# Appendix for "EquiGraspFlow: SE(3)-Equivariant 6-DoF Grasp Pose Generative Flows"

**Anonymous Author(s)**
Affiliation
Address
email

## Contents

## A Proof of $\mathrm{SE}(3)$-Invariant Conditional Distributions

In this section, we restate and prove Proposition 1 in section 4.1.

**Proposition 1.** *Suppose a prior distribution $p_0(T|\mathcal{P})$ is $\mathrm{SE}(3)$-invariant. If the angular and linear velocity fields $\omega, v$ are $\mathrm{SE}(3)$-equivariant, then the conditional distribution $p_t(T|\mathcal{P})$ at any time $t \geq 0$ defined via the flow of ODE $(\dot{R}, \dot{x}) = ([\omega_\theta(t, \mathcal{P}, T)]R, v_\phi(t, \mathcal{P}, T))$ is $\mathrm{SE}(3)$-invariant.*

To this end, we first introduce the concept of a conditional flow derived from the velocity fields and define the $\mathrm{SE}(3)$-equivariance of the conditional flow. Subsequently, we demonstrate that $\mathrm{SE}(3)$-equivariant time-dependent conditional velocity fields induce an $\mathrm{SE}(3)$-equivariant conditional flow. Finally, we prove Proposition 1 by establishing that, starting from an $\mathrm{SE}(3)$-invariant prior, an $\mathrm{SE}(3)$-equivariant conditional flow preserves the invariance over time.

### A.1 $\mathrm{SE}(3)$-Equivariant Conditional Flows

Consider a trajectory on the $\mathrm{SE}(3)$ manifold, starting from an initial point $T \in \mathrm{SE}(3)$ and guided by the time-dependent angular and linear velocity fields, $\omega$ and $v$, conditioned on a point cloud $\mathcal{P}$. This trajectory is called an *integral curve* for $\omega$ and $v$ conditioned on $\mathcal{P}$ and starting at $T$, and is denoted by $\gamma : \mathbb{R} \to \mathrm{SE}(3)$. By decomposing the $\mathrm{SO}(3)$ and $\mathbb{R}^3$ components such that $\gamma(t) = (\gamma_R(t), \gamma_x(t))$, the integral curve is defined via the following ordinary differential equations (ODEs) of the velocity fields:

$$\dot{\gamma}_R(t) = [\omega(t, \mathcal{P}, \gamma(t))]\gamma_R(t), \quad \dot{\gamma}_x(t) = v(t, \mathcal{P}, \gamma(t)), \quad \gamma(0) = T. \tag{1}$$

Denoting the space of point clouds by $\mathcal{X}$, a *conditional flow* of the velocity fields $\omega$ and $v$ conditioned on $\mathcal{P}$ is defined as a mapping $f : \mathbb{R} \times \mathcal{X} \times \mathrm{SE}(3) \to \mathrm{SE}(3)$. Here, $f(t, \mathcal{P}, T) = \gamma(t)$ where $\gamma$ is the integral curve for $\omega$ and $v$ conditioned on $\mathcal{P}$ and starting at $T$.

Now, we define the $\mathrm{SE}(3)$-equivariance of a conditional flow as follows:

**Definition 3.** *A flow on* $\mathrm{SE}(3)$ *conditioned on a point cloud, denoted by $f(t, \mathcal{P}, T)$, is* $\mathrm{SE}(3)$*-equivariant if, for an arbitrary $T' \in \mathrm{SE}(3)$, $f(t, T'\mathcal{P}, T'T) = T' f(t, \mathcal{P}, T)$.*

Next, We demonstrate that $\mathrm{SE}(3)$-equivariant time-dependent conditional velocity fields induce the $\mathrm{SE}(3)$-equivariance of their conditional flow through the following Proposition:

**Proposition 2.** *For any time-dependent conditional angular and linear velocity fields $\omega$ and $v$, their conditional flow $f$ is* $\mathrm{SE}(3)$*-equivariant if $\omega$ and $v$ are* $\mathrm{SE}(3)$*-equivariant.*

*Proof.* Consider an arbitrary point cloud $\mathcal{P}$ and fix $T = (R_0, x_0)$ as an arbitrary element in $\mathrm{SE}(3)$. Then, $f(t, \mathcal{P}, T) = \gamma(t) = (\gamma_R(t), \gamma_x(t))$ represents the integral curve for the velocity fields conditioned on $\mathcal{P}$ and starting at $T$. The ODEs governing this integral curve are given by the same equations as (1).

For any $T' = (R', x') \in \mathrm{SE}(3)$, $f(t, T'\mathcal{P}, T'T) = \tilde{\gamma}(t) = (\tilde{\gamma}_R(t), \tilde{\gamma}_x(t))$ where $\tilde{\gamma}$ is the integral curve for the velocity fields conditioned on $T'\mathcal{P}$ and starting at $T'T$. The ODEs for this integral curve are given by:

$$\dot{\tilde{\gamma}}_R(t) = [\omega(t, T'\mathcal{P}, \tilde{\gamma}(t))]\tilde{\gamma}_R(t), \quad \dot{\tilde{\gamma}}_x(t) = v(t, T'\mathcal{P}, \tilde{\gamma}(t)), \quad \tilde{\gamma}(0) = T'T. \tag{2}$$

Now, consider an integral curve $\hat{\gamma}$ defined as $\hat{\gamma}(t) = (\hat{\gamma}_R(t), \hat{\gamma}_x(t)) := (R'\gamma_R(t), R'\gamma_x(t) + x') = T'(\gamma_R(t), \gamma_x(t)) = T'\gamma(t) = T' f(t, \mathcal{P}, T)$. This integral curve results from transforming the integral curve $(\gamma_R(t), \gamma_x(t))$ by $T'$.

To prove the $\mathrm{SE}(3)$-equivariance of the conditional flow, we need to show that $\tilde{\gamma}$ and $\hat{\gamma}$ are the same integral curve. Specifically, we need to show that $\tilde{\gamma}(t) = f(t, T'\mathcal{P}, T'T) = T' f(t, \mathcal{P}, T) = \hat{\gamma}(t)$.

Noting that $R[a]R^T = [Ra]$ for any $R \in \mathrm{SO}(3)$ and $a \in \mathbb{R}^3$, we analyze $\dot{\hat{\gamma}}_R(t)$ as follows:

$$\begin{aligned}
\dot{\hat{\gamma}}_R(t) &= \frac{d}{dt}(R'\gamma_R(t)) = R'\dot{\gamma}_R(t) \\
&= R'[\omega(t, \mathcal{P}, \gamma(t))]\gamma_R(t) \\
&= [R'\omega(t, \mathcal{P}, \gamma(t))]R'\gamma_R(t) \\
&= [\omega(t, T'\mathcal{P}, T'\gamma(t))]R'\gamma_R(t) \\
&= [\omega(t, T'\mathcal{P}, \hat{\gamma}(t))]\hat{\gamma}_R(t).
\end{aligned} \tag{3}$$

Similarly, for $\hat{\gamma}_x(t)$, we have:

$$\begin{aligned}
\dot{\hat{\gamma}}_x(t) &= \frac{d}{dt}(R'\gamma_x(t) + x') \\
&= R'\dot{\gamma}_x(t) \\
&= R'v(t, \mathcal{P}, \gamma(t)) \\
&= v(t, T'\mathcal{P}, T'\gamma(t)) \\
&= v(t, T'\mathcal{P}, \hat{\gamma}(t)).
\end{aligned} \tag{4}$$

Finally, note that $\hat{\gamma}(0) = T'\gamma(0) = T'T$. Thus, $\tilde{\gamma}(t)$ and $\hat{\gamma}(t)$ satisfy the same ODEs, and the uniqueness of the solution of the ODE ensures that $\tilde{\gamma}$ and $\hat{\gamma}$ are the same integral curve. Consequently, we have $f(t, T'\mathcal{P}, T'T) = T' f(t, \mathcal{P}, T)$ for any $T' \in \mathrm{SE}(3)$, demonstrating that $f$ is $\mathrm{SE}(3)$-equivariant. □

### A.2 $\mathrm{SE}(3)$-Invariant Conditional Distributions

To demonstrate that an $\mathrm{SE}(3)$-equivariant conditional flow preserves the invariance of an $\mathrm{SE}(3)$-invariant prior, we present the following proposition.

**Proposition 3.** *Suppose a prior distribution $p_0(T|\mathcal{P})$ is* $\mathrm{SE}(3)$*-invariant. If the conditional flow $f$ is* $\mathrm{SE}(3)$*-equivariant, then the conditional distribution $p_t(T|\mathcal{P})$ at any time $t \geq 0$ defined via the flow is* $\mathrm{SE}(3)$*-invariant.*

*Proof.* To prove this proposition, we first extend Theorem 3 from [1], which involves a general Riemannian manifold and a general group, to a conditional version.

Consider a Riemannian manifold $(\mathcal{M}, h)$ with a group $G$. Denote the action of an element $g \in G$ on $\mathcal{M}$ by the map $L_g : \mathcal{M} \to \mathcal{M}$. The map $L_g$ is isometric if, for any tangent vectors $u$ and $v$ at any point $x \in \mathcal{M}$, the following condition holds: $h(d(L_g)_x(u), d(L_g)_x(v)) = h(u, v)$, where $d(L_g)_x$ represents the differential of $L_g$ at $x$. If $L_g$ is isometric, then $\left|\det J_{L_g}(x)\right| = 1$ for any $x \in \mathcal{M}$, where $J_{L_g}(x)$ denotes the Jacobian matrix of the map $L_g$ evaluated at $x$ and expressed in local coordinates.

Let $c$ denote a condition variable. The conditional flow at time $t$ is represented by the map $f_{t,c} : \mathcal{M} \to \mathcal{M}$. This flow transforms a prior conditional distribution $p_0(x|c)$ into the conditional distribution $p_t(x|c)$. The likelihood of the transformed conditional distribution is given by the following change of variables formula:

$$p_t(x|c) = p_0\left(f_{t,c}^{-1}(x)\middle|c\right)\left|\det J_{f_{t,c}^{-1}}(x)\right|. \tag{5}$$

A conditional distribution $p(x|c)$ is $G$-invariant if $p(L_g(x)|g \cdot c) = p(x|c)$ for any $g \in G$. Assuming the action of $g \in G$ on $c$ is well-defined and denoted by $g \cdot c$, the conditional flow $f_{t,c}$ is $G$-equivariant if, $f_{t,g\cdot c}(L_g(x)) = L_g(f_{t,c}(x))$ for any $g \in G$, i.e., $f_{t,g\cdot c} \circ L_g = L_g \cdot f_{t,c}$ and $L_g^{-1} \circ f_{t,g\cdot c}^{-1} = f_{t,c}^{-1} \circ L_g^{-1}$.

Assuming that the map $L_g$ is isometric for any $g \in G$, we can prove Proposition 3 in a general Riemannian manifold $\mathcal{M}$ and a general group $G$ as follows:

$$\begin{aligned}
&p_t(L_g(x)|g \cdot c) \\
&= p_0\left(f_{t,g\cdot c}^{-1}(L_g(x))\middle|g \cdot c\right)\left|\det J_{f_{t,g\cdot c}^{-1}}(L_g(x))\right| \\
&= p_0\left(L_{g^{-1}}\left(f_{t,g\cdot c}^{-1}(L_g(x))\right)\middle|c\right)\left|\det J_{f_{t,g\cdot c}^{-1}}(L_g(x))\right| && \text{(invariant prior)} \\
&= p_0\left(\left(L_{g^{-1}} \circ f_{t,g\cdot c}^{-1} \circ L_g\right)(x)\middle|c\right) \\
&\qquad \underbrace{\left|\det J_{L_{g^{-1}}}\left(\left(f_{t,g\cdot c}^{-1} \circ L_g\right)(x)\right)\right|}_{=1}\left|\det J_{f_{t,g\cdot c}^{-1}}(L_g(x))\right|\underbrace{\left|\det J_{L_g}(x)\right|}_{=1} \\
&= p_0\left(\left(L_{g^{-1}} \circ f_{t,g\cdot c}^{-1} \circ L_g\right)(x)\middle|c\right) \\
&\qquad \left|\det J_{L_{g^{-1}}}\left(\left(f_{t,g\cdot c}^{-1} \circ L_g\right)(x)\right) J_{f_{t,g\cdot c}^{-1}}(L_g(x)) J_{L_g}(x)\right| && \text{(multiplicativity)} \\
&= p_0\left(\left(L_{g^{-1}} \circ f_{t,g\cdot c}^{-1} \circ L_g\right)(x)\middle|c\right)\left|\det J_{L_{g^{-1}} \circ f_{t,g\cdot c}^{-1} \circ L_g}(x)\right| && \text{(chain rule)} \\
&= p_0\left(\left(L_g^{-1} \circ f_{t,g\cdot c}^{-1} \circ L_g\right)(x)\middle|c\right)\left|\det J_{L_g^{-1} \circ f_{t,g\cdot c}^{-1} \circ L_g}(x)\right| && (L_{g^{-1}} = L_g^{-1}) \\
&= p_0\left(\left(f_{t,c}^{-1} \circ L_g^{-1} \circ L_g\right)(x)\middle|c\right)\left|\det J_{f_{t,c}^{-1} \circ L_g^{-1} \circ L_g}(x)\right| \\
&= p_0\left(f_{t,c}^{-1}(x)\middle|c\right)\left|\det J_{f_{t,c}^{-1}}(x)\right| \\
&= p_t(x|c).
\end{aligned} \tag{6}$$

Proposition 3 is a special case where $\mathcal{M} = \mathrm{SE}(3)$, $G = \mathrm{SE}(3)$, and $c = \mathcal{P}$, and the group action of $T' \in \mathrm{SE}(3)$ on $T \in \mathrm{SE}(3)$, denote by $L_{T'}(T) = T'T$, is the left translation map which is isometric. Hence, Proposition 3 is proved. □

We now prove Proposition 1 by utilizing Proposition 2 and Proposition 3.

*Proof of Proposition 1.* Since the angular and linear velocity fields $\omega_\theta$ and $v_\phi$ are $\mathrm{SE}(3)$-equivariant, it follows from Proposition 2 that their flow $f$ is also $\mathrm{SE}(3)$-equivariant. Consequently, by Proposition 3, the conditional distribution $p_t(T|\mathcal{P})$, which is defined via the flow of the velocity fields, is $\mathrm{SE}(3)$-invariant. □

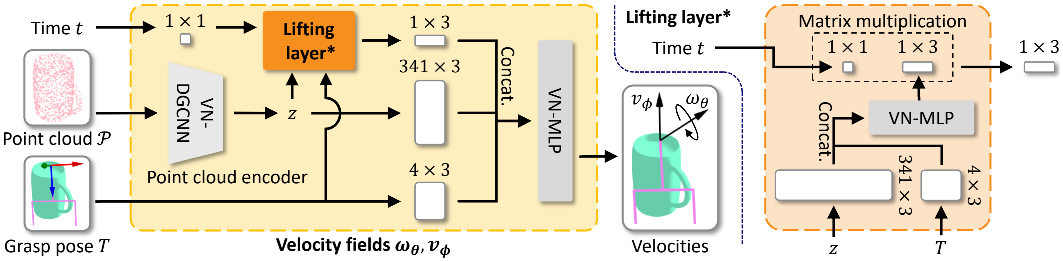

Figure 1: The structure of the velocity fields and lifting layer. The VN-DGCNN encodes the point cloud $\mathcal{P}$ into a representation $z$ consisting of 341 three-dimensional vectors. The VN-MLP in the lifting layer utilizes this representation along with the grasp pose T, to produce a matrix of size $1 \times 3$. This matrix lifts the time variable t (of size $1 \times 1$) to a three-dimensional vector. Finally, the VN-MLP takes as input the concatenated list of the lifted time, representation, and grasp pose, and outputs the angular and linear velocities.

# B Implementation Details

## B.1 Details for Networks

To model the time-dependent conditional velocity fields $\omega_\theta(t, \mathcal{P}, T)$ and $v_\phi(t, \mathcal{P}, T)$ with $\mathrm{SO}(3)$-equivariance, we employ Vector Neuron (VN) architectures [2], which are specifically designed for $\mathrm{SO}(3)$-equivariance. The structure of these velocity fields is illustrated in Figure 1.

To encode the point cloud $\mathcal{P}$, we utilize the backbone of the VN-DGCNN designed for classification tasks. We use the network before the invariant layer, excluding the batch normalization layers. We add an EdgeConv module with a size of 170 at the sixth module position. Following the backbone, a mean pooling layer is applied to pool the point dimension, extracting a representation $z$ consisting of 341 three-dimensional vectors. The grasp pose $T$ is reconfigured into a form that concatenates the three column vectors of the rotation part and one vector of the translation part, resulting in four three-dimensional vectors. Time $t$ is converted into a single three-dimensional vector through the lifting layer. Subsequently, the VN-MLP concatenates these lists of three-dimensional vectors as input and outputs the angular and linear velocities. The VN-MLP consists of five hidden VN-Linear layers, each followed by VN-LeakyReLU activation with a negative slope 0.2, and one output VN-Linear layer. The sizes of the hidden layers are (256, 256, 128, 128, 128), and the output layer size is 2, as the network's output is two three-dimensional vectors.

The lifting layer uses the representation $z$ and the grasp pose $T$ to convert the scalar time $t$ into a three-dimensional vector. This process involves a VN-MLP that consists of a single VN-Linear layer with a size of 1, producing an output matrix of size $1 \times 3$. This output matrix is then multiplied to the scalar $t$ (size $1 \times 1$), resulting in a single three-dimensional vector.

## B.2 Details for Training and Inference

**Dataset Split** We use a dataset of 101 mugs and 83 bowls obtained from the ACRONYM dataset [3] to train the networks. The dataset is split as follows: 61 mugs and 51 bowls are randomly selected for training, 20 mugs and 16 bowl for validation, and the remaining 20 mugs and 16 bowl for testing.

**Flow Matching** We employ the Flow Matching (FM) framework [4, 5] to train our continuous normalizing flow model. The core element of FM involves designing the *per-sample* target vector field $u_t^*(T|T_1)$ and the corresponding probability path $p_t(T|T_1)$, where $T_1 = (R_1, x_1)$ represents a particular sample from the target distribution $q(T|\mathcal{P})$. In our approach, we separate the rotation and translation components in $u_t^*(T|T_1) = (\omega_t^*(R|R_1), v_t^*(x|x_1))$. We then define the target angular and linear velocity fields $\omega_t^*(R|R_1)$ and $v_t^*(x|x_1)$ as follows:

$$[\omega_t^*(R|R_1)] = \frac{\log(R^T R_1)}{1-t}, \quad v_t^*(x|x_1) = \frac{x_1 - x}{1-t}. \tag{7}$$

Consequently, the training objective for EquiGraspFlow is designed as

$$\mathcal{L} = \mathbb{E}_{t, T_1 \sim q(T|\mathcal{P}), T \sim p_t(T|T_1)} \left[ \frac{1}{2} \left\| [\omega_\theta(t, \mathcal{P}, T)] - \frac{\log(R^T R_1)}{1-t} \right\|_F^2 + \left\| v_\phi(t, \mathcal{P}, T) - \frac{x_1 - x}{1-t} \right\|^2 \right] \quad (8)$$

where $|| \cdot ||_F$ denotes the Frobenius norm, and $T = (R, x)$ and $T_1 = (R_1, x_1)$.

One thing to note is that it might seem natural to design the vector field on the $\mathrm{SE}(3)$ manifold instead of separating the rotation and translation components, similarly to how we design the angular velocity field on the $\mathrm{SO}(3)$ manifold as shown in (7). However, this approach results in screw motion-shaped paths of grasp poses, where the translation may not follow a straight line toward the target grasp pose. In the context of our grasp pose generation task, separating the rotation and translation and ensuring that the translation motion directly heads toward the target grasp pose is a more intuitive and appropriate vector field formulation.

**Guided Flows** Guided Flows [6] is a technique that enhances the sample quality and efficiency of conditional generative models by integrating classifier-free guidance [7] into Flow Matching models. This method employs a guided velocity field during sampling, defined as a weighted sum of unconditional and conditional velocity fields. Using an empty set $\varnothing$ as a null condition for the point cloud input, we define the guided angular and linear velocity fields $\tilde{\omega}_\theta$ and $\tilde{v}_\phi$ as follows, utilizing the weight parameter $\beta$:

$$\begin{aligned} \tilde{\omega}_\theta(t, \mathcal{P}, T) &= (1 - \beta)\omega_\theta(t, \varnothing, T) + \beta\omega_\theta(t, \mathcal{P}, T), \\ \tilde{v}_\phi(t, \mathcal{P}, T) &= (1 - \beta)v_\phi(t, \varnothing, T) + \beta v_\phi(t, \mathcal{P}, T). \end{aligned} \quad (9)$$

When $\varnothing$ is input, the point cloud encoder outputs a list of zero vectors as $z$. To train the unconditional velocity fields, we randomly replace $\mathcal{P}$ with the empty set $\varnothing$ with a probability of 20% during training. For inference, we use $\beta = 1.25$ to evaluate average performance and $\beta = 2$ to assess the consistency of performance.

**Optimizer** Adam optimizer [8] with learning rate $1 \times 10^{-4}$ is utilized to train the baselines and our network. L2 regularization with hyperparameter $1 \times 10^{-5}$ and $1 \times 10^{-6}$ are employed for training 6-DOF GraspNet (GAN) [9] and EquiGraspFlow.

### B.3 Details for Grasping Motion in Real-World Experiments

The robot motion for grasping an object in real-world experiment is designed as follows. To prevent collisions with the object during the movement of the gripper toward the generated grasp pose, we first move the gripper to a pre-grasp pose. This pre-grasp pose is offset from the grasp pose by a small distance in the $-z$ direction in the gripper's frame (the $z$-axis of the gripper's frame represents the direction of gripper's palm). Next, we move the gripper to the grasp pose and execute the grasp. Once the object is grasped, the gripper is lifted by 10cm. The success of the grasp is manually determined based on whether the object is held securely by the gripper. After each grasping attempt, we manually reset the position and orientation of the object to its initial state.

## C Additional Results for Grasp Pose Generation

Figures 2 to 5 illustrate the additional visualizations of the generated grasp poses. These figures show the generated grasp poses of two mugs and two bowls for ten object rotations, along with the Earth Mover's Distance (EMD) and grasp success rate values. The objects are rotated and input into each model, but in these figures, both the objects and the generated grasp poses are inversely rotated to align all scenes. Successful and failed grasp poses are indicated in green and red, respectively.

The grasp poses generated by EquiGraspFlow are widely distributed across various parts of the objects, demonstrating that our model generates more diverse grasp poses compared to the baselines. The values indicate that EquiGraspFlow generates grasp poses similar to the ground truth with high success rate. Additionally, the variance in the values indicates that EquiGraspFlow exhibits more consistent results across different object rotations. Notably, our model maintains identical value across the ten object rotations, demonstrating the perfect equivariance of our approach.

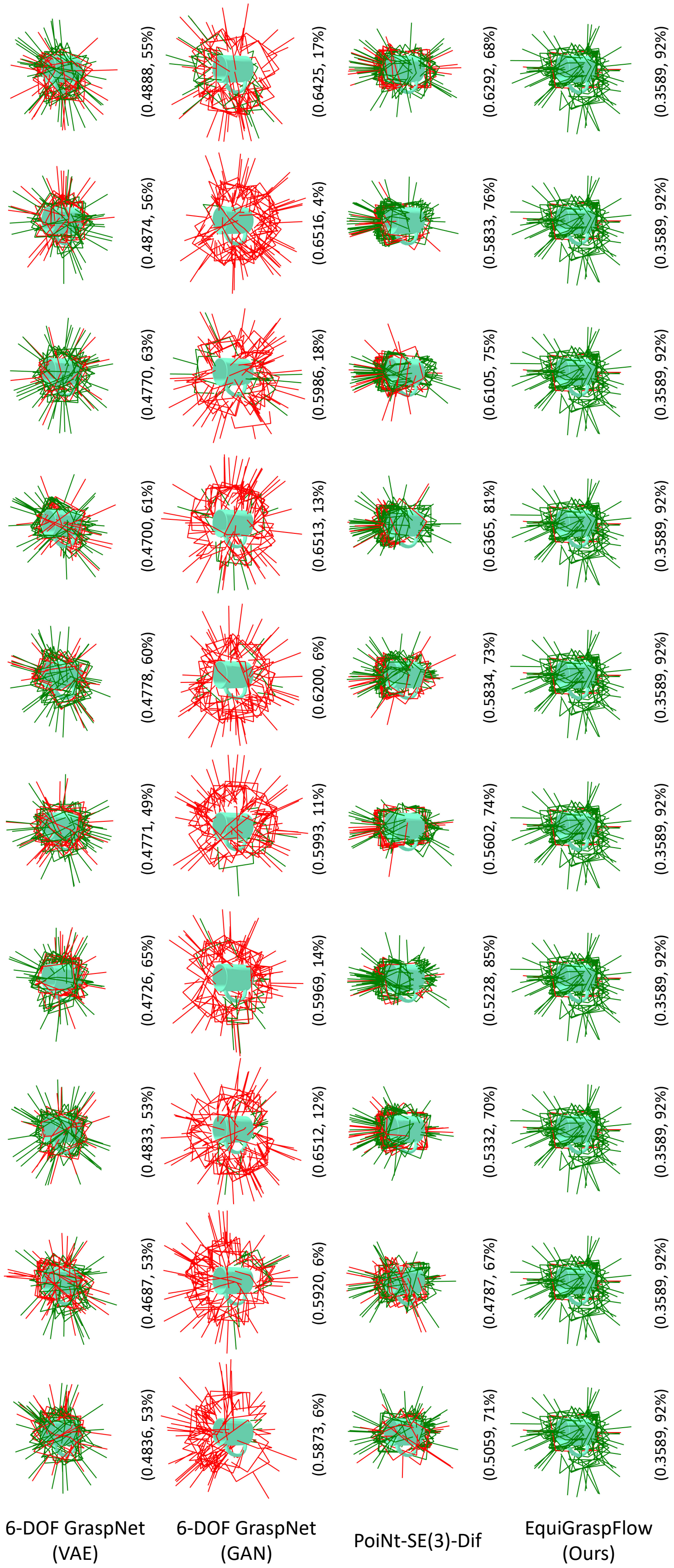

Figure 2: The generated grasp poses for the first mug across ten rotations.

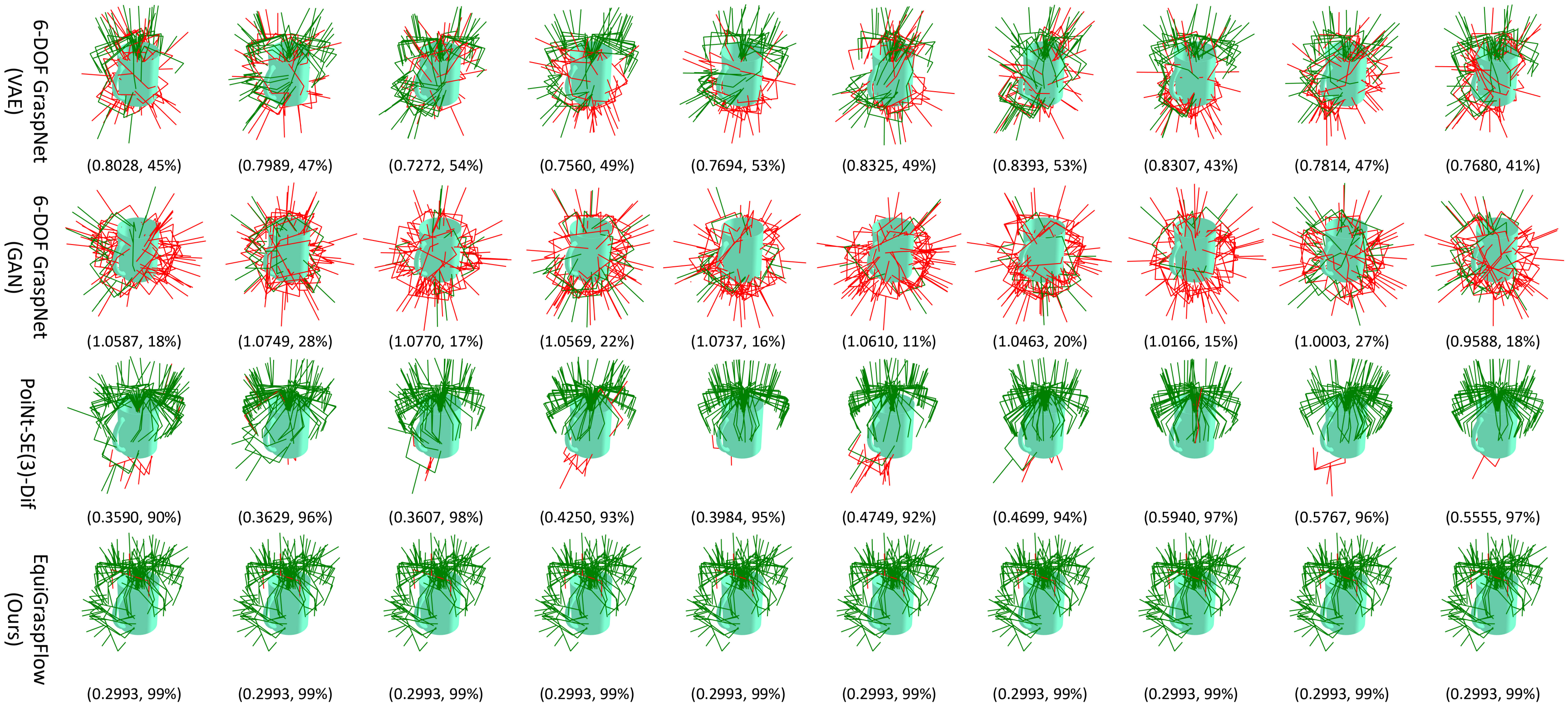

Figure 3: The generated grasp poses for the second mug across ten rotations.

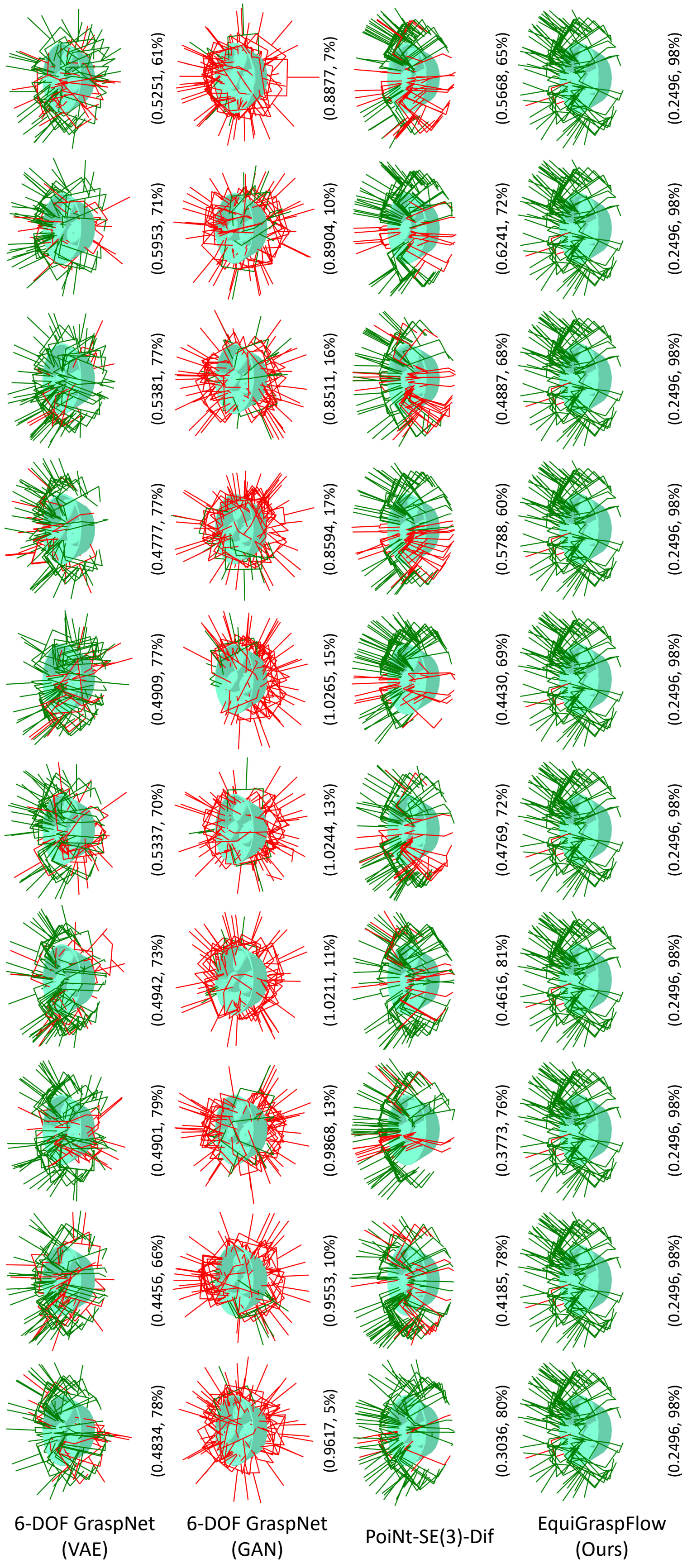

Figure 4: The generated grasp poses for the first bowl across ten rotations.

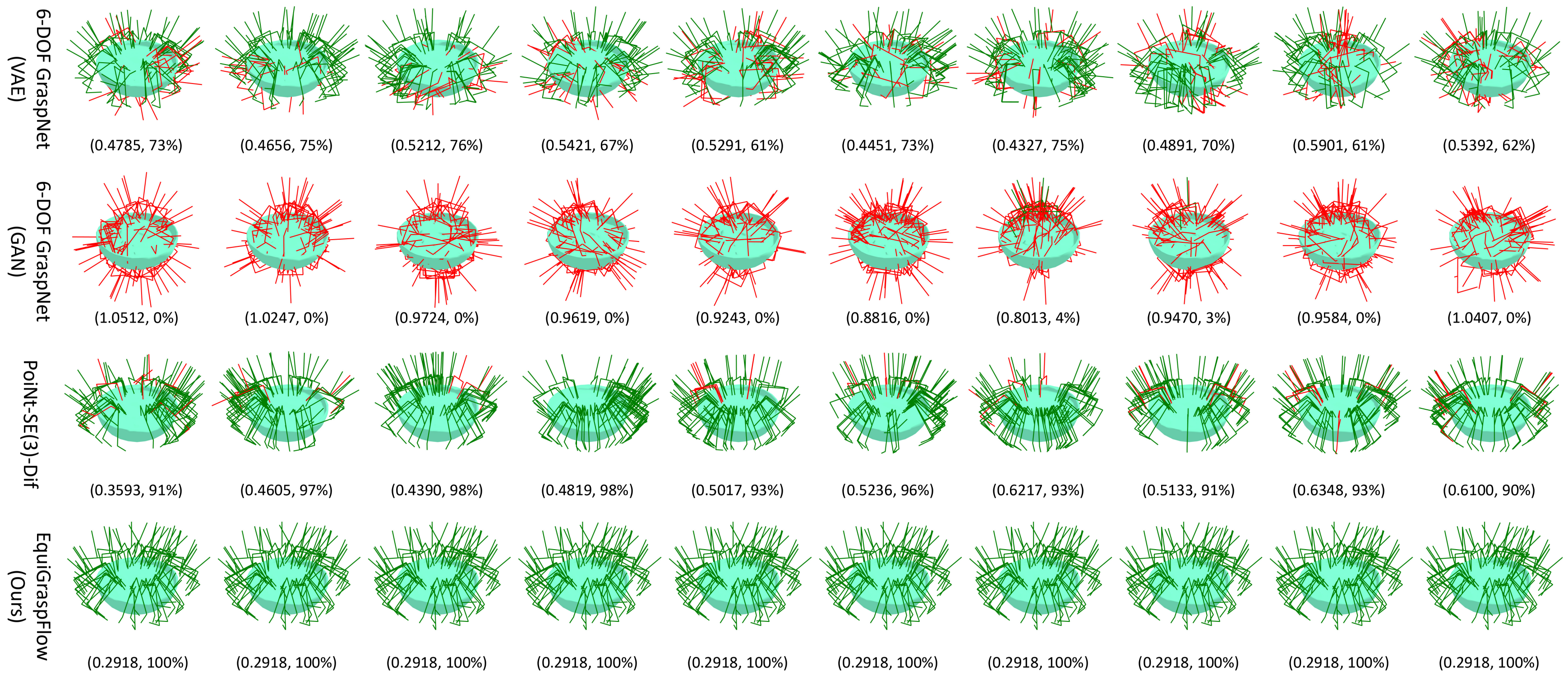

Figure 5: The generated grasp poses for the second bowl across ten rotations.

## References

[1] I. Katsman, A. Lou, D. Lim, Q. Jiang, S. N. Lim, and C. M. De Sa. Equivariant manifold flows. *Advances in Neural Information Processing Systems*, 34:10600–10612, 2021.

[2] C. Deng, O. Litany, Y. Duan, A. Poulenard, A. Tagliasacchi, and L. J. Guibas. Vector neurons: A general framework for so (3)-equivariant networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12200–12209, 2021.

[3] C. Eppner, A. Mousavian, and D. Fox. Acronym: A large-scale grasp dataset based on simulation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6222–6227. IEEE, 2021.

[4] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[5] R. T. Chen and Y. Lipman. Riemannian flow matching on general geometries. *arXiv preprint arXiv:2302.03660*, 2023.

[6] Q. Zheng, M. Le, N. Shaul, Y. Lipman, A. Grover, and R. T. Chen. Guided flows for generative modeling and decision making. *arXiv preprint arXiv:2311.13443*, 2023.

[7] J. Ho and T. Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[8] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *The Thrid International Conference on Learning Representations*, 2015.

[9] A. Mousavian, C. Eppner, and D. Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2901–2910, 2019.