GeoDEX: Geometric Diffusion Transformers for Multi-View Dexterous Pick-and-Place

Qilong Cheng
New York University
qc1007@nyu.edu

Yipeng Wang
New York University
yw6514@nyu.edu

Team [1]

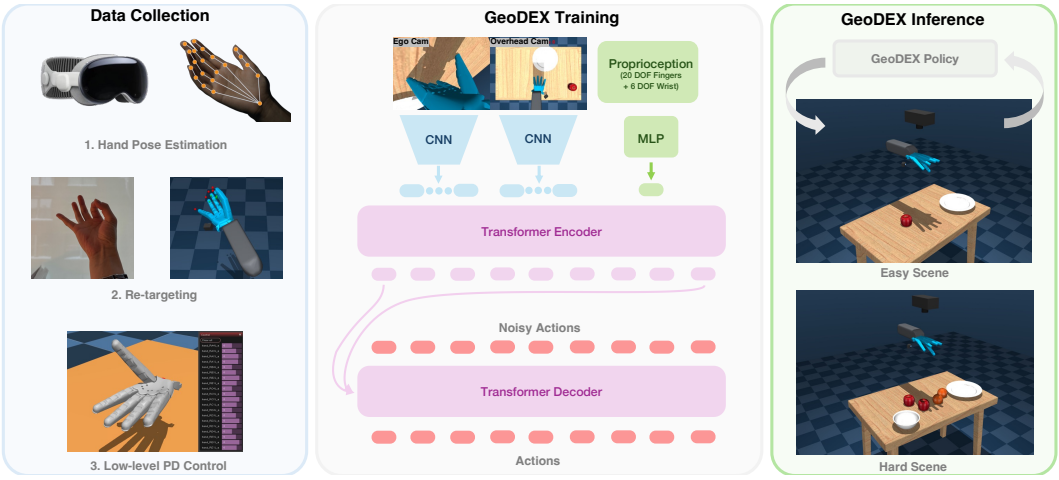

Figure 1: Overview of the project pipeline

Abstract

Dexterous pick-and-place is a contact-rich, partially observable visuomotor problem requiring geometric reasoning and stable interaction under occlusion and object pose variations. We propose **GeoDEX**, a geometric diffusion-transformer policy for dexterous pick-and-place that predicts temporally correlated action chunks from dual-view RGBD and proprioceptive observations for a 26 DoF robot (20 DoF hand + 6 DoF wrist). Human demonstrations are captured via Apple Vision Pro and retargeted through nonlinear optimization that preserves $\mathrm{SE}(3)$ wrist alignment and fingertip geometry in MuJoCo. The policy embeds geometric structure directly into diffusion training through a geodesic $\mathrm{SO}(3)$ loss, relative fingertip consistency constraints, and contact-aware regularization. We evaluate performance under progressively increasing scene complexity and conduct ablations on modality, action outputs, and geometric modeling. GeoDEX will demonstrate that incorporating explicit $\mathrm{SE}(3)$ structure into diffusion-transformer policies improves stability and generalization in contact-rich dexterous manipulation.

1 Introduction

Dexterous manipulation is a contact-rich, partially observable visuomotor problem that needs geometric reasoning, stable interaction, and robust perception under occlusion. The challenge increases under pose variation, clutter, and category-level sorting, where policies must simultaneously learn grasp mechanics and implicit object classification without explicit supervision. We study progressive pick-and-place, from single-object placement to multi-object and category-based sorting. Our approach employs a transformer-based diffusion policy that predicts temporally correlated action chunks from multi-view visual and proprioceptive observations.

Experiments are conducted on a 26 DoF robot hand (20 DoF fingers + 6 DoF floating wrist) equipped with two RGBD cameras: a wrist-mounted egocentric camera and an overhead camera. Human demonstrations are captured via Apple Vision Pro (27 $\mathrm{SE}(3)$ keypoints) and retargeted through nonlinear optimization. Policies are trained in MuJoCo simulation, that map RGBD and proprioceptive inputs to action output, enabling end-to-end visuomotor control.

2 Related Work

Diffusion-based visuomotor policies have become a strong paradigm for multimodal action generation in robot manipulation. Diffusion Policy [1] showed that denoising-based action modeling improves stability over autoregressive methods, while transformer backbones enable long-horizon and chunked prediction [2, 3, 4]. Large-scale vision-language-action (VLA) models such as RT-1 and RT-2 [5, 3] further demonstrated the benefits of pretrained visual representations for manipulation generalization. In dexterous settings, prior work emphasizes reinforcement learning (RL) with domain randomization [6], demonstration-augmented policy learning [7], and cross-embodiment transfer via morphology-aware or shared latent representations [8, 9]. Geometric control methods establish stable tracking on $\mathrm{SE}(3)$ and $\mathrm{SO}(3)$ manifolds [10, 11], yet most generative manipulation policies regress Euclidean pose parameters and neglect Lie group structure.

Building on diffusion transformers [12, 13, 14], GeoDEX incorporates explicit $\mathrm{SE}(3)$ geometry, relative fingertip constraints, and contact-aware regularization into chunked diffusion policies for multi-view dexterous manipulation.

3 Proposed Approach

Figure 1 provides an overview of the project pipeline and method we are proposing.

3.1 Data Collection

Human demonstrations are captured via Apple Vision Pro hand tracking using the MIT CSAIL Streamer application [15, 16]. The 27 tracked keypoints are then retargeted to the 20 DoF robot hand using nonlinear optimization. [17].

Each trajectory $\tau = \{(o_t, a_t)\}_{t=1}^{T}$ contains synchronized observations and desired action targets. The observation is defined as $o_t = \left(I_t^{\text{ego}}, I_t^{\text{over}}, q_t, x_t^w, c_t, s_t^{obj}\right)$, where $I_t^{\text{ego}}, I_t^{\text{over}}$ are RGBD images, $q_t \in \mathbb{R}^{20}$ finger joint positions, $x_t^w \in \mathrm{SE}(3)$ wrist pose. The c_t contact signals, and s_t^{obj} object states are also collected for potential architecture changes and multi-modality explorations. The action $a_t = (q_t^*, x_t^{w*})$ corresponds to desired joint and wrist targets prior to PD control.

3.2 Robot Hand Retargeting

Retargeting to the 20-DoF Kyber hand is formulated as a nonlinear optimization problem, following prior optimization-based dexterous retargeting approaches [18, 9]. Let $q \in \mathbb{R}^{20}$ denote joint angles and $x_i(q)$ the forward-kinematics fingertip positions. Given human fingertip targets h_i, we minimize

$$\mathcal{L}_{IK} = W_c(1 - \cos(\mathbf{r}, \mathbf{h})) + W_r \sum_i \|x_i(q) - h_i\|^2 + W_s\|q - q_{prev}\|^2 + W_{reg}\|q\|^2, \quad (1)$$

enforcing wrist alignment, fingertip consistency, temporal smoothness, and joint regularization. This produces smooth, geometrically consistent teleoperation trajectories across embodiments.

3.3 Low-Level PD Controller

Finger joints are controlled via independent PD control, $\tau_i = k_p(q_i^* - q_i) - k_d\dot{q}_i$. The wrist is modeled as a free-floating joint and controlled through Cartesian impedance. With $e_p = p^* - p$ and $e_R = \log(R^\top R^*)$, the spatial wrench $F = K_p e_p - K_v v$, $T = K_r e_R - K_\omega \omega$ is applied at the palm frame.

3.4 Policy Formulation

Dexterous pick-and-place is modeled as a partially observable visuomotor problem. The policy learns a conditional diffusion model $\pi_\theta(a_{t:t+H} \mid o_{t-k:t})$, predicting temporally correlated action chunks of horizon H from observation history k. Observations include dual-view RGBD images and proprioceptive state (20 finger joints + 6-DoF wrist pose). Actions are 29D per timestep (20 joint targets + 9D wrist command). Diffusion is performed in normalized action space to capture multimodal contact-rich trajectories.

3.5 Architecture

Both camera streams are encoded with a shared ResNet18 backbone [19]. Visual features are fused with proprioception and diffusion timestep embeddings, forming a token sequence processed by a transformer encoder [20]. A diffusion head predicts denoised action chunks following DDPM training [12, 1, 2], with objective $\mathcal{L}_{\text{diffusion}} = \|\hat{\epsilon} - \epsilon\|^2$. To improve cross-embodiment robustness, we introduce structured finger dropout [8, 21]. With probability p_{drop}, a finger's proprioceptive inputs are masked and a binary mask token appended. The diffusion loss is computed only over active action dimensions $\mathcal{A}_{active}$:

$$\mathcal{L}_{\text{diffusion}}^{\text{mask}} = \|\hat{\epsilon}_{\mathcal{A}_{active}} - \epsilon_{\mathcal{A}_{active}}\|^2. \tag{2}$$

3.6 Loss Functions

We incorporate geometric regularization on $\mathrm{SE}(3)$ [10, 11]. Rotational alignment uses a geodesic loss $\mathcal{L}_{\text{rot}} = \|\log(R_{\text{pred}} R_{\text{gt}}^\top)\|_2$. Relative fingertip geometry is enforced via $\mathcal{L}_{\text{rel}} = \sum_{i,j} \|(x_i - x_j) - (x_i^* - x_j^*)\|^2$ [8, 9], restricted to active fingers under dropout. Contact stability is regularized using MuJoCo contact signals [17] to penalize slip and abrupt transitions, consistent with prior dexterous RL approaches [6, 7]. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{diffusion}}^{\text{mask}} + \lambda_1 \mathcal{L}_{\text{rot}} + \lambda_2 \mathcal{L}_{\text{contact}} + \lambda_3 \mathcal{L}_{\text{rel}}^{\text{mask}}. \tag{3}$$

4 Expected Results and Experiments

We plan to evaluate performance of the method under increasing scene complexity from single-object placement to category-level sorting. Metrics used for comparison will include task success, grasp success, contact stability, and trajectory smoothness. In addition, we will perform controlled ablation studies to isolate representation and architectural effects, comparing single- versus dual-view perception, RGB versus RGBD inputs, absolute versus relative action output, Euclidean versus $\mathrm{SE}(3)$ modeling, with and without contact-aware regularization, and structured finger dropout to evaluate cross-embodiment robustness. Finally, we plan to conduct real-world deployment experiments in collaboration with the Courant, using their off-the-shelf robot hands. It would help us evaluate sim-to-real transfer performance, cross-embodiment generalization, real-world policy robustness and grasp stability.

5 Conclusion

GeoDEX plan to investigate how geometric structure can be embedded into diffusion-based visuomotor policies for dexterous manipulation. By integrating multi-view perception, SE(3) representation, cross-embodiment generalization, and spatially structured objectives, we aim to improve stability, precision, and generalization in contact-rich pick-and-place tasks.

References

[1] Cheng Chi, Siyuan Feng, Yilun Du, Zhengwu Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems (RSS)*, 2023.

[2] Tony Z. Zhao, Aviral Kumar, Sergey Levine, and Chelsea Finn. Action chunking with transformers. In *Conference on Robot Learning (CoRL)*, 2023.

[3] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In Jie Tan, Marc Toussaint, and Kourosh Darvish, editors, *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR, 2023.

[4] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning (CoRL)*, 2023.

[5] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Giorgos Dabisias, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Dmitry Kalashnikov, Sergey Levine, and Vincent Vanhoucke. Rt-1: Robotics transformer for real-world control at scale. In *Robotics: Science and Systems (RSS)*, 2022.

[6] Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Józefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research (IJRR)*, 39(1):3–20, 2020.

[7] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. In *Robotics: Science and Systems (RSS)*, 2018.

[8] Heng Zhang, Kevin Yuchen Ma, Mike Zheng Shou, Weisi Lin, and Yan Wu. Cross-embodiment dexterous hand articulation generation via morphology-aware learning, 2025.

[9] Wentao Yuan, Yang Gao, Jian Huang, Rong Yang, Kuan-Ting Hsiao, Rui Huang, Fu-Jen Chang, Hsueh-Ming Su, Hongzhao Cai, Abhinav Gupta, Chelsea Finn, Yuke Tian, Junfeng Zhou, Yuke Zhu, Tianhao Liu, Yuke Zhao, Abhinav Gupta, and Edward H. Adelson. Cross-embodiment dexterous manipulation via shared latent representations. In *International Conference on Learning Representations (ICLR)*, 2024.

[10] Francesco Bullo and Andrew D. Lewis. *Geometric Control of Mechanical Systems*. Springer, 2005.

[11] Taeyoung Lee, Melvin Leok, and N. Harris McClamroch. Geometric tracking control of a quadrotor uav on se(3). In *IEEE Conference on Decision and Control (CDC)*, 2010.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[13] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

[14] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[15] Apple Inc. Apple vision framework and arkit hand tracking documentation. https://developer.apple.com/documentation/visionos/, 2023. Accessed 2026.

[16] MIT CSAIL. Mit csail vision pro streamer: Real-time hand tracking and pose streaming. https://github.com/mit-csail/vision-pro-streamer, 2024. Accessed 2026.

[17] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5026–5033. IEEE, 2012.

[18] Rui Huang, Wentao Yuan, Zipeng Fu, Yuke Zhu, and Abhinav Gupta. Dexretarget: Learning cross-hand dexterous manipulation via optimization-based retargeting. In *Conference on Robot Learning (CoRL)*, 2022.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[21] Zhenyu Wei, Yunchao Yao, and Mingyu Ding. One hand to rule them all: Canonical representations for unified dexterous manipulation. *arXiv preprint arXiv:2602.16712*, 2026.