontrivial: $\mathbb{R}^3$ admits no translation-invariant probability measure, ruling out an observation-independent compatible source in absolute translation coordinates.

Unlike molecular conformation and protein-backbone generation, where structures can be modeled with center-of-mass-free coordinates and noise (Xu et al., 2022; Hoogeboom et al., 2022; Yim et al., 2023), world-frame robot actions must be located relative to an observed scene. Existing equivariant policies address this translation dependence by restricting their guarantees to rotational subgroups (Wang et al., 2024; Chang et al., 2026) or by using canonicalization and relative actions (Yang et al., 2024a). With the symmetry or action coordinates fixed, these works focus on equivariant denoisers or velocity fields rather than characterizing compatible source distributions for world-frame actions.

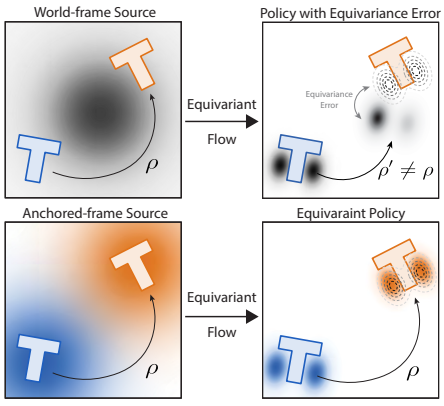

Figure 1: **Equivariance begins at the source.** A flow-matching policy predicts fingertip locations for grasping a T-shaped object. Both panels use the same equivariant flow. **Top:** A fixed world-frame source (gray) induces equivariance error. **Bottom:** An object-anchored source transforms with the observation, yielding an equivariant output distribution.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Flow matching makes the source distribution an explicit design variable and realizes generation through an invertible transport (Tong et al., 2024; Albergo et al., 2023). As flow matching is increasingly adopted for generative policy learning, recent work has explored source design for task information, transport geometry, action history, and proprioception, as well as direct optimization of source samples (Dong et al., 2025; Machado et al., 2026; Patil et al., 2026; Chen et al., 2024a; Kang et al., 2026; Dai et al., 2026). What remains unresolved is how much equivariance error an incompatible source induces under equivariant transport, and which source distributions are compatible.

To answer the first question, we show that an equivariant generation kernel cannot increase the source's equivariance error, and that an invertible equivariant flow preserves it exactly (Theorem 1). For the standard observation-independent Gaussian source, this error has a closed form that grows quadratically with translation distance (Corollary 1). To answer the second question, we characterize compatible source distributions and identify an asymmetry between translation and rotation: a compatible translational source in world coordinates must depend on the observation, whereas normalized Haar measure is the unique observation-independent compatible distribution on $SO(3)$ (Propositions 1–2).

This characterization motivates GARLIC (**G**eometric **A**lgebra for **R**obotic po**LI**cies with **C**ompatible sources), which conditions only its translational source on the observation by anchoring isotropic noise at the end effector, uses a Haar-distributed rotational source, and couples the resulting compatible source with an equivariant geometric-algebra transformer under flow matching.

We summarize our contributions as follows:

- We bound policy equivariance error by source equivariance error for any equivariant generation kernel and show that the bound is tight for invertible equivariant flows (Theorem 1). We further derive a closed-form expression for the error induced by an observation-independent isotropic Gaussian source (Corollary 1).
- We characterize all compatible source distributions through a frame-relative normal form (Proposition 1). This characterization implies that no observation-independent translational probability distribution is compatible in world coordinates and that normalized Haar measure is the unique observation-independent compatible distribution on $SO(3)$ (Proposition 2).
- We introduce GARLIC as a direct instantiation of this characterization, combining an end-effector-anchored translational source, a Haar-distributed rotational source, and an equivariant geometric-algebra transformer under flow matching. This construction yields an idealized policy map that is exactly $SE(3)$-equivariant end-to-end (Corollary 2).
- We validate the equivariance-error propagation result in a controlled planar study that varies only the source. We evaluate GARLIC on four RLBench tasks and four real-robot tasks, where it achieves the highest success rate across all evaluated data budgets and maintains performance under simulated $SE(2)$ scene shifts.

2 Related Work

Compatible sources in equivariant generative models. Equivariant generative models typically pair a source distribution compatible with the group action with an equivariant transport (Köhler et al., 2020; Katsman et al., 2021; Klein et al., 2023; Bose et al., 2024). For Euclidean symmetries, molecular models obtain translation compatibility by using centered coordinates and/or mean-free noise (Xu et al., 2022; Hoogeboom et al., 2022; Yim et al., 2023). GP-EquiFlow instead constructs an $SE(3)$-equivariant source conditioned on observed trajectories for geometric prediction (Wang et al., 2026b). These works focus on compatible source constructions rather than on how source incompatibility propagates through an otherwise equivariant transport. Related work bounds equivariance error arising from non-equivariant drift (Chen et al., 2024b); our analysis instead isolates error originating from the source and shows that invertible equivariant flows preserve it exactly.

Symmetry-aware generative robot policies. Some field-based robot policies target equivariance to particular subgroups, for which an observation-independent invariant source can be compatible (Wang et al., 2024; Chang et al., 2026). Other methods incorporate symmetry through trajectory-level processes, relative poses, or hierarchical frames (Tie et al., 2025; Ryu et al., 2024; Funk et al., 2024; Zhao et al., 2025). Recent symmetry-aware generative policies span flow-based visuomotor control, point-cloud manipulation, vision-language-action policies, and dexterous grasp generation (Yang et al., 2024a; Zhu et al., 2025; Zhang et al., 2026b; Zhang & She, 2026; Ha et al., 2026;

Enwerem et al., 2026; Zhong & Allen-Blanchette, 2025). These approaches do not generally characterize source compatibility for policies predicting absolute poses in world coordinates, which is the setting addressed here.

The source as a design variable. Flow matching permits non-Gaussian sources (Tong et al., 2024), conditional diffusion adapts sources for efficiency (Lee et al., 2022), and text-to-image flows condition sources on the prompt (Kim et al., 2026). In robot policies, sources have been conditioned on behavior policies (Chen et al., 2024a), recent actions (Kang et al., 2026), proprioception (Dai et al., 2026), user-intervenable endpoints (Zhang et al., 2026a), and task semantics (Dong et al., 2025). Other work selects or optimizes sources according to transport efficiency or downstream task performance, for example by selecting among learned priors (Machado et al., 2026) or optimizing individual source samples (Patil et al., 2026). These criteria do not generally test compatibility with the group action: transforming the observation and then sampling the source should agree in distribution with sampling first and then transforming the sample. Our frame-relative characterization provides this criterion for equivariant policies.

Canonicalization, frames, and symmetrization. Canonicalization and symmetrization provide alternative mechanisms for enforcing equivariance, including canonical frames (Kaba et al., 2023; Shumaylov et al., 2025; Sareen et al., 2025; Zhou et al., 2026a;b; Wang et al., 2026a), frame averaging (Puny et al., 2022), kernel symmetrization (Cornish, 2024), and randomized canonicalization (Lawrence et al., 2025). For some group actions, globally continuous canonicalization is impossible, motivating weighted or probabilistic frames (Dym et al., 2024). In robotic settings, partial observations can also make scene-derived frames unreliable, motivating learned or predicted frames (Yang et al., 2024b;a). End-effector proprioception provides another route: the observed end-effector pose directly defines an $SE(3)$ frame. Prior work has found end-effector-frame representations beneficial relative to world- or robot-base-frame alternatives (Liu et al., 2022). Viewed in world coordinates, exact proprioceptive canonicalization induces an observation-dependent, frame-relative source of the form characterized by Proposition 1. Thus, canonicalization and anchored world-frame sources are two mechanisms covered by the same compatibility characterization.

3 PRELIMINARIES

Problem setup. We frame visuomotor manipulation as imitation learning from expert demonstrations. The observation $O = (X, s)$ combines a point cloud $X \in \mathbb{R}^{N\times 3}$ with proprioception $s = (T^{\text{ee}}, g)$, where $T^{\text{ee}} = (R^{\text{ee}}, p^{\text{ee}}) \in SE(3)$ is the end-effector pose and $g \in \{0, 1\}$ is the gripper state. The action $a = (T^a, g^a) \in SE(3) \times \{0, 1\}$, with $T^a = (R^a, p^a)$, specifies a target end-effector pose in *absolute world coordinates*, together with a gripper command.

Group action and the equivariance target. For $\rho = (R, v) \in SE(3)$, rigid transformations act componentwise as

$$\rho \cdot x = Rx + v,\ x \in X,\quad \rho \cdot T^{\text{ee}} = (RR^{\text{ee}}, Rp^{\text{ee}} + v),\quad \rho \cdot g = g, \tag{1}$$

and analogously on actions. The desired policy equivariance is π_θ to be equivariant, $\pi_\theta(\cdot \mid \rho \cdot O) = \rho_{\#}\, \pi_\theta(\cdot \mid O)$ for all $\rho \in SE(3)$, where $\rho_{\#}$ denotes the pushforward under the action transformation and leaves the gripper coordinate unchanged.

Conditional flow matching. Conditional flow matching (CFM) models the conditional action distribution through the ODE $\frac{\mathrm{d}}{\mathrm{d}t} z(t) = v_\theta(z(t), t, O)$ with $z(0) \sim p_0(\cdot \mid O)$, where p_0 is the source distribution and $z = (z_p, z_r, z_g)$ denotes the continuous action representation used by the flow, comprising translation, rotation, and gripper coordinates (Lipman et al., 2023; Liu et al., 2023; Albergo et al., 2023). For the straight path $z_t = (1 - t)z_0 + tz_1$, CFM regresses $v_\theta(z_t, t, O)$ toward $z_1 - z_0$. Integrating the learned velocity field pushes p_0 forward to the policy distribution. Unlike standard diffusion formulations, flow matching leaves the source distribution as an explicit design choice (Liu et al., 2023; Albergo et al., 2023; Chen et al., 2024a).

Conditions for Policy Equivariance A compatible source together with an equivariant generation mechanism is sufficient for policy equivariance (Köhler et al., 2020; Xu et al., 2022; Yang et al., 2024a). Let $K_O(\cdot \mid z)$ denote the conditional generation kernel mapping a source sample z to an action, so that $\pi_\theta(\cdot \mid O) = \int K_O(\cdot \mid z)\, p_0(dz \mid O)$. Sufficient conditions for policy equivariance are

$$\textbf{(A1)}\quad p_0(\rho \cdot A \mid \rho \cdot O) = p_0(A \mid O),\quad \textbf{(A2)}\quad K_{\rho\cdot O}(\rho \cdot B \mid \rho \cdot z) = K_O(B \mid z), \tag{2}$$

for measurable sets A and B; equivalently, $p_0(\cdot \mid \rho \cdot O) = \rho_{\#}\, p_0(\cdot \mid O)$ and $K_{\rho \cdot O}(\cdot \mid \rho \cdot z) = \rho_{\#}\, K_O(\cdot \mid z)$. A1 is source compatibility, and A2 is equivariance of the generation kernel. For flow matching, A2 follows from equivariance of the velocity field, $v_\theta(\rho \cdot z, t, \rho \cdot O) = \rho_* v_\theta(z, t, O)$, where ρ_* denotes the differential of $z \mapsto \rho \cdot z$.

Many equivariant robot policies enforce A2 architecturally using equivariant backbones such as $SE(3)$-Transformers or GATr (Zhong & Allen-Blanchette, 2025; Yang et al., 2024a;b; Tie et al., 2025; Fuchs et al., 2020; Brehmer et al., 2023), often with isotropic Gaussian sources Wang et al. (2024); Sun et al. (2025). Section 4 therefore focuses on A1 and asks: (Q1) how does source equivariance error propagate through an equivariant generation mechanism? (Q2) which source distributions are compatible?

4 Analyzing Source Compatibility

In this section, we first quantify how source incompatibility propagates through an equivariant generation mechanism (Sections 4.1 and 4.2), then characterize all compatible source distributions (Section 4.3). All proofs are provided in Appendix A.

4.1 Propagation of Equivariance Error

For a conditional distribution $q(\cdot \mid O)$, we define its *equivariance error* at (ρ, O) as

$$\mathcal{E}_{\rho,O}[q] = \mathrm{TV}\left(q(\cdot \mid \rho \cdot O), \rho_{\#} q(\cdot \mid O)\right), \tag{3}$$

so that $\mathcal{E}_{\rho,O}[q] = 0$ exactly when q is equivariant at (ρ, O).

Theorem 1 (Equivariance-error propagation)**.** *Assume the generation kernel is equivariant (A2). Then for every $\rho \in SE(3)$ and observation O:* (i) $\mathcal{E}_{\rho,O}[\pi_\theta] \le \mathcal{E}_{\rho,O}[p_0]$*;* (ii) *if K_O is induced by an invertible map, as in an exact ODE flow, then $\mathcal{E}_{\rho,O}[\pi_\theta] = \mathcal{E}_{\rho,O}[p_0]$.*

Thus, an equivariant generation kernel can only reduce source equivariance error, whereas an invertible equivariant flow preserves it. Consequently, a compatible source yields an equivariant policy, but under an invertible flow any source incompatibility is inherited unchanged by the policy.

4.2 Equivariance Error of the Gaussian Source

To make the propagation result explicit, consider the commonly used observation-independent isotropic Gaussian source $p_0 = \mathcal{N}(0, \sigma^2 I)$ in world coordinates. The same data-processing argument used in Theorem 1 applies to KL divergence (Remark 1 in Appendix A.1). We therefore define $\mathcal{E}^{\mathrm{KL}}_{\rho,O}$ by replacing TV with KL; for the Gaussian source, this error has a closed form.

Corollary 1 (Gaussian source equivariance error)**.** *For every $\rho = (R, v) \in SE(3)$, $\mathcal{E}^{\mathrm{KL}}_{\rho,O}[p_0] = \mathrm{KL}(p_0 \,\|\, \rho_{\#} p_0) = \|v\|^2 / 2\sigma^2$. Under an invertible equivariant flow, the policy has the same KL equivariance error. The error depends only on translation and vanishes iff $v = 0$; hence the fixed Gaussian is compatible with rotations but not with nonzero translations.*

Thus, when an observation-independent isotropic Gaussian source is paired with an invertible $SE(3)$-equivariant flow, the resulting policy remains compatible with rotations but inherits the source's translation error. Increasing σ reduces this error as $1/\sigma^2$, but no finite variance eliminates it. This motivates the next question: *which source distributions are compatible?*

4.3 Characterizing Compatible Sources

An *equivariant frame* is a Borel map τ from observations to $SE(3)$ satisfying $\tau(\rho \cdot O) = \rho\, \tau(O)$. In our setting, the end-effector pose provides one such frame. The corresponding canonicalized observation is $M_\tau(O) = \tau(O)^{-1} \cdot O$, which expresses O relative to this frame. The following proposition characterizes compatible sources in terms of distributions defined in these canonical coordinates.

Proposition 1 (Canonical-frame characterization)**.** *Let $\mathcal{C}_\tau = \{O : \tau(O) = e\}$ denote the* canonical slice*. A source distribution p_0 is compatible (A1) if and only if there exists a Markov kernel $\tilde{q}(\cdot \mid c)$, indexed by $c \in \mathcal{C}_\tau$, such that*

$$p_0(\cdot \mid O) \;=\; \tau(O)_{\#}\, \tilde{q}\big(\cdot \mid M_\tau(O)\big) \qquad \textit{for all } O. \tag{4}$$

For a fixed frame τ, the kernel $\tilde{q}$ is unique.

Table 1: **Representative compatible source constructions.** Columns denote construction mechanisms; they are neither mutually exclusive nor exhaustive, and different mechanisms may realize the same compatible source distribution.

	(a) Observation-Independent	(b) Anchored	(c) Canonicalization-Induced
Rotation	Haar (unique)	$m(O) \cdot \kappa_R$	Transform back through $\tau(O)$
Translation	None exists	$b(O) + \kappa_T$	
Gripper	Any law	Not needed	

Canonical-frame constructions of compatible sources are classical (Eaton, 1989; Bloem-Reddy & Teh, 2020); Proposition 1 shows that, for a fixed equivariant frame, every compatible source admits such a representation.

To obtain componentwise consequences, write the action as $a = (p^a, R^a, g^a)$ with translation, rotation, and gripper components. Under $\rho = (R, v)$, the action transforms as $\rho \cdot a = (Rp^a + v, RR^a, g^a)$. We now restrict attention to product sources whose translation, rotation, and gripper factors are conditionally independent given the observation. Lemma 1 shows that the product source satisfies A1 if and only if each factor obeys its corresponding transformation law, allowing the three components to be analyzed separately. We first ask *which compatible factor distributions can be chosen independently of the observation?*

Proposition 2 (Observation-independent compatibility)**.** *For observation-independent factors, (i) no distribution on $\mathbb{R}^3$ is compatible with translations; (ii) the normalized Haar measure is the* unique *probability distribution on $SO(3)$ compatible with rotations; and (iii) every gripper distribution is compatible.*

Thus, among observation-independent factors, compatible distributions range from none for translation, to a unique distribution for rotation, to every distribution for the gripper. Translation is therefore the only factor that requires observation dependence. We next characterize a simple class of observation-dependent sources that restores compatibility.

Corollary 2 (Equivariantly anchored sources)**.** *Let $b(O) \in \mathbb{R}^3$ and $m(O) \in SO(3)$ be equivariant anchors, satisfying $b(\rho \cdot O) = R\, b(O) + v$ and $m(\rho \cdot O) = R\, m(O)$.* (i) *For any fixed rotation-invariant distribution κ_T on $\mathbb{R}^3$, $b(O) + \epsilon$ with $\epsilon \sim \kappa_T$ defines a compatible translation source.* (ii) *For any fixed distribution κ_R on $SO(3)$, $m(O)\, S$ with $S \sim \kappa_R$ defines a compatible rotation source.*

The two cases differ in how the residual transforms. For translation, $b(O)$ accounts for the transformed position, but the residual becomes $R\epsilon$, so κ_T must be rotation-invariant. For rotation, $m(O)$ absorbs the full left action of R, leaving κ_R unrestricted. The gripper is unchanged under $SE(3)$ and therefore requires no anchor. Table 1 summarizes these constructions.

Proposition 1 also captures canonicalization: the observation is expressed in a canonical frame, a source is defined in those coordinates, and the resulting distribution is transformed back to world coordinates through $\tau(O)$ (Yang et al., 2024a; Lawrence et al., 2025; Sareen et al., 2025; Zhou et al., 2026a). In Table 1, canonicalization corresponds to column (c), and equivariantly anchored sources to column (b). Appendix E.2 classifies existing policies according to these constructions.

5 METHOD

Guided by the source characterization in Section 4, we introduce **G**eometric **A**lgebra for **R**obotic po**LI**cies with **C**ompatible sources (GARLIC). GARLIC combines a compatible source constructed according to Corollary 2 with GATr (Brehmer et al., 2023) as the equivariant transport backbone. Its idealized policy map is exactly $SE(3)$-equivariant end to end. Figure 2 gives an overview of the method, and Appendix C provides implementation details.

5.1 SOURCE DISTRIBUTION CONSTRUCTION

Guided by Section 4.3, we construct a compatible source that introduces observation dependence only in translation. We represent an action as (p^a, r, g^a), where $r = [r_1, r_2] \in \mathbb{R}^6$ is the 6D

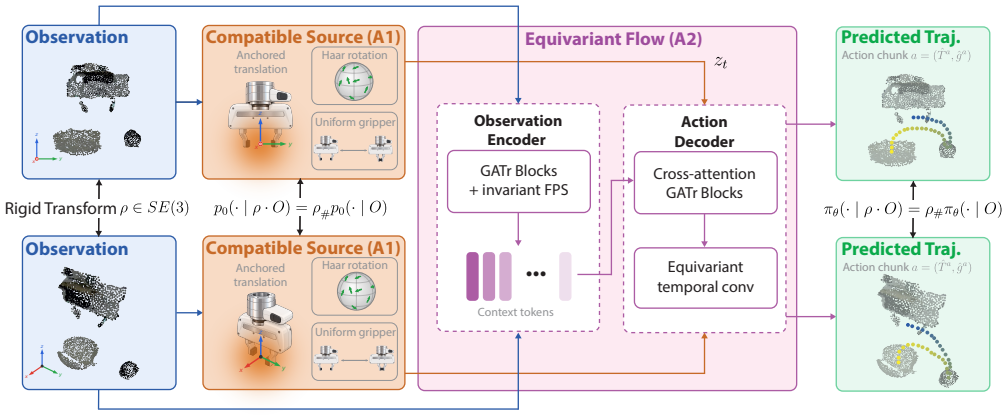

Figure 2: **Method overview. Top:** The policy receives a point cloud and end-effector state; the latter anchors the compatible source in Eq. 5. Self-attention GATr blocks with $SE(3)$-invariant FPS encode the observation history into context tokens; flow states z_t then cross-attend to this context before equivariant temporal convolution and action decoding. **Bottom:** Transforming the observation by $\rho \in SE(3)$ correspondingly transforms the source and predicted trajectory, yielding an $SE(3)$-equivariant policy.

representation of R^a (Zhou et al., 2019). The source factorizes over these components as

$$p_0(\cdot \mid O) = \mathcal{N}\big(p^{\text{ee}}, \sigma^2 I_3\big) \otimes \mathcal{N}\big(0, \sigma^2 I_6\big) \otimes \mathrm{U}\big[-\tfrac{1}{2}, \tfrac{1}{2}\big]. \tag{5}$$

We use the same σ for the translational and rotational Gaussian components. The translational source is anchored at the end-effector position p^{ee}, analogous to centering constructions for equivariant molecular diffusion (Hoogeboom et al., 2022). After Gram–Schmidt projection, the isotropic 6D Gaussian induces the normalized Haar measure on $SO(3)$ (Corollary 3 in Appendix A.4). For flow matching, we represent the binary gripper command by a real-valued coordinate, encoding the two states as $-\frac{1}{2}$ and $\frac{1}{2}$, and use a uniform source on the interval between them. The resulting product source is compatible and therefore satisfies A1.

5.2 Observations in Geometric Algebra

GATr (Brehmer et al., 2023) is an $E(3)$-equivariant transformer that represents 3D geometry in projective geometric algebra. For an observation $O = (X, s)$ with $s = (T^{\text{ee}}, g)$, each point $x \in X$ is encoded as a trivector, yielding N single-channel tokens. The end-effector pose T^{ee} is encoded as a single four-channel token. One channel represents the position as a trivector; the remaining three represent planes through that position, with normals given by the two vectors of the 6D rotation representation and their cross product (see Figure 7 in Appendix C). Because the gripper state g is $SE(3)$-invariant, an MLP encodes it into GATr's invariant scalar channel.

5.3 Flow-Matching Process

Encoder with Invariant Downsampling. The observation history is encoded with self-attention GATr blocks, using $SE(3)$-invariant farthest-point sampling between blocks to reduce the point tokens. We seed FPS using the point farthest from the centroid, making the selected set invariant to rigid transformations and, away from exact ties, input ordering. The resulting tokens from all observation frames form the context representation.

Decoder with Equivariant Temporal Convolution. Each step of the predicted action chunk is represented by a query token that cross-attends to the observation context. Pose coordinates use the geometric representation of Section 5.2, while flow time and step index enter through invariant scalar channels. We couple neighboring action steps with an $SE(3)$-equivariant temporal convolution before the output head; Appendix C.4 gives the construction and proof of equivariance.

Training and Inference. Given z_0 from Eq. 5 and demonstrated action z_1, we sample $t \in [0, 1)$ and form the straight path $z_t = (1-t)z_0 + tz_1$. Following JiT (Li & He, 2025), the network predicts the endpoint $\hat{z}_1$ rather than the velocity, from which we obtain the implied velocity $(\hat{z}_1 - z_t)/(1-t)$. At inference, we sample from the same source and integrate this velocity field with Euler steps.

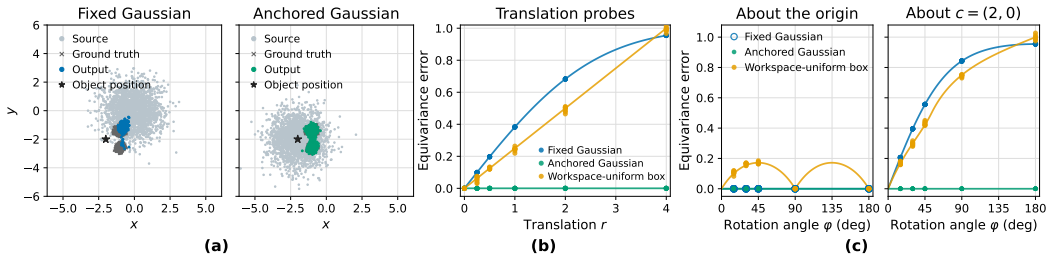

Figure 3: **Controlled validation of source-induced equivariance error. (a) Prediction on an unseen scene.** Predictions from models using fixed and anchored Gaussian sources. **(b) Equivariance error under translation.** TV equivariance error as the scene is translated by r. Markers denote learned-policy errors; curves show the closed-form source errors. **(c) Equivariance error under rotation.** TV equivariance error as the scene is rotated by φ about either the origin or $c = (2, 0)$. The anchored Gaussian remains equivariant for both rotation centers.

6 Experiments

Our experiments address four questions: **(Q1)** Do learned-policy equivariance errors match the source-side predictions from Section 4? **(Q2)** How does GARLIC compare in task performance, sample efficiency, parameter efficiency, and OOD generalization? **(Q3)** How do its main structural design choices affect performance and equivariance? **(Q4)** How does GARLIC perform on real-world manipulation tasks?

6.1 Validating the Equivariance Error on a Planar Toy Task

Section 4 gives exact predictions for how source equivariance error propagates to the policy. We test these predictions in a planar $SE(2)$ setting, where the relevant distributions and errors can be evaluated directly.

World and demonstrations. The action is a planar grasp location $a \in \mathbb{R}^2$, with no orientation or gripper output, allowing the policy densities and equivariance errors to be evaluated directly and visualized. The observation $O = (o, h)$ consists of an object position o and unit heading h; R_o is the planar rotation defined by h. The demonstration distribution is an asymmetric bimodal mixture whose component means are rotated by R_o and translated by o. This construction is equivariant, avoids additional symmetries that could mask source effects, and yields a nontrivial transport problem. Appendix D.1 gives the exact mixture and training distribution.

One equivariant flow, three sources. All conditions use the same frame-conjugated velocity field architecture $v_\theta(z, t, O) = R_o\, f_\theta\big(R_o^\top (z - o),\, t\big)$, where f_θ is a small MLP trained by conditional flow matching with the same optimizer and training budget across conditions. This parameterization satisfies A2 for every θ. Under the resulting invertible ODE flow, Theorem 1(ii) therefore implies that the policy equivariance error equals the source equivariance error. The conditions differ only in the source. *Fixed Gaussian* uses $\mathcal{N}(0, I)$ (Corollary 1), and *Anchored Gaussian* uses $\mathcal{N}(o, I)$ centered at the object (Corollary 2). An observation-independent uniform measure on $\mathbb{R}^2$ would be translation-invariant but cannot be normalized to a probability distribution. *Workspace-uniform box* instead samples uniformly from the training workspace $B = [-2, 2]^2$.

Results. Figure 3 (a) shows that on an unseen observation, the anchored Gaussian captures the demonstrated target distribution, whereas the fixed Gaussian model does not. Panel (b) measures TV equivariance error under translations $v = (r, 0)$ over eight unseen observations. The measured policy errors match the closed-form source-side curves derived in Appendix D.1. Error increases with translation distance for both the fixed Gaussian and the workspace-uniform box, while the anchored Gaussian remains equivariant. The fixed Gaussian error approaches one asymptotically, whereas the workspace-uniform error reaches one when the translation equals the box width ($r = 4$). Thus, restricting a uniform source to the training workspace does not make it compatible. Panel (c) rotates the scene about the origin and about $c = (2, 0)$. Consistent with Corollary 1, the fixed Gaussian has zero equivariance error only for rotations about its center. The workspace-uniform box has zero error only for rotations that preserve the square, whereas the anchored Gaussian remains equivariant about either rotation center. Across both translation and rotation, the measured policy errors follow the source-side predictions of Theorem 1(ii).

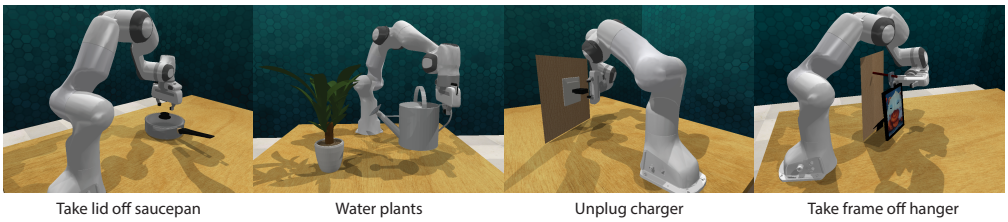

Figure 4: **RLBench tasks.** `Take Lid Off Saucepan` is a lift on a centered handle. `Water Plants` grasps the can and rotates it to pour. `Unplug Charger` and `Take Frame Off Hanger` require precise motion around a small charger and a narrow hanger opening.

6.2 RLBench Performance, Efficiency, and OOD Generalization

We evaluate GARLIC on four simulated RLBench tasks (Figure 4) with 10, 50, and 100 expert demonstrations per task. Each method is trained with three seeds. For each seed, we average performance over the final five checkpoints and report the mean and standard deviation across seeds, using the same 100 randomized evaluation initializations. Setup and training details are given in Appendix D.2, and videos are provided in the supplementary material.

Task performance and sample efficiency. Table 2 compares GARLIC with four baselines: non-equivariant *DP3* (Ze et al., 2024) (DDPM) and *PFM* (Chisari et al., 2024) (flow matching), *EquiBot* (Yang et al., 2024a) with $SO(3)$-equivariant layers and centroid canonicalization, and *DP3-Canon.* with fixed end-effector-frame canonicalization. GARLIC achieves the highest success rate for every task and demonstration. With 50 demonstrations, GARLIC outperforms all 100-demonstration baselines on three of the four tasks. On the fourth, it performs comparably to the best 100-demonstration baseline, indicating greater sample efficiency.

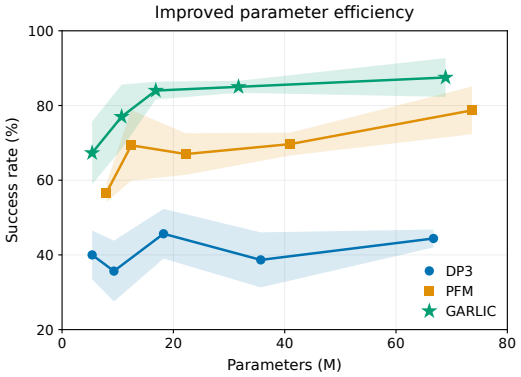

Figure 5: **Parameter efficiency.**

Parameter efficiency. Figure 5 compares success rates across parameter budgets on `Unplug Charger`, using the same training and evaluation protocol as Table 2. Across parameter budgets, GARLIC consistently achieves the highest success rate. Even with fewer than 10M parameters, it matches or exceeds the performance of larger baseline models.

Table 2: **Success rates on four RLBench tasks with 10, 50, and 100 demonstrations.**

# Demos	Method	Take lid off saucepan	Water plants	Unplug charger	Take frame off hanger
10	DP3	34.0 ± 3.0	31.0 ± 3.7	9.7 ± 4.6	24.9 ± 4.2
	PFM	5.7 ± 2.3	11.0 ± 3.6	29.9 ± 3.8	33.7 ± 2.4
	EquiBot	38.8 ± 4.1	28.3 ± 5.6	25.7 ± 4.4	21.0 ± 2.2
	DP3-Canon.	34.0 ± 7.5	34.3 ± 6.0	23.0 ± 3.6	16.8 ± 1.2
	GARLIC	$\mathbf{40.7 \pm 5.5}$	$\mathbf{42.1 \pm 6.2}$	$\mathbf{33.7 \pm 3.1}$	$\mathbf{38.7 \pm 4.2}$
50	DP3	78.7 ± 8.5	69.0 ± 8.4	24.8 ± 4.0	26.9 ± 7.4
	PFM	84.7 ± 2.1	59.3 ± 9.3	50.4 ± 2.7	51.9 ± 5.7
	EquiBot	62.4 ± 5.4	44.9 ± 2.6	44.0 ± 3.0	27.5 ± 4.1
	DP3-Canon.	84.0 ± 5.0	62.7 ± 2.1	42.0 ± 5.0	27.9 ± 4.2
	GARLIC	$\mathbf{92.0 \pm 2.4}$	$\mathbf{81.3 \pm 1.5}$	$\mathbf{56.5 \pm 3.4}$	$\mathbf{63.7 \pm 2.9}$
100	DP3	85.0 ± 5.0	73.3 ± 6.2	44.4 ± 2.4	38.3 ± 5.9
	PFM	92.3 ± 3.1	73.0 ± 3.5	53.3 ± 4.2	55.2 ± 4.1
	EquiBot	65.3 ± 5.4	67.1 ± 5.7	48.3 ± 4.0	42.4 ± 2.1
	DP3-Canon.	89.7 ± 1.2	58.0 ± 5.3	43.3 ± 5.8	38.7 ± 4.0
	GARLIC	$\mathbf{95.7 \pm 1.7}$	$\mathbf{90.7 \pm 2.1}$	$\mathbf{85.7 \pm 5.2}$	$\mathbf{71.3 \pm 4.0}$

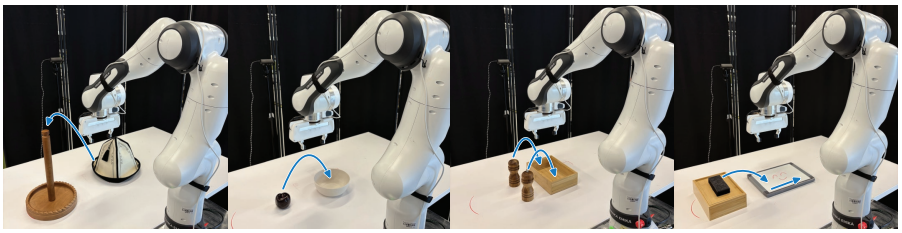

Figure 6: **Real world experiments.** `Hang hat on rack` and `Place cherry in bowl` are single-stage pick-and-place tasks. `Store pepper grinder` and `Erase whiteboard` require two consecutive manipulation stages. Arrows indicate the motion of the grasped object.

OOD generalization. We evaluate generalization to OOD scenes obtained by rigidly transforming in-distribution scenes. Each evaluation scene is transformed by a random $\rho \in SE(2)$, consisting of rotation about the gravity axis and translation in the horizontal plane; gravity prevents full $SE(3)$ symmetry. We evaluate the checkpoints from Table 2 without fine-tuning. On `Unplug Charger`, Table 3 shows that GARLIC retains its high ID performance under these transformations, with an ID–OOD gap of only -1.2 percentage points. DP3-Canon. exhibits a similarly small gap but at substantially lower absolute success, while DP3 and PFM degrade sharply. Results on the remaining three tasks are reported in Table 8.

Table 3: **OOD generalization.**

Method	ID	OOD	gap Δ
DP3	44.0 ± 2.0	2.7 ± 3.1	-41.3
PFM	54.2 ± 3.9	5.3 ± 4.2	-48.9
DP3-Canon.	42.7 ± 1.2	41.3 ± 3.0	-1.4
GARLIC	$\mathbf{85.3 \pm 4.1}$	$\mathbf{84.1 \pm 3.1}$	$\mathbf{-1.2}$

6.3 Ablation Study

We ablate the design choices of GARLIC that follow from Section 4. *Fixed Gaussian* replaces our observation-dependent source with the observation-independent $\mathcal{N}(0, I)$. *Bingham prior* replaces the rotation source with a compatible observation-dependent Bingham distribution on S^3 (Corollary 2). *Remove GATr* replaces the equivariant GATr backbone with PFM, removing the architectural guarantee of A2. *DP3+Aug* replaces explicit equivariance with random $SE(2)$ data augmentation. Table 4 reports results on `Unplug Charger`. The unchanged model achieves the highest ID and OOD success rates, followed by the Bingham variant, which also has a small ID–OOD gap.

Table 4: **Ablation study.** We ablate GARLIC to violate conditions A1 and A2 from Section 3.

Ablation	A1	A2	ID	OOD	gap Δ
Fixed Gaussian	–	🌰	50.7 ± 5.5	31.3 ± 5.0	-19.4
Bingham prior	🌰	🌰	$\underline{71.2 \pm 7.4}$	$\underline{68.4 \pm 8.9}$	$\underline{-2.8}$
Remove GATr	🌰	–	61.7 ± 2.1	8.7 ± 5.0	-53.1
DP3 + Aug	–	–	51.5 ± 4.9	4.0 ± 0.3	-47.5
Unchanged	🌰	🌰	$\mathbf{84.7 \pm 5.1}$	$\mathbf{83.2 \pm 4.7}$	$\mathbf{-1.5}$

6.4 Real-World Experiment

We evaluate GARLIC on four real-world manipulation tasks (Figure 6) using a Franka arm and everyday objects. For each task, we collect 20 human teleoperation demonstrations using a Meta Quest and evaluate on 10 object poses not seen during training, with one trial per pose. GARLIC achieves the highest success rate on all four tasks (Table 5). Hardware and preprocessing details are given in Appendix D.3, and videos are provided in the supplementary material.

Table 5: **Success rates on real-world tasks.** GARLIC outperforms baselines on every task.

Method	Hat	Cherry	Grinder	Eraser
DP3	0.7	0.6	0.4	0.3
PFM	0.6	0.5	0.2	0.4
GARLIC	**1.0**	**0.8**	**0.7**	**0.8**

7 Conclusion

We show that, under an equivariant generation kernel, policy equivariance error is bounded by source equivariance error and is preserved exactly by invertible equivariant flows. We characterize all compatible conditional sources through a frame-relative normal form: for robot actions in world coordinates, translation requires observation dependence, whereas rotation admits the unique observation-independent normalized Haar source. Guided by this characterization, GARLIC combines an end-effector-anchored translational source, a Haar-distributed rotational source, and an equivariant GATr transport. Across our RLBench and real-world evaluations, GARLIC achieves the highest success rates and shows strong sample efficiency, parameter efficiency, and OOD generalization.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs and implement methods. We have not used generative AI tools to provide critical ingredients for proving mathematical claims, help develop theoretical models or conceptual frameworks, generate synthetic data sets, formulate mathematical claims, design or provide feedback on research methodology or experiments, interpret results. Propose or refine hypotheses, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to create or modify scientific figures or images, creation of artifacts, summarize or analyse existing literature, and edit a research paper to improve readability. We have reviewed all AI-assisted work. The authors reviewed the implementation and experimental procedures, verified the reported results, and checked the manuscript for accuracy. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We describe the experimental settings and implementation details to support reproducibility. The theoretical setup and assumptions are stated in Sections 3 and 4, with proofs provided in Appendix A. Section 5 describes the source construction and policy, while Appendix C details the observation embeddings, network architecture, training objective, and inference procedure. Architecture and optimization hyperparameters are summarized in Table 7. Appendix D.1 specifies the demonstration distribution, controlled comparisons, and evaluation procedure for the planar experiment. Section 6 and Appendix D.2 describe the RLBench tasks, demonstration budgets, baselines, and evaluation protocol. Appendix D.3 documents the hardware, demonstration collection, preprocessing, baseline training, closed-loop execution, and task success criteria for the real-world experiments. Supplementary videos illustrate the simulation and real-world evaluations. The supplementary code includes the model, source, objective, and sampler implementations, along with an end-to-end equivariance test in which the anchored source reaches the finite-sample error floor. The full codebase, training data, and trained checkpoints will be released with the paper upon acceptance.

REFERENCES

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Benjamin Bloem-Reddy and Yee Whye Teh. Probabilistic symmetries and invariant neural networks. *Journal of Machine Learning Research*, 21(90):1–61, 2020.

Avishek Joey Bose, Tara Akhound-Sadegh, Guillaume Huguet, Kilian Fatras, Jarrid Rector-Brooks, Cheng-Hao Liu, Andrei Cristian Nica, Maksym Korablyov, Michael Bronstein, and Alexander Tong. SE(3)-stochastic flow matching for protein backbone generation. In *International Conference on Learning Representations (ICLR)*, 2024. Spotlight; arXiv:2310.02391.

Johann Brehmer, Pim De Haan, Sönke Behrends, and Taco Cohen. Geometric algebra transformer. *Advances in neural information processing systems*, 36:35472–35496, 2023.

Jianlei Chang, Ruofeng Mei, Wei Ke, and Xiangyu Xu. EfficientFlow: Efficient equivariant flow policy learning for embodied AI. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. arXiv:2512.02020.

Kaiqi Chen, Eugene Lim, Kelvin Lin, Yiyang Chen, and Harold Soh. Don't start from scratch: Behavioral refinement via interpolant-based policy diffusion. In *Robotics: Science and Systems*, 2024a.

Ziyu Chen, Markos A. Katsoulakis, and Benjamin J. Zhang. Robustness and structure preservation in flow-based generative models via Wasserstein path-space divergences. *arXiv preprint arXiv:2410.01244*, 2024b. arXiv:2410.01244.

Eugenio Chisari, Nick Heppert, Max Argus, Tim Welschehold, Thomas Brox, and Abhinav Valada. Learning robotic manipulation policies from point clouds with conditional flow matching. *arXiv preprint arXiv:2409.07343*, 2024.

Rob Cornish. Stochastic neural network symmetrisation in markov categories. *arXiv preprint arXiv:2406.11814*, 2024.

Imre Csiszár. On information-type measure of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

Meipo Dai, Qiyuan Zhuang, He-Yang Xu, Ying-Jie Shuai, Yijun Wang, Qi Dou, and Xiu-Shen Wei. Where should action generation begin? a learnable source prior for generative robot policies. *arXiv preprint arXiv:2606.17408*, 2026.

Zibin Dong, Yicheng Liu, Yinchuan Li, Hang Zhao, and Jianye Hao. Conditioning matters: Training diffusion policies is faster than you think. *arXiv preprint arXiv:2505.11123*, 2025.

Leo Dorst and Steven De Keninck. A guided tour to the plane-based geometric algebra pga. *URL https://bivector. net/PGA4CS. html*, 2, 2022.

Nadav Dym, Hannah Lawrence, and Jonathan W. Siegel. Equivariant frames and the impossibility of continuous canonicalization. In *International Conference on Machine Learning*, 2024.

Morris L Eaton. Group invariance applications in statistics. IMS, 1989.

Clinton Enwerem, John S. Baras, and Calin Belta. EquiDexFlow: Contact-grounded SE(3)-equivariant dexterous grasp generative flows. *arXiv preprint arXiv:2606.12728*, 2026.

Gerald B Folland. *A course in abstract harmonic analysis*. CRC press, 2016.

Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Advances in neural information processing systems*, 33:1970–1981, 2020.

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

Thien-Loc Ha, Quang-Tan Nguyen, Trong-Bao Ho, Long Dinh, Minh Duc Nguyen, Gia-Binh Nguyen, Pham Tri Quang, Minh N. Vu, Duy M. H. Nguyen, An Thai Le, and Ngo Anh Vien. EquiVLA: A general framework for rotationally equivariant vision-language-action models. *arXiv preprint arXiv:2606.19784*, 2026.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pp. 8867–8887. PMLR, 2022.

Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. Equivariance with learned canonicalization functions. In *International Conference on Machine Learning (ICML), PMLR 202*, pp. 15546–15566, 2023.

Olav Kallenberg. *Foundations of modern probability*. Springer, 1997.

Sinjae Kang, Chanyoung Kim, Kaixin Wang, Li Zhao, and Kimin Lee. WarmPrior: Straightening flow-matching policies with temporal priors. *arXiv preprint arXiv:2605.13959*, 2026.

Isay Katsman, Aaron Lou, Derek Lim, Qingxuan Jiang, Ser-Nam Lim, and Christopher De Sa. Equivariant manifold flows. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. arXiv:2107.08596.

Junwan Kim, Jiho Park, Seonghu Jeon, and Seungryong Kim. Better source, better flow: Learning condition-dependent source distribution for flow matching. *arXiv preprint arXiv:2602.05951*, 2026.

Leon Klein, Andreas Krämer, and Frank Noé. Equivariant flow matching. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 59886–59910, 2023. arXiv:2306.15030.

Jonas Köhler, Leon Klein, and Frank Noé. Equivariant flows: exact likelihood generative learning for symmetric densities. In *International conference on machine learning*, pp. 5361–5370. PMLR, 2020.

Hannah Lawrence, Vasco Portilheiro, Yan Zhang, and Sékou-Oumar Kaba. Improving equivariant networks with probabilistic symmetry breaking. *arXiv preprint arXiv:2503.21985*, 2025.

Sang-gil Lee, Heeseung Kim, Chaehun Shin, Xu Tan, Chang Liu, Qi Meng, Tao Qin, Wei Chen, Sungroh Yoon, and Tie-Yan Liu. PriorGrad: Improving conditional denoising diffusion models with data-dependent adaptive prior. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2106.06406.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Minghua Liu, Xuanlin Li, Zhan Ling, Yangyan Li, and Hao Su. Frame mining: a free lunch for learning robotic manipulation from 3d point clouds. In *Conference on Robot Learning*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Bruno Machado, Alexandre Chapin, Emmanuel Dellandrea, and Liming Chen. Flowing with purpose: Latent action guided flow matching policies for robotic manipulation. *arXiv preprint arXiv:2606.23420*, 2026.

Omkar Patil, Ondrej Biza, Thomas Weng, Karl Schmeckpeper, Wil Thomason, Xiaohan Zhang, Kausik Sivakumar, Robin Walters, Nakul Gopalan, Sebastian Castro, Stephen Hart, and Eric Rosen. You've got a golden ticket: Improving generative robot policies with a single noise vector. *arXiv preprint arXiv:2603.15757*, 2026.

Omri Puny, Matan Atzmon, Heli Ben-Hamu, Ishan Misra, Aditya Grover, Edward J. Smith, and Yaron Lipman. Frame averaging for invariant and equivariant network design. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2110.03336.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

Hyunwoo Ryu, Jiwoo Kim, Hyunseok An, Junwoo Chang, Joohwan Seo, Taehan Kim, Yubin Kim, Chaewon Hwang, Jongeun Choi, and Roberto Horowitz. Diffusion-EDFs: Bi-equivariant denoising generative modeling on SE(3) for visual robotic manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18007–18018, 2024. Highlight; arXiv:2309.02685.

Kusha Sareen, Daniel Levy, Arnab Kumar Mondal, Sékou-Oumar Kaba, Tara Akhound-Sadegh, and Siamak Ravanbakhsh. Symmetry-aware generative modeling through learned canonicalization. *arXiv preprint arXiv:2501.07773*, 2025. NeurReps 2024 workshop version.

Zakhar Shumaylov, Peter Zaika, James Rowbottom, Ferdia Sherry, Melanie Weber, and Carola-Bibiane Schönlieb. Lie algebra canonicalization: Equivariant neural operators under arbitrary Lie groups. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.02698.

Xiatao Sun, Yuxuan Wang, Shuo Yang, Yinxing Chen, and Daniel Rakita. Hybrid diffusion policies with projective geometric algebra for efficient robot manipulation learning. *arXiv preprint arXiv:2507.05695*, 2025.

Chenrui Tie, Yue Chen, Ruihai Wu, Boxuan Dong, Zeyi Li, Chongkai Gao, and Hao Dong. Etseed: Efficient trajectory-level se (3) equivariant diffusion policy. In *International Conference on Learning Representations*, volume 2025, pp. 60114–60132, 2025.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research (TMLR)*, 2024. arXiv:2302.00482.

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

Keqin Wang, Tao Zhong, David Chang, and Christine Allen-Blanchette. Local-canonicalization equivariant graph neural networks for sample-efficient and generalizable swarm robot control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2026a. arXiv:2509.14431.

Xuyang Wang, Xinzhe Zhou, Xiaoming Duan, and Jianping He. SE(3)-equivariant flow matching with Gaussian process priors for geometric trajectory prediction. In *International Conference on Machine Learning (ICML)*, 2026b.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=PzcvxEMzvQC.

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

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

Jason Yim, Brian L. Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. SE(3) diffusion model with application to protein backbone generation. In *International Conference on Machine Learning (ICML)*, 2023. arXiv:2302.02277.

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

He Zhang, Ying Sun, Pengteng Li, Ziyang Chen, Yiren Zhao, Ziyang Rao, Weiyu Guo, Yandong Guo, and Hui Xiong. Source-lifted flow matching for intervenable multimodal imitation. *arXiv preprint arXiv:2607.10206*, 2026a.

Qinglun Zhang, Shen Cheng, Tian Dan, Haoqiang Fan, Guanghui Liu, and Shuaicheng Liu. Efficient hybrid SE(3)-equivariant visuomotor flow policy via spherical harmonics for robot manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026b. Highlight; arXiv:2603.23227.

Zhiyuan Zhang and Yu She. EquiForm: Noise-robust SE(3)-equivariant policy learning from 3D point clouds. *arXiv preprint arXiv:2601.17486*, 2026.

Haibo Zhao, Dian Wang, Yizhe Zhu, Xupeng Zhu, Owen Lewis Howell, Linfeng Zhao, Yaoyao Qian, Robin Walters, and Robert Platt. Hierarchical equivariant policy via frame transfer. In *Forty-second International Conference on Machine Learning*, 2025.

Tao Zhong and Christine Allen-Blanchette. Gagrasp: Geometric algebra diffusion for dexterous grasping. *arXiv preprint arXiv:2503.04123*, 2025.

Cai Zhou, Zijie Chen, Zian Li, Jike Wang, Kaiyi Jiang, Pan Li, Rose Yu, Muhan Zhang, Stephen Bates, and Tommi Jaakkola. Rethinking diffusion models with symmetries through canonicalization with applications to molecular graph generation. *arXiv preprint arXiv:2602.15022*, 2026a.

Xinzhe Zhou, Xuyang Wang, Xiaoming Duan, and Jianping He. BridgeFlow: Fast and robust SE(2)-equivariant motion planning with flow matching. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2026b. arXiv:2607.14725.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5745–5753, 2019.

Xupeng Zhu, Fan Wang, Robin Walters, and Jane Shi. SE(3)-equivariant diffusion policy in spherical fourier space. In *International Conference on Machine Learning (ICML)*, 2025. arXiv:2507.01723.

Yifeng Zhu, Abhishek Joshi, Peter Stone, and Yuke Zhu. Viola: Imitation learning for vision-based manipulation with object proposal priors. *arXiv preprint arXiv:2210.11339*, 2022. doi: 10.48550/arXiv.2210.11339.

A Proofs for Section 4

A.1 Proof of Theorem 1

Theorem 1 (Equivariance-error propagation) Assume A2. Then for every $\rho \in SE(3)$ and observation O: *(i)* $\mathcal{E}_{\rho,O}[\pi_\theta] \le \mathcal{E}_{\rho,O}[p_0]$; *(ii)* if K_O is induced by an invertible map, as for an exact ODE flow, then $\mathcal{E}_{\rho,O}[\pi_\theta] = \mathcal{E}_{\rho,O}[p_0]$.

Proof. Write $\mu = p_0(\cdot \mid \rho \cdot O)$ and $\nu = \rho_{\#}\, p_0(\cdot \mid O)$ for the two measures compared by $\mathcal{E}_{\rho,O}[p_0]$, and write $Km = \int K_{\rho\cdot O}(\cdot \mid z)\, m(dz)$ for a measure m. For every measurable set B, by definition $\pi_\theta(B \mid \rho \cdot O) = \int K_{\rho\cdot O}(B \mid z)\, p_0(dz \mid \rho \cdot O) = K\mu(B)$, and

$$
\begin{aligned}
\rho_{\#}\,\pi_\theta(B \mid O) &= \pi_\theta(\rho^{-1} \cdot B \mid O) && \big((\rho_{\#}m)(B) = m(\rho^{-1}\cdot B)\big) \\
&= \int K_O(\rho^{-1} \cdot B \mid z)\, p_0(dz \mid O) \\
&= \int K_{\rho\cdot O}(\rho \cdot \rho^{-1} \cdot B \mid \rho \cdot z)\, p_0(dz \mid O) && \text{(A2 condition)} \\
&= \int K_{\rho\cdot O}(B \mid \rho \cdot z)\, p_0(dz \mid O) \\
&= \int K_{\rho\cdot O}(B \mid z')\, \big(\rho_{\#}\, p_0(\cdot \mid O)\big)(dz') && \text{(pushforward change of variables)} \\
&= K\nu(B).
\end{aligned}
\tag{6}
$$

The two measures compared by $\mathcal{E}_{\rho,O}[\pi_\theta]$ are therefore $K\mu$ and $K\nu$.

(i) Total variation does not increase under a common Markov kernel, $\mathrm{TV}(K\mu, K\nu) \le \mathrm{TV}(\mu, \nu)$, which is the data-processing inequality (Csiszár, 1967).

(ii) Suppose $K_O(\cdot \mid z) = \delta_{\Phi_O(z)}$ for a bimeasurable bijection Φ_O. By A2, $K_{\rho\cdot O}$ is induced by the bimeasurable bijection $\Phi_{\rho\cdot O}$ satisfying $\Phi_{\rho\cdot O}(\rho \cdot z) = \rho \cdot \Phi_O(z)$. Thus, for the common kernel K above, the data-processing inequality holds with equality, $\mathrm{TV}(K\mu, K\nu) = \mathrm{TV}(\mu, \nu)$ (Csiszár, 1967; Liese & Vajda, 2006). This covers exact ODE flows whenever the ODE is globally well posed both forward and backward on $[0, 1]$. Existence and uniqueness make the time-one map Φ_O bijective with inverse given by the backward flow, while continuous dependence makes both maps continuous and hence measurable. □

Remark 1 (Beyond total variation)**.** *According to Csiszár (1967); Liese & Vajda (2006), Theorem 1 can be extended to any f-divergence in place of* TV*. In Corollary 1, we use the KL case.*

A.2 Proof of Corollary 1

Corollary 1 (Gaussian source equivariance error) For every $\rho = (R, v) \in SE(3)$, $\mathcal{E}^{\mathrm{KL}}_{\rho,O}[p_0] = \mathrm{KL}(p_0 \,\|\, \rho_{\#}p_0) = {}^{\|v\|^2}\!/_{2\sigma^2}$. Under an exact equivariant flow, the policy has the same KL equivariance error. The error depends only on translation and vanishes iff $v = 0$; thus, the fixed Gaussian is equivariant only to rotations about its center.

Proof. By definition, $\mathcal{E}^{\mathrm{KL}}_{\rho,O}[p_0] = \mathrm{KL}\big(p_0(\cdot \mid \rho \cdot O) \,\|\, \rho_{\#}\, p_0(\cdot \mid O)\big)$. Since p_0 does not depend on O, both arguments reduce to p_0 and $\rho_{\#}\, p_0$,

$$
\mathcal{E}^{\mathrm{KL}}_{\rho,O}[p_0] = \mathrm{KL}(p_0 \,\|\, \rho_{\#}\, p_0), \tag{7}
$$

so it remains to compute $\rho_{\#}\, p_0$. In the coordinates $z = (z_p, z_r, z_g)$ of Section 3, $p_0 = \mathcal{N}(0, \sigma^2 I)$ is the product of centered isotropic Gaussians on the three blocks, and $\rho = (R, v)$ acts blockwise, by $z_p \mapsto Rz_p + v$, by an orthogonal linear map Q on z_r of dimension d, and trivially on z_g. The pushforward of a product measure under a blockwise map is the product of the blockwise pushforwards,

$$
(z_p \mapsto Rz_p + v)_{\#}\, \mathcal{N}(0, \sigma^2 I_3) = \mathcal{N}(v, \sigma^2 RR^\top) = \mathcal{N}(v, \sigma^2 I_3), \tag{8}
$$

$$
Q_{\#}\, \mathcal{N}(0, \sigma^2 I_d) = \mathcal{N}(0, \sigma^2 QQ^\top) = \mathcal{N}(0, \sigma^2 I_d), \tag{9}
$$

and the gripper block is unchanged, so $\rho_{\#}\, p_0$ differs from p_0 only in the translational block. Since KL is additive over independent blocks and vanishes on identical ones,

$$\mathrm{KL}(p_0 \,\|\, \rho_{\#}\, p_0) = \mathrm{KL}\big(\mathcal{N}(0, \sigma^2 I_3) \,\|\, \mathcal{N}(v, \sigma^2 I_3)\big) = \frac{\|v\|^2}{2\sigma^2}. \tag{10}$$

The value depends on ρ only through $\|v\|$ and vanishes iff $v = 0$, that is, iff ρ is a rotation about the origin, the center of the source. By Theorem 1(ii) in its KL form (Remark 1), the policy of an exact equivariant flow has the same KL error at every (ρ, O). □

A.3 Proof of Proposition 1

Proposition 1 (Canonical-frame characterization) Let $\mathcal{C}_\tau = \{O : \tau(O) = e\}$ denote the *canonical slice*, where e is the identity element of $SE(3)$. A source distribution p_0 satisfies A1 if and only if there exists a Markov kernel $\tilde{q}$ on $\mathcal{C}_\tau$ with

$$p_0(\cdot \mid O) \;=\; \tau(O)_{\#}\, \tilde{q}\big(\cdot \mid M_\tau(O)\big) \qquad \text{for all } O. \tag{11}$$

For a fixed frame τ, the kernel $\tilde{q}$ is unique.

Proof. Preliminaries. In our setting, observations and actions lie in finite products of Euclidean and pose spaces. These are standard Borel spaces, and the usual $SE(3)$ actions on them are continuous and hence jointly measurable. Recall that τ is an equivariant Borel frame and that $M_\tau(O) = \tau(O)^{-1} \cdot O$. Since inversion in $SE(3)$ is continuous, M_τ is Borel. Moreover, $\mathcal{C}_\tau$ is Borel because it is the preimage of the closed set $\{e\}$ under τ. Frame equivariance gives $\tau(M_\tau(O)) = \tau(O)^{-1}\tau(O) = e$, so $M_\tau(O) \in \mathcal{C}_\tau$. Finally, $M_\tau(\rho{\cdot}O) = (\rho\,\tau(O))^{-1}{\cdot}\rho{\cdot}O = M_\tau(O)$, so M_τ is invariant to global rigid motions.

Sufficiency: the representation implies A1. Suppose the stated representation holds. For every ρ,

$$\begin{aligned} p_0(\cdot \mid \rho \cdot O) &= \tau(\rho \cdot O)_{\#}\, \tilde{q}\big(\cdot \mid M_\tau(\rho \cdot O)\big) \\ &= (\rho\,\tau(O))_{\#}\, \tilde{q}\big(\cdot \mid M_\tau(O)\big) \qquad \text{(equivariance of } \tau\text{; invariance of } M_\tau) \\ &= \rho_{\#}\, p_0(\cdot \mid O), \end{aligned} \tag{12}$$

which is A1 in pushforward form.

Necessity: A1 implies the representation. Suppose p_0 satisfies A1 and define $q(\cdot \mid O) = (\tau(O)^{-1})_{\#}\, p_0(\cdot \mid O)$, that is, $q(A \mid O) = \int \mathbf{1}_A(\tau(O)^{-1} \cdot a)\, p_0(da \mid O)$. Since τ and the group action are Borel, q is a Markov kernel by standard measurability of parameterized integrals (Kallenberg, 1997). For every measurable A, every ρ, and every O,

$$\begin{aligned} q(A \mid \rho \cdot O) &= p_0\big(\tau(\rho \cdot O) \cdot A \mid \rho \cdot O\big) \\ &= p_0\big(\rho \cdot (\tau(O) \cdot A) \mid \rho \cdot O\big) \quad \text{(frame equivariance)} \\ &= p_0\big(\tau(O) \cdot A \mid O\big) \qquad\quad \text{(A1 on } \tau(O) \cdot A) \\ &= q(A \mid O), \end{aligned} \tag{13}$$

Thus, q is constant along each $SE(3)$ orbit. Since this identity holds for every pair (ρ, O), we may set $\rho = \tau(O)^{-1}$, obtaining $q(\cdot \mid O) = q(\cdot \mid M_\tau(O))$. Hence the pulled-back source depends only on the canonicalized observation, not on its original world pose. Let $\tilde{q}$ be the restriction of q to $\mathcal{C}_\tau$, a Markov kernel on the slice. Then

$$\begin{aligned} \tau(O)_{\#}\, \tilde{q}\big(\cdot \mid M_\tau(O)\big) &= \tau(O)_{\#}\, q\big(\cdot \mid M_\tau(O)\big) \\ &= \tau(O)_{\#}\, q(\cdot \mid O) \\ &= \tau(O)_{\#}\, (\tau(O)^{-1})_{\#}\, p_0(\cdot \mid O) \\ &= p_0(\cdot \mid O). \end{aligned} \tag{14}$$

Uniqueness. If $\tilde{q}_1$ and $\tilde{q}_2$ both represent p_0, then for $c \in \mathcal{C}_\tau$ we have $\tau(c) = e$ and $M_\tau(c) = c$, so $\tilde{q}_i(\cdot \mid c) = p_0(\cdot \mid c)$ for both i. The representing kernel is therefore determined on the slice by p_0 itself. Hence, for a fixed frame τ, the correspondence between compatible sources and canonical kernels is one-to-one rather than merely existential. □

Relation to canonicalization. Constructing an equivariant conditional distribution by canonicalizing the observation, defining a distribution in canonical coordinates, and transforming it back is classical (Eaton, 1989; Bloem-Reddy & Teh, 2020). Proposition 1 goes further: for a fixed equivariant frame τ, every A1-compatible source must arise in this way, and its canonical kernel is unique.

To place this characterization in context, the related result of Bloem-Reddy & Teh (2020) is formulated for a compact group and an invariant input law. Their result is distributional, characterizing equivariance relative to a group-invariant distribution of inputs. This framework does not directly cover $SE(3)$ policies, since $SE(3)$ contains unbounded translations and is therefore noncompact, while no probability distribution over absolute world-frame positions can remain invariant under every translation. We instead impose A1 pointwise for every observation and use proprioception to supply the equivariant frame, so neither compactness nor an invariant observation law is required.

Factorized sources. The main text reduces the design of a compatible source to its translational, rotational, and gripper factors. The following reduction is exact for product sources.

Lemma 1 (Blockwise reduction)**.** *Decompose the action as $a = (p^a, R^a, g^a)$ and write $\rho = (R, v)$, where $R \in SO(3)$ and $v \in \mathbb{R}^3$. Suppose that for every O the source is a product, $p_0(\cdot \mid O) = p_p(\cdot \mid O) \otimes p_R(\cdot \mid O) \otimes p_g(\cdot \mid O)$. Then p_0 satisfies A1 iff*

$$\begin{aligned} p_p(\cdot \mid \rho \cdot O) &= (x \mapsto Rx + v)_{\#}\, p_p(\cdot \mid O), \\ p_R(\cdot \mid \rho \cdot O) &= (R^a \mapsto RR^a)_{\#}\, p_R(\cdot \mid O), \\ p_g(\cdot \mid \rho \cdot O) &= p_g(\cdot \mid O) \end{aligned} \tag{15}$$

for all ρ and all O.

Proof. The action $\rho \cdot a = (Rp^a + v,\, RR^a,\, g^a)$ is the product of the three block actions, so the pushforward of a product measure is the product of the blockwise pushforwards. If Eq. 15 holds, $\rho_{\#}\, p_0(\cdot \mid O)$ is the product of the three transformed factors, which is $p_0(\cdot \mid \rho \cdot O)$. Conversely, if A1 holds, each marginal of $p_0(\cdot \mid \rho \cdot O)$ equals the corresponding marginal of $\rho_{\#}\, p_0(\cdot \mid O)$, which is the blockwise pushforward of the marginal of $p_0(\cdot \mid O)$, and the marginals of a product are its factors. □

A.4 Proof of Proposition 2

Proposition 2 (Observation-independent compatibility) For observation-independent factors, (i) no distribution on $\mathbb{R}^3$ is translation-equivariant; (ii) the normalized Haar measure is the *unique* rotation-equivariant distribution on $SO(3)$; and (iii) *every* gripper distribution is equivariant.

Proof of Proposition 2. When a factor does not depend on the observation, Eq. 15 simply asks whether its distribution is unchanged—that is, invariant—under the corresponding group action.

(i) Suppose μ is a probability measure on $\mathbb{R}^3$ satisfying $(x \mapsto x + v)_{\#}\, \mu = \mu$ for every $v \in \mathbb{R}^3$. Let $C_k = k + [0, 1)^3$ for $k \in \mathbb{Z}^3$; these cubes partition $\mathbb{R}^3$. Translation invariance gives $\mu(C_k) = c$ for every k. Countable additivity would then give $1 = \mu(\mathbb{R}^3) = \sum_{k \in \mathbb{Z}^3} c$, but this sum is zero if $c = 0$ and infinite if $c > 0$. Thus no observation-independent translational factor can satisfy A1, even for pure translations.

(ii) Necessity: compatibility implies Haar. Suppose an observation-independent rotational factor satisfies Eq. 15, and denote its fixed law by κ, so $p_R(\cdot \mid O) = \kappa$ for every O. Then Eq. 15 gives $(R^a \mapsto RR^a)_{\#}\, \kappa = \kappa$ for every $R \in SO(3)$. Thus, κ is left-invariant. Since $SO(3)$ is compact, uniqueness of the left-invariant Borel probability measure implies that κ is the normalized Haar measure (Folland, 2016).

Sufficiency: Haar is compatible. Conversely, let κ be the normalized Haar measure on $SO(3)$ and set $p_R(\cdot \mid O) = \kappa$ for every O. Left invariance gives $(R^a \mapsto RR^a)_{\#}\, \kappa = \kappa$ for every $R \in SO(3)$. Hence both sides of the rotational condition in Eq. 15 equal κ, so the condition holds.

(iii) Since $SE(3)$ acts trivially on g^a, the condition in Eq. 15 reduces to $\kappa_g = \kappa_g$ for any observation-independent gripper law κ_g. Hence every gripper distribution is compatible. □

Haar rotations from the 6D representation. The rotational source in Eq. 5 is defined in the 6D representation rather than directly on $SO(3)$. At execution, Gram-Schmidt maps these coordinates to a rotation. The following corollary connects this construction to Proposition 2(ii).

Corollary 3 (Haar law of the 6D rotational source)**.** *Let $\sigma > 0$ and let $r = [r_1, r_2] \sim \mathcal{N}(0, \sigma^2 I_6)$ be the rotational factor in Eq. 5. Let $\mathrm{GS}(r_1, r_2) \in SO(3)$ denote the standard 6D-to-rotation map (Zhou et al., 2019), which is well-defined almost surely. Then $\mathrm{GS}_{\#}\,\mathcal{N}(0, \sigma^2 I_6)$ is the normalized Haar measure on $SO(3)$. Thus, its induced law on $SO(3)$ is the unique observation-independent compatible rotation law characterized by Proposition 2(ii).*

Proof. Proper rotations preserve norms and inner products and transform cross products equivariantly, so $\mathrm{GS}(Rr_1, Rr_2) = R\,\mathrm{GS}(r_1, r_2)$ for every $R \in SO(3)$. Write $\gamma = \mathcal{N}(0, \sigma^2 I_6)$. The block-diagonal map $R \oplus R\colon (r_1, r_2) \mapsto (Rr_1, Rr_2)$ is orthogonal on $\mathbb{R}^6$, so it leaves γ invariant. For the pushforward $\nu = \mathrm{GS}_{\#}\,\gamma$ and a Borel set $A \subseteq SO(3)$, we therefore have

$$\nu(RA) = \gamma\big(\mathrm{GS}^{-1}(RA)\big) = \gamma\big((R \oplus R)\,\mathrm{GS}^{-1}(A)\big) = \gamma\big(\mathrm{GS}^{-1}(A)\big) = \nu(A). \tag{16}$$

Thus, ν is a left-invariant probability measure on $SO(3)$ and equals the Haar measure by part (ii) of the proposition. This is the QR-of-Gaussian construction of Haar-distributed rotations restricted to two columns. □

This corollary supplies the 6D-to-Haar link used in the discussion following Proposition 2 and in the source construction of Section 5.1.

Rotation coordinates and compatibility. Proposition 2(ii) concerns the law induced on $SO(3)$, rather than the distribution in a chosen coordinate space. Coordinate-space isotropy alone therefore does not ensure observation-independent compatibility. A non-Haar law can still define a compatible anchored source through Corollary 2(ii). By contrast, the isotropic 6D Gaussian is invariant when both 3D blocks are rotated together, and Corollary 3 shows that its Gram-Schmidt pushforward is exactly Haar. It is therefore compatible while remaining observation-independent.

A.5 Proof of Corollary 2

Corollary 2 (Equivariantly anchored sources) Let $b(O) \in \mathbb{R}^3$ and $m(O) \in SO(3)$ be equivariant anchors, satisfying $b(\rho \cdot O) = R\,b(O) + v$ and $m(\rho \cdot O) = R\,m(O)$. *(i)* For any fixed rotation-invariant law κ_T on $\mathbb{R}^3$, $b(O) + \epsilon$ with $\epsilon \sim \kappa_T$ defines a compatible translation source. *(ii)* For any fixed law κ_R on $SO(3)$, $m(O)\,S$ with $S \sim \kappa_R$ defines a compatible rotation source.

Proof. *(i)* Fix $\rho = (R, v)$, let $\epsilon \sim \kappa_T$ with $Q_{\#}\,\kappa_T = \kappa_T$ for every $Q \in SO(3)$, and write $p_p(\cdot \mid O) = \mathrm{Law}(b(O) + \epsilon)$. Then

$$\begin{aligned}
(x \mapsto Rx + v)_{\#}\,p_p(\cdot \mid O) &= \mathrm{Law}\left(R\,b(O) + v + R\epsilon\right) \\
&= \mathrm{Law}\left(R\,b(O) + v + \epsilon\right) && (\kappa_T \text{ is rotation-invariant}) \\
&= \mathrm{Law}\left(b(\rho \cdot O) + \epsilon\right) && (\text{equivariance of } b) \\
&= p_p(\cdot \mid \rho \cdot O).
\end{aligned} \tag{17}$$

This is the translational condition in Lemma 1.

(ii) Fix $\rho = (R, v)$, let $S \sim \kappa_R$, and write $p_R(\cdot \mid O) = \mathrm{Law}(m(O)\,S)$. Then

$$\begin{aligned}
(R^a \mapsto RR^a)_{\#}\,p_R(\cdot \mid O) &= \mathrm{Law}\left(R\,m(O)\,S\right) \\
&= \mathrm{Law}\left(m(\rho \cdot O)\,S\right) && (\text{equivariance of } m) \\
&= p_R(\cdot \mid \rho \cdot O).
\end{aligned} \tag{18}$$

This is the rotational condition in Lemma 1; no property of κ_R is used. □

B A PRIMER ON PROJECTIVE GEOMETRIC ALGEBRA

B.1 PROJECTIVE GEOMETRIC ALGEBRA AND GEOMETRIC OBJECTS

GARLIC represents point-cloud locations and end-effector poses as Euclidean points and oriented planes. We therefore use the plane-based projective geometric algebra (PGA) $\mathbb{G}_{3,0,1}$ (Dorst & De Keninck, 2022), in which these objects and their rigid motions share a 16-dim representation.

The basis vectors e_0, e_1, e_2, e_3 satisfy $e_0^2 = 0$, $e_1^2 = e_2^2 = e_3^2 = 1$, and $e_i e_j = -e_j e_i$ for $i \neq j$. Their ordered products $e_A = e_{i_1} \cdots e_{i_k}$ are basis blades of grade k. A multivector is a linear combination $x = \sum_A x_A e_A$, and $\langle x \rangle_k$ denotes its grade-k part. For homogeneous multivectors a and b of grades k and l, respectively, the exterior product is $a \wedge b = \langle ab \rangle_{k+l}$.

Planes are vectors and points are trivectors in this model. Let $\pi(n, q)$ represent the Euclidean plane $\{x : n \cdot (x - q) = 0\}$ through $q \in \mathbb{R}^3$ with normal $n \in \mathbb{R}^3$, and let $P(p)$ represent the finite point $p = (p_x, p_y, p_z)$. Under our sign convention,

$$\pi(n, q) := -(n \cdot q)e_0 + \sum_{i=1}^{3} n_i e_i, \qquad P(p) := e_{123} - p_x e_{023} + p_y e_{013} - p_z e_{012}. \tag{19}$$

These representations are projective: nonzero scalar multiples encode the same object. Fixing the e_{123} coefficient to one is only a convenient input gauge for finite points; any representative with a nonzero e_{123} coefficient can be rescaled to this gauge.

B.2 RIGID MOTIONS AS VERSORS

A normalized plane vector represents reflection in that plane, and products of such vectors, called versors, compose reflections into rigid motions (Brehmer et al., 2023). For a multivector x, the reverse $\tilde{x}$ reverses the vector factors of each blade, while the grade involution $\hat{x}$ multiplies its grade-k part by $(-1)^k$. A normalized versor satisfies $u^{-1} = \tilde{u}$ and acts by

$$\rho_u(x) = \begin{cases} uxu^{-1}, & u \text{ even}, \\ u\hat{x}u^{-1}, & u \text{ odd}. \end{cases} \tag{20}$$

For an odd versor consisting of one plane vector, this is a reflection; composing reflections yields rotations and translations. In particular, every $\rho = (R, v) \in SE(3)$ has an even-versor representative M satisfying $\rho_M(P(p)) = P(Rp + v)$.

Products of normalized plane vectors form $\mathrm{Pin}(3, 0, 1)$, a double cover of $E(3)$, while even products form $\mathrm{Spin}(3, 0, 1)$, a double cover of $SE(3)$. The action in Eq. 20 is linear, grade-preserving, and an algebra homomorphism: $\rho_u(xy) = \rho_u(x)\rho_u(y)$. Thus geometric products and grade projections are equivariant, whereas scalar parts are invariant. These properties underlie GATr's equivariance.

B.3 EQUIVARIANT PRIMITIVES IN GATR

GATr builds equivariant transformer blocks from linear maps, geometric bilinear operations, gated nonlinearities, normalization, and attention (Brehmer et al., 2023). Each token contains multivector channels, which transform under ρ_u, and auxiliary invariant scalar channels. Its equivariant linear maps are spanned by the five grade projections $\langle x \rangle_k$, $k = 0, \ldots, 4$, and the four maps $e_0 \langle x \rangle_k$, $k = 0, \ldots, 3$; a multichannel layer learns their coefficients for each input–output channel pair.

Geometric bilinear layers. GATr introduces interactions between multivectors through the geometric product and the equivariant join. Write $I = e_{0123} := e_0 e_1 e_2 e_3$, and let x^* denote the right-complement dual, defined on basis blades by $e_A \wedge e_A^* = I$ and extended linearly. If z_{0123} is the pseudoscalar coefficient of a reference multivector z, the geometric bilinear layer is

$$\begin{aligned} \mathrm{EquiJoin}(x, y; z) &:= z_{0123}(x^* \wedge y^*)^*, \\ \mathrm{Geometric}(x, y; z) &:= \mathrm{Concat}_{\mathrm{channels}}\big(xy, \mathrm{EquiJoin}(x, y; z)\big). \end{aligned} \tag{21}$$

The geometric product is equivariant by the algebra-homomorphism property above. The reference coefficient supplies the sign needed by the join under an odd versor, so the combined operation is also equivariant when the reference transforms with the inputs.

Nonlinearities and normalization. Define the invariant inner product $\langle x, y\rangle_{\mathrm{GA}} := \langle \tilde{x} y\rangle_0$. GATr uses scalar-gated GELU, $\mathrm{GatedGELU}(x) := \mathrm{GELU}(\langle x\rangle_0)x$, and equivariant LayerNorm, $\mathrm{LayerNorm}(x) := x/\sqrt{\mathbb{E}_c[\langle x, x\rangle_{\mathrm{GA}}]}$, where the expectation is over multivector channels. Because the scalar part and inner product are invariant, both operations commute with the versor action.

Attention. Let Q, K, and V denote the multivector query, key, and value tensors, respectively; let i, i' index tokens, let c, c' index channels, and let n_c be the number of multivector channels. GATr defines

$$\mathrm{Attention}(Q, K, V)_{i'c'} := \sum_i \mathrm{Softmax}_i\left(\frac{\sum_c \langle Q_{i'c}, K_{ic}\rangle_{\mathrm{GA}}}{\sqrt{8n_c}}\right) V_{ic'}. \tag{22}$$

The inner products give invariant attention weights, whose weighted sums of multivector values remain equivariant. Auxiliary scalar channels use ordinary scalar operations and contribute standard dot products to the same attention weights.

C Method Details

C.1 Embedding the observation history

For an observation $O = (X, s)$, each point $x \in X$ is embedded with the point row as $P(x)$, yielding a one-channel token. Let the robot state $s = (T^{\mathrm{ee}}, g)$ contain the end-effector pose $T^{\mathrm{ee}} = (R^{\mathrm{ee}}, p^{\mathrm{ee}})$ and gripper state g. GARLIC writes the 6D rotation block as $r = (r_1, r_2)$ and sets $r_3 = r_1 \times r_2$. The point channel directly represents the position p^{ee}, while three planes through that point encode the orientation through their normals r_1, r_2, and r_3. Applying the point row once and the plane row three times gives the four multivector channels

$$\begin{aligned}\mathcal{E}(p^{\mathrm{ee}}, r_1, r_2) = \big(&e_{123} - p_x^{\mathrm{ee}} e_{023} + p_y^{\mathrm{ee}} e_{013} - p_z^{\mathrm{ee}} e_{012},\\ &- (r_1 \cdot p^{\mathrm{ee}}) e_0 + \textstyle\sum_{j=1}^3 (r_1)_j e_j,\\ &- (r_2 \cdot p^{\mathrm{ee}}) e_0 + \textstyle\sum_{j=1}^3 (r_2)_j e_j,\\ &- (r_3 \cdot p^{\mathrm{ee}}) e_0 + \textstyle\sum_{j=1}^3 (r_3)_j e_j\big).\end{aligned} \tag{23}$$

Thus channel 0 occupies the grade-3 coefficients highlighted by the point row of Table 6, while channels 1–3 occupy the grade-1 coefficients highlighted by the plane row; every gray coefficient is zero at input. The gripper state g is encoded separately in GATr's auxiliary scalar channels.

Figure 7 visualizes this four-channel embedding for the observed end-effector pose. Here, r_1 and r_2 are the first two columns of $R^{\mathrm{ee}} \in SO(3)$, and $r_3 = r_1 \times r_2$, so the three plane normals form an orthonormal orientation frame. The resulting pose embedding is $SE(3)$-equivariant. Similar point-and-plane pose embeddings have been used in Zhong & Allen-Blanchette (2025).

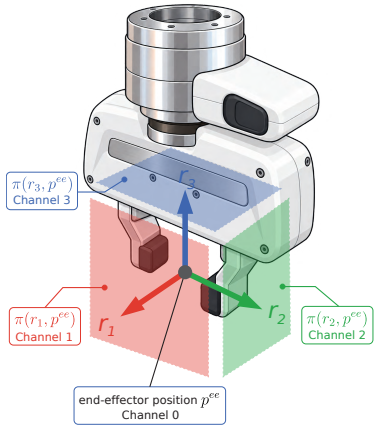

Figure 7: **Geometric interpretation of the pose embedding.** Channel 0 encodes the end-effector position and channels 1–3 encode the orientation. The plane patches are displaced from the point for visibility.

For an observation history, GARLIC applies the same embedding independently to each frame, and the resulting per-frame token sets are passed to the encoder.

Table 6: **The coefficients used by the plane and point embeddings in GARLIC**. Gripper state and other conditioning use GATr's auxiliary scalar channels, not the scalar multivector component.

	Scalar *grade 0*	**Vector** *grade 1*				**Bivector** *grade 2*		**Trivector** *grade 3*				**Pseudoscalar** *grade 4*
Object	1	e_0	e_1	e_2	e_3	e_{0i}	e_{ij}	e_{012}	e_{013}	e_{023}	e_{123}	e_{0123}
Plane $\pi(n, q)$	0	$-n\cdot q$	n_x	n_y	n_z	0	0	0	0	0	0	0
Point $P(p)$	0	0	0	0	0	0	0	$-p_z$	p_y	$-p_x$	1	0

C.2 Encoder and invariant downsampling

The encoder converts the observation history into a compact set of context tokens for the flow decoder. It encodes and progressively downsamples each frame's point cloud with shared weights, then assembles the resulting point and end-effector tokens across frames into a shared context set.

Encoder architecture. First, an equivariant linear layer lifts each point token into the encoder hidden representation. We then process the point tokens with GATr blocks (Brehmer et al., 2023), interleaved with invariant downsampling. Because point clouds have no natural ordering, we omit positional encodings within each frame to preserve permutation equivariance with respect to point order. The equivariant MLP in each block begins with the geometric bilinear layer in Eq. 21. For the join in this layer, we use GATr's data-derived reference (Brehmer et al., 2023), obtained for each frame by embedding the centroid of its raw point cloud.

Invariant downsampling. To keep the encoder equivariant under rigid transformations, the downsampling rule must be invariant to the scene's global pose. We implement this using farthest-point sampling (FPS) (Qi et al., 2017) on the original point coordinates, which we carry alongside the learned token features throughout the encoder. At each stage, we initialize FPS with the point farthest from the centroid of the current points. Greedy FPS then selects the point farthest from the selected set until the target number of tokens remains, and we use the selected indices to gather the corresponding multivector and scalar features.

The seed and subsequent selections depend only on Euclidean distances, which are preserved by rigid motions and unaffected by input order. Thus, in exact arithmetic and away from distance ties, FPS selects the same physical points regardless of global pose or point ordering. Distance ties require arbitrary tie-breaking and can therefore make the selected set depend on point order.

Context construction. We first map the surviving point tokens to the context representation using an equivariant linear layer. The end-effector token bypasses the self-attention stack and is mapped separately from the pose embedding in Eq. 23 to the same representation. We encode the observed gripper state with an MLP and pass the result as scalar input to this mapping. To distinguish observation frames, we transform a learned frame-index embedding and add it to the scalar channels of every token from the corresponding frame, leaving the multivector channels unchanged. The resulting context tokens serve as the keys and values for decoder cross-attention in Appendix C.3, where the point-cloud and end-effector information first interact.

C.3 Flow queries and cross-attention decoder

Given the context tokens from Appendix C.2, the decoder embeds each step of the current flow state as a query token. The queries attend to the shared context, interact through an equivariant temporal convolution, and are mapped to one output token per action step.

Flow-query embedding. Each action step is represented as $z = (z_p, r_1, r_2, z_g)$, where $z_p, r_1, r_2 \in \mathbb{R}^3$ and $z_g \in \mathbb{R}$. We form the source chunk z_0 by sampling each step from the source in Eq. 5, and z_1 is the demonstrated action chunk in the same coordinates. For a flow time t shared across the chunk, the straight path gives

$$z_t = (1 - t)z_0 + tz_1. \tag{24}$$

Each step of z_t becomes one decoder query token.

For step k, write $z_t^{(k)} = (z_p^{(k)}, r_1^{(k)}, r_2^{(k)}, z_g^{(k)})$. We embed its pose with the same four-channel map used for the observed end-effector pose, namely $\mathcal{E}(z_p^{(k)}, r_1^{(k)}, r_2^{(k)})$ in Eq. 23. Channel 0 is the point $P(z_p^{(k)})$, while channels 1–3 are the planes through $z_p^{(k)}$ with normals $r_1^{(k)}$, $r_2^{(k)}$, and $r_1^{(k)} \times r_2^{(k)}$. Although the interpolated rotation vectors need not be orthonormal, Eq. 23 remains $SE(3)$-equivariant for arbitrary plane normals. Using the same embedding places the observation and action geometry in compatible multivector channels for equivariant cross-attention.

The flow time t, step index k, and gripper coordinate $z_g^{(k)}$ enter only through the scalar channels. We encode t sinusoidally, k with a learned embedding, and $z_g^{(k)}$ with an MLP, then concatenate and project their features into the scalar channels. Because t, k, and $z_g^{(k)}$ are invariant under $SE(3)$, this conditioning preserves equivariance. Finally, an equivariant linear layer lifts the pose embedding and scalar features to the decoder representation.

Cross-attention decoder. We use a stack of cross-attention blocks. In each block, the action-step tokens form the query tensor Q, while the context tokens from Appendix C.2 form the key and value tensors K and V in Eq. 22. Each block applies equivariant layer normalization, geometric cross-attention, and a residual connection, with no MLP. We add no further positional encoding in cross-attention because frame and step identities are already encoded in the scalar channels.

The decoder has no self-attention among action steps. Conditioned on the shared context, the action steps are processed independently until they interact through the temporal convolution.

C.4 Equivariant temporal convolution

The cross-attention stack in Appendix C.3 processes the action-step tokens separately and therefore does not exchange information across steps. The temporal convolution couples these tokens to promote temporal coherence while preserving equivariance. It sits between the cross-attention stack and the output head of Appendix C.5.

Temporal block structure. After applying equivariant layer normalization, we process the multivector and scalar features in separate branches along the action-step axis. For the multivector path, let $x_c^{(k)}$ and $y_{c'}^{(k)}$ denote the input and output multivectors in channels c and c', respectively, at action step k. The convolution outputs

$$y_{c'}^{(k)} = \sum_c \sum_j w_{c'cj}\, x_c^{(k+jd)}, \tag{25}$$

where j ranges over the temporal kernel offsets, the learned weights $w_{c'cj} \in \mathbb{R}$ are real scalars, d is the dilation rate, and indices outside the action chunk are zero-padded. The weights carry no blade index, so the same temporal kernel is shared across all blade coefficients. We use neither a bias nor a nonlinearity in this branch. The scalar path uses a standard one-dimensional convolutional block along the action-step axis, with two convolutions separated by a GELU activation. Each branch output is then added to its corresponding input.

Equivariance. The multivector action in Eq. 20 is linear. Since the weights in Eq. 25 are real scalars shared across blade coefficients, applying the same rigid motion to every channel and action step gives

$$\rho_M\big(y_{c'}^{(k)}\big) = \sum_c \sum_j w_{c'cj}\, \rho_M\big(x_c^{(k+jd)}\big), \tag{26}$$

so the multivector convolution commutes with the rigid-motion action. An arbitrary multivector bias, a coefficient-wise nonlinearity, or a blade-dependent kernel would not generally commute with this action. Zero padding preserves equivariance because the zero multivector is fixed by ρ_M. The layer normalization is the equivariant one described in Appendix B.3, the scalar branch acts only on invariant features, and residual addition preserves equivariance. Hence each block, and therefore the entire module, commutes with the $SE(3)$ action applied simultaneously to all action steps.

C.5 Decoder output and endpoint prediction

This subsection describes how the output head maps each action-step token to a pose and gripper value, and how the resulting action is interpreted as a flow endpoint.

Decoder output head. After the temporal convolution, each action-step token passes through equivariant layer normalization and a residual two-layer geometric MLP with scalar-gated GELU. The MLP begins with the geometric bilinear layer in Eq. 21. The geometric bilinear layer requires a join reference. We use GATr's fixed canonical reference, the constant multivector $e_{123} + e_{0123}$, rather than the data-derived reference used by the encoder. Because this reference is fixed instead of transforming with the inputs, the output head and hence the decoder, is equivariant under proper rigid motions but not reflections; Appendix E.1 discusses this distinction. Finally, an equivariant linear layer produces four multivector channels and one scalar channel for the readback below.

Endpoint readback. For action step k, write the readout in the flow coordinates of Appendix C.3 as $\hat{z}_1^{(k)} = (\hat{z}_p^{(k)}, \hat{r}_1^{(k)}, \hat{r}_2^{(k)}, \hat{z}_g^{(k)})$, and let $x_A^{(\ell)}$ denote the coefficient of blade e_A in output multivector channel ℓ. We recover the position from the trivector coefficients of channel 0:

$$\hat{z}_p^{(k)} = \frac{1}{x_{123}^{(0)}}\big(-x_{023}^{(0)},\, x_{013}^{(0)},\, -x_{012}^{(0)}\big). \tag{27}$$

Because a PGA point is projective and the network does not constrain $x_{123}^{(0)}$ to one, this ratio is required for equivariant readback. We clamp the denominator away from zero for numerical stability. The Euclidean vector coefficients of channels 1 and 2 give the two vectors of the 6D rotation representation:

$$\hat{r}_1^{(k)} = \big(x_1^{(1)}, x_2^{(1)}, x_3^{(1)}\big), \qquad \hat{r}_2^{(k)} = \big(x_1^{(2)}, x_2^{(2)}, x_3^{(2)}\big). \tag{28}$$

The scalar channel directly gives the gripper value $\hat{z}_g^{(k)}$, without a further readout head. The network directly predicts the position, 6D rotation, and gripper state at the flow endpoint; together, they form $\hat{z}_1^{(k)}$ rather than a velocity.

C.6 Training and inference

Training objective. Guided by JiT (Li & He, 2025), the network predicts the endpoint $\hat{z}_1$ but is supervised in velocity space. With the shared remaining-time denominator $\delta(t) = \max(1-t, c_{\text{train}})$, we derive the predicted and demonstrated velocities as

$$\hat{v}_t = \frac{\hat{z}_1 - z_t}{\delta(t)}, \qquad v_t^\star = \frac{z_1 - z_t}{\delta(t)}. \tag{29}$$

We directly minimize their squared difference in velocity space:

$$\mathcal{L} = \|\hat{v}_t - v_t^\star\|_{\text{act}}^2 + \lambda_{\text{rep}} \underbrace{\left[\big(|x_{123}^{(0)}| - 1\big)^2 + \frac{1}{2}\sum\nolimits_{\ell=1}^{2}\big(\|\hat{r}_\ell\| - 1\big)^2\right]}_{\mathcal{R}_{\text{rep}}}. \tag{30}$$

Here $\|\cdot\|_{\text{act}}$ applies the fixed action-component weights in Table 7. The representation regularizer $\mathcal{R}_{\text{rep}}$ is applied independently to every action-step token and keeps the point channel's homogeneous coefficient near unit magnitude and the two rotation vectors near unit norm. The constant c_{train} prevents numerical instability near $t = 1$ and is also listed in Table 7.

Inference. We draw the initial chunk z_0 from the source in Eq. 5 and encode the observation history once, reusing the resulting context tokens throughout integration. We integrate the flow with a front-loaded exponential Euler schedule. At each step, the network predicts the endpoint position, 6D rotation, and gripper state $\hat{z}_1$. We convert this endpoint to a velocity and apply the Euler update

$$\hat{v}_i = \frac{\hat{z}_1 - z_{t_i}}{\max(1 - t_i,\, c_{\text{inf}})}, \qquad z_{t_{i+1}} = z_{t_i} + \Delta t_i\, \hat{v}_i. \tag{31}$$

The denominator floor c_{inf} prevents the velocity from becoming numerically unstable near $t = 1$. After integration, each 6D rotation is projected to $SO(3)$ by Gram-Schmidt. Since the network and readback are equivariant, the Euler update uses only scalar coefficients, and Gram-Schmidt commutes with proper rotations, the discrete sampler remains $SE(3)$-equivariant.

C.7 Architecture and optimization details

Table 7 reports the key architecture, training, and inference hyperparameters for reproducibility.

Table 7: Key architecture, training, and inference hyperparameters of GARLIC. Channel widths are reported as multivector/scalar.

Component	Hyperparameter	Value
Encoder	Self-attention blocks	6
	Token schedule	$512 \to 512 \to 256 \to 128 \to 64 \to 32 \to 16$
	Width	$32/32$
	Attention	4 heads, multi-query
	MLP width	$64/64$
	Context width	$16/32$
Decoder	Cross-attention blocks	16
	Width	$256/256$
	Attention	4 heads, multi-query ($8/16$ per head)
	Scalar conditioning	t, k, z_g; 256 each $\to 256$
	Temporal convolution	4 blocks; kernel 5; dilations $1, 2, 4, 8$
	Output MLP width	$512/512$
	Output width	$4/1$
Flow matching	Source standard deviation σ	1
	Flow time	$t \sim \mathrm{U}[0, 1)$, one per chunk
	Remaining-time floor c_{train} in Eq. 29	0.05
	Loss weights	$\lambda_{\mathrm{pose}} = 10$, $\lambda_g = 1$, $\lambda_{\mathrm{reg}} = 0.1$
	Time embedding	256-d sinusoidal; t scaled by 20
Inference	Euler steps n	5
	Schedule	front-loaded exponential
Optimization	Optimizer	AdamW, $\beta = (0.95, 0.999)$, $\epsilon = 10^{-8}$
	Learning rate	3.52×10^{-5}
	Weight decay	10^{-6}
	Global batch size	352
	Training epochs	1500
	EMA decay	0.999

D Experiment Details

D.1 Planar $SE(2)$ Toy Task

Purpose and controlled comparison. This experiment isolates source-induced equivariance error in a setting where both source and policy densities can be evaluated directly. We independently train three flow-matching policies with the same demonstration distribution, frame-conjugated velocity-field architecture, and optimization protocol, varying only the source distribution.

World and demonstrations. The action $a \in \mathbb{R}^2$ is a planar grasp position, and the observation $O = (o, h)$ consists of an object position $o \in \mathbb{R}^2$ and a unit heading $h \in \mathbb{S}^1$. A transformation $\rho = (R, v) \in SE(2)$ acts on the observation as $(o, h) \mapsto (Ro + v, Rh)$ and on the action as $a \mapsto Ra + v$. Let

$$R_o = [\, h,\ Jh \,], \qquad J = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix},$$

where R_o is the rotation matrix whose first column is h.

Conditioned on O, the demonstrated action follows

$$a \mid O \sim 0.7\mathcal{N}\big(o + R_o\mu_+,\, 0.15^2 I_2\big) + 0.3\mathcal{N}\big(o + R_o\mu_-,\, 0.15^2 I_2\big), \qquad \mu_\pm = \begin{bmatrix} 1.0 \\ \pm 0.6 \end{bmatrix}.$$

This conditional distribution is exactly $SE(2)$-equivariant. The unequal mixture weights avoid an accidental reflection symmetry, while the two modes make the learned transport nontrivial. For each training seed, we pre-generate 100,000 scenes by sampling o uniformly from $[-2, 2]^2$, the heading angle α uniformly from $[0, 2\pi)$ with $h = (\cos\alpha, \sin\alpha)$, and one demonstrated action from the conditional distribution above.

Flow model, source conditions, and training. Let z_0 denote a source sample, and let $z_1 = a$ denote a demonstrated action. All conditions use the frame-conjugated velocity field

$$v_\theta(z_t, t, O) = R_o f_\theta\big(R_o^\top (z_t - o), t\big)\,,$$

where f_θ is a two-hidden-layer MLP with SiLU activations. This field satisfies A2 for every θ, independently of training, so differences between the conditions isolate the effect of the source.

The three conditions use the following sources:

$$\begin{aligned} \text{fixed Gaussian:} \quad & z_0 \sim \mathcal{N}(0, I_2), \\ \text{anchored Gaussian:} \quad & z_0 \mid O \sim \mathcal{N}(o, I_2), \\ \text{workspace uniform:} \quad & z_0 \sim \mathrm{U}\big([-2,2]^2\big)\,. \end{aligned}$$

The fixed Gaussian and workspace-uniform sources are observation-independent, whereas the anchored Gaussian depends only on the object position.

Each model is trained by conditional flow matching along the straight path

$$z_t = (1-t)z_0 + t z_1, \qquad \mathcal{L} = \mathbb{E}\Big[\|v_\theta(z_t, t, O) - (z_1 - z_0)\|_2^2\Big]\,, \qquad t \sim \mathrm{U}[0,1).$$

For each of three seeds, all source conditions use the same demonstrations, minibatch indices, and sampled flow times; only the source samples differ. We evaluate the final checkpoint without validation-based model selection.

Evaluation protocol. We evaluate the equivariance error on eight observations held out from training. Translation probes use $v = (r, 0)$ with $r \in \{0, 0.25, 0.5, 1, 2, 4\}$. Rotation probes use $\varphi \in \{15^\circ, 30^\circ, 45^\circ, 90^\circ, 180^\circ\}$ about either the origin or the off-center point $c = (2, 0)$. We report TV for all three sources.

For each probe, we independently evaluate the two policy densities $\pi_\theta(\cdot \mid \rho \cdot O)$ and $\rho_\# \pi_\theta(\cdot \mid O)$ using the continuous change-of-variables formula along the learned ODE, and compute TV by numerical quadrature. Thus, the policy-side measurements do not assume the equivariance relation being tested. At each transformation, Figure 3 shows one marker for each of the eight held-out observations, averaged across three seeds. The curves show the closed-form source errors derived below.

Closed-form source errors. Let Φ denote the standard normal CDF. For the fixed Gaussian, define the effective displacement

$$d(\rho) = \begin{cases} |r|, & \rho = (I, (r, 0)), \\ 2\|c\|\,|\sin(\varphi/2)|, & \rho = (R_\varphi, (I - R_\varphi)c), \end{cases}$$

where the second case is a rotation by φ about c. The fixed Gaussian's source errors are

$$\mathcal{E}_{\rho,O}[\mathcal{N}(0, I_2)] = 2\Phi\left(\frac{d(\rho)}{2}\right) - 1, \qquad \mathcal{E}^{\mathrm{KL}}_{\rho,O}[\mathcal{N}(0, I_2)] = \frac{d(\rho)^2}{2}.$$

For rotations about the origin, $c = 0$, so both errors vanish. The anchored Gaussian $\mathcal{N}(o, I_2)$ has zero TV and KL error under every probe.

For the workspace-uniform source, let $B = [-2, 2]^2$ and $B_\rho = RB + v$. Since both supports have area 16,

$$\mathcal{E}_{\rho,O}[\mathrm{U}(B)] = 1 - \frac{\mathrm{area}(B \cap B_\rho)}{16}.$$

An axial translation therefore gives $\min(|r|/4, 1)$, while a rotation about c uses $B_\rho = R_\varphi B + (I - R_\varphi)c$. For rotations about the origin, the error vanishes at multiples of 90°, reflecting the square's residual rotational symmetry. For a generic box transformation, KL is infinite because the supports differ. The TV expressions give the curves in Figure 3; for the exact invertible flow, the policy and source TV errors are equal by Theorem 1(ii), while invertible changes of variables likewise preserve KL.

D.2 Simulation Experiments

Observation and evaluation protocol. Each RLBench environment uses a 7-DoF Franka Emika arm and five RGB-D cameras: four fixed cameras surrounding the workspace and one wrist-mounted camera. At each observation step, we transform the point clouds from all five views into the simulator world frame, concatenate them, and uniformly sample 512 XYZ points without replacement. RGB is not provided to the policies. The sampled point cloud is paired with the end-effector pose and gripper state defined in Section 3, and all methods receive the same raw observations.

For each task and demonstration budget, we train three seeds. We evaluate the last five checkpoints of each run on the same 100 randomized task initializations and use RLBench's binary task-success signal. This gives $3 \times 5 \times 100 = 1{,}500$ evaluation episodes per method and data budget. We report the mean and standard deviation over the checkpoint success rates. Architecture, training, and inference hyperparameters for GARLIC are reported in Table 7.

Parameter budgets. For Figure 5, we vary only the model width to obtain the reported parameter counts and otherwise use the same data, training schedule, and evaluation protocol as Table 2.

OOD evaluation. We evaluate the checkpoints from Table 2 without fine-tuning. For each randomized base scene, we construct paired ID and OOD trials. The OOD trial applies a planar rigid transformation $\rho = (R_\theta, v) \in SE(2)$, where θ is sampled uniformly from $[0, \pi)$ about the gravity axis through the workspace center and $v = (v_x, v_y, 0)$, with $v_x \sim \mathcal{U}(-0.152\,\mathrm{m}, 0.152\,\mathrm{m})$ and $v_y \sim \mathcal{U}(-0.163\,\mathrm{m}, 0.163\,\mathrm{m})$. The magnitude bound on each axis is one eighth of the corresponding workspace extent. Each checkpoint is evaluated on the same 50 paired ID/OOD scenes, with the base scenes and sampled transformations shared across methods and seeds. We report the success rate in each setting and their difference $\Delta = \mathrm{OOD} - \mathrm{ID}$. Table 8 reports the results for the remaining three tasks, while Table 3 in the main text highlights `Unplug Charger`.

Table 8: OOD generalization on the remaining three RLBench tasks. ID and OOD success rates are reported as percentages, with $\Delta = \mathrm{OOD} - \mathrm{ID}$.

Task	Method	ID (%)	OOD (%)	Δ
`Take Lid Off Saucepan`	DP3	86.0 ± 1.6	2.0 ± 0.0	-84.0
	PFM	91.0 ± 1.4	4.0 ± 1.6	-87.0
	DP3-Canon.	85.3 ± 2.5	82.7 ± 2.5	-2.6
	GARLIC	$\mathbf{95.2 \pm 2.6}$	$\mathbf{93.4 \pm 1.5}$	$\mathbf{-1.8}$
`Water Plants`	DP3	68.7 ± 3.4	0.0 ± 0.0	-68.7
	PFM	70.0 ± 5.9	1.3 ± 0.9	-68.7
	DP3-Canon.	61.7 ± 3.4	60.7 ± 7.7	-1.0
	GARLIC	$\mathbf{89.8 \pm 1.9}$	$\mathbf{90.4 \pm 2.5}$	$\mathbf{+0.6}$
`Take Frame Off Hanger`	DP3	40.8 ± 7.6	6.1 ± 5.0	-34.7
	PFM	60.6 ± 3.8	4.1 ± 1.7	-56.5
	DP3-Canon.	39.5 ± 1.0	38.8 ± 0.0	-0.7
	GARLIC	$\mathbf{73.2 \pm 2.6}$	$\mathbf{72.8 \pm 1.5}$	$\mathbf{-0.4}$

Ablation protocol. All ablations in Table 4 are trained on `Unplug Charger` with 100 demonstrations and use the training, checkpoint-selection, and paired ID/OOD evaluation protocols described above. All settings not involved in the stated intervention remain unchanged. DP3+Aug draws its training-time transformations from the same $SE(2)$ distribution used for OOD evaluation and applies each transformation jointly to the observation and action.

For the Bingham-prior variant, the translation and gripper factors in Eq. 5 remain unchanged. Let R_{ref} denote the end-effector rotation in the last observation, recovered from its 6D representation by Gram–Schmidt. We sample an identity-centered isotropic Bingham perturbation δ with concentration $\kappa = 8$ and set

$$R_0 = R_{\mathrm{ref}} R(\delta), \qquad r_0 = [R_0 e_1, R_0 e_2].$$

Equivalently, its quaternion density is $p(q_0 \mid O) \propto \exp\big(8(q_0^\top q_{\mathrm{ref}})^2\big)$. Because the perturbation is composed on the right, a world rotation G sends R_0 to $G R_{\mathrm{ref}} R(\delta) = G R_0$ using the same δ; hence the source remains compatible.

D.3 Real-world Experiment Setup

Hang hat on rack. The wooden rack remains fixed, while the hat is placed at a random position and orientation on one side of the rack. The robot must pick up the hat and hang it on the rack. A trial is successful if the hat is hanging on the rack within the time limit.

Place cherry in bowl. The bowl remains fixed, while the cherry is placed at a random position and orientation. The robot must pick up the cherry and place it in the bowl. A trial is successful if the cherry is inside the bowl within the time limit.

Store pepper grinders. The wooden box remains fixed, while two pepper grinders are placed at random positions on the same side of the box. The robot must place the two grinders into the box one after the other. A trial is successful if both grinders are inside the box within the time limit and remain upright.

Erase whiteboard. The whiteboard remains fixed, while the eraser is placed at a random position on one side of the board. The robot must pick up the eraser and use it to remove the writing. A trial is successful if the writing is erased within the time limit.

Hardware and data collection. Figure 6 shows our real-world setup. A Franka FR3 arm equipped with a Franka Hand performs all four tasks. We command absolute end-effector poses at 20 Hz through the operational-space controller in Deoxys (Zhu et al., 2022), with a proportional gain of $k_p = 350$. A fixed third-person Orbbec Gemini 335 camera records aligned RGB-D images at 640×480 and 30 Hz, although the policies use only depth. We calibrate the camera extrinsics and tabletop plane once and keep them fixed throughout data collection and evaluation. Policy inference runs on a separate machine with an NVIDIA RTX Pro 6000 Blackwell GPU, an AMD Ryzen Threadripper 7970X CPU, and 256 GB of DDR5 ECC memory, which communicates with the robot-control PC over TCP.

For each task, we collect 20 demonstrations by teleoperating the arm with a Meta Quest 2 controller. The robot uses the same operational-space controller and gains during demonstration collection and evaluation. Each demonstration lasts 10–35 s. We record the end-effector pose and gripper width at 100 Hz and camera frames at 30 Hz on a shared clock, then resample both streams to a uniform 10 Hz grid using nearest-neighbor matching, with a maximum time offset below 30 ms. We hold out one demonstration per task for validation and use the remaining 19 for training.

Shared policy interface. All methods use the same observation and action interface. Each observation contains the current frame and one recorded 300 ms earlier. To obtain each point cloud, we back-project the depth image, transform the points into the robot base frame, remove points within 8 mm of the calibrated table plane, apply an axis-aligned workspace crop, and uniformly sample 512 points. Each policy predicts a 32-step action chunk at 300 ms intervals, giving a 9.6 s horizon. The same preprocessing is used during training and deployment.

Closed-loop execution. All methods use the same inference server and robot-side executor. The executor interpolates each predicted action chunk onto its 50 ms control grid, using linear interpolation for position, spherical interpolation for rotation, and zero-order hold for the gripper. It then converts the targets to the robot base frame and assigns their execution times. The predicted gripper channel is thresholded at 0.5 to produce open and close commands.

The robot continues executing the current chunk while the policy predicts the next one, roughly every 4.8 s. When a new chunk arrives, the executor joins it to the current trajectory by timestamp, so inference does not pause the robot. To reduce sampling variation across replanning steps, we reuse the same underlying random draw within each episode.

End-to-end inference takes approximately 90 ms for DP3, 290 ms for PFM, and 400 ms for our method, all below the replanning interval. For safety, an episode terminates if a commanded target leaves the workspace.

Baseline Implementations and Training. We compare against DP3 (Ze et al., 2024) and Point-FlowMatch (Chisari et al., 2024), using their official implementations and default hyperparameters. For a controlled comparison, we train both baselines for 1,500 epochs with a batch size of 128 and set their prediction and execution horizons to 32 and 16 steps, respectively. We additionally set DP3's U-Net channel widths to 256, 512, and 1024 to keep its parameter count comparable to Point-

FlowMatch. We train a separate policy from scratch for each task, using one seed per method and task.

E Additional Discussion

E.1 From an $E(3)$ Backbone to an $SE(3)$ Policy

A reflection acts on the entire scene, including the manipulated objects, robot arm, and gripper. For generic asymmetric objects and robots, the reflected scene contains mirrored bodies that cannot be obtained by moving the original physical system. The relevant symmetry of this fixed embodiment is therefore $SE(3)$: rotations and translations preserve the physical bodies and their handedness, whereas reflections do not.

Although the GATr backbone is $E(3)$-equivariant (Brehmer et al., 2023), we deliberately design the complete policy to preserve handedness. The pose embedding and Gram–Schmidt readback use cross products to construct right-handed frames, and the output head uses a fixed canonical join reference. These operations commute with proper rigid motions but not reflections, so the end-to-end policy matches the $SE(3)$ symmetry of the manipulation problem.

E.2 Existing Policies in the Characterization of Proposition 1

Table 9 interprets existing equivariant policies using the factorwise source constructions of Table 1 (Lemma 1). For absolute-pose actions, Proposition 2 shows that only translation requires observation dependence.

Table 9: Existing policies in the factorwise design space of Table 1.

Policy	Translational construction	Mechanism
EquiDiff (Wang et al., 2024)	excluded from its rotation-only guarantee	(a), restricted group
EquiBot (Yang et al., 2024a)	global translation removed by the velocity representation	(a)+(c), factorwise
DP3-Canon.	full end-effector-frame canonicalization	(c)
GARLIC (Eq. 5)	source anchored at the end-effector position	(a)+(b), factorwise

EquiBot factorizes actions into linear velocity, angular velocity, and gripper components. Global translations cancel in its velocity representation. Viewed through our characterization, this is factorwise canonicalization: translation uses column (c), while rotation and gripper retain observation-independent sources from column (a). DP3-Canon. (Table 2) instead applies column (c) to the complete pose. GARLIC retains world-coordinate absolute poses and anchors only the translational source through column (b), while using column (a) for rotation and gripper. Thus, EquiBot, DP3-Canon., and GARLIC handle translation through the action representation, a complete canonical frame, and the source distribution, respectively.