Learning Geometry-Aware Nonprehensile Pushing and Pulling with Dexterous Hands

Anonymous Submission

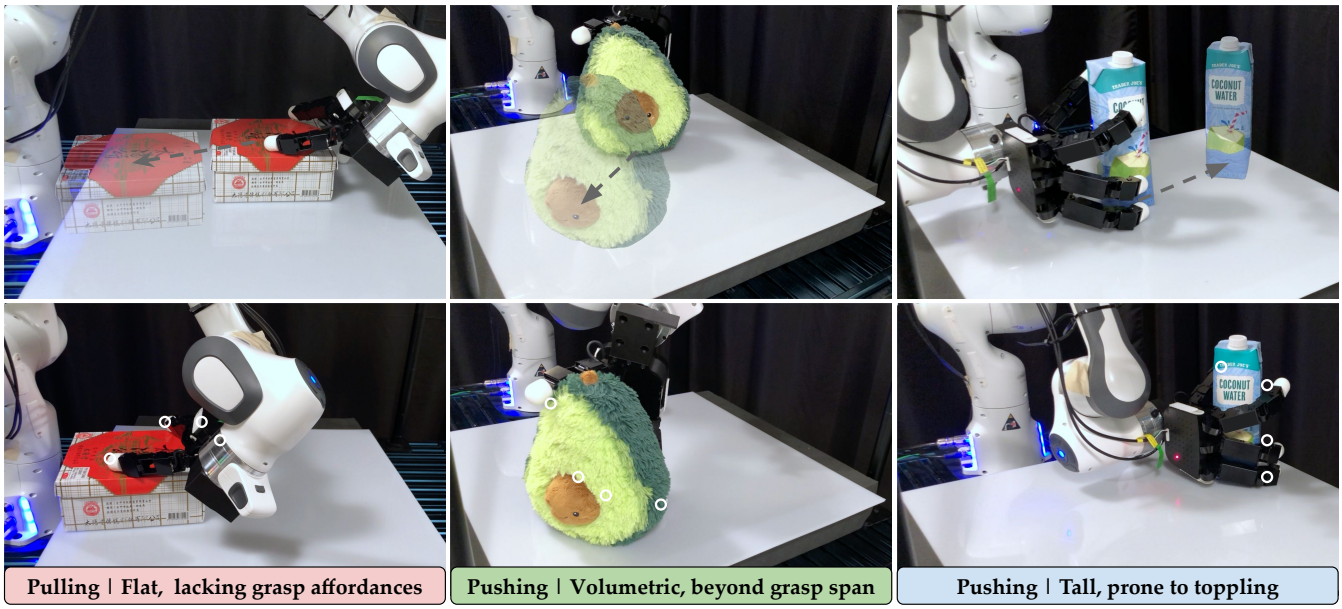


Fig. 1: Three examples of nonprehensile manipulation using GD2P with a 4-finger, 16-DOF Allegro Hand. The top row shows the starting object configuration with its goal rendered as a transparent overlay, while the bottom row shows the result after the robot's motion. GD2P synthesizes diverse hand poses conditioned on object geometry, handling flat (left), volumetric (middle), and tall (right) objects. Grey arrows represent the transporting direction, whereas white volumetric dots mark the estimated fingertip contact with the object.

***Abstract*— Nonprehensile manipulation, such as pushing and pulling, enables robots to move, align, or reposition objects that may be difficult to grasp due to their geometry, size, or relationship to the robot or the environment. Much of the existing work in nonprehensile manipulation relies on parallel-jaw grippers or tools such as rods and spatulas. In contrast, multi-fingered dexterous hands offer richer contact modes and versatility for handling diverse objects to provide stable support over the objects, which compensates for the difficulty of modeling the dynamics of nonprehensile manipulation. Therefore, we propose Geometry-aware Dexterous Pushing and Pulling (GD2P) for nonprehensile manipulation with dexterous robotic hands. We study pushing and pulling by framing the problem as synthesizing and learning pre-contact dexterous hand poses that lead to effective manipulation. We generate diverse hand poses via contact-guided sampling, filter them using physics simulation, and train a diffusion model conditioned on object geometry to predict viable poses. At test time, we sample hand poses and use standard motion planners to select and execute pushing and pulling actions. We perform extensive real-world experiments with an Allegro Hand and a LEAP Hand, demonstrating that GD2P offers a scalable route for generating dexterous nonprehensile manipulation motions with its applicability to different hand morphologies. Our project website is available at: *geodex2p.github.io*.**

I. Introduction

Nonprehensile actions are fundamental to how humans and robots interact with the physical world [4], [5], [6], [7]. These actions permit the manipulation of objects that may be too large, heavy, or geometrically complex to grasp directly. While there has been tremendous progress in nonprehensile robot manipulation [8], [9], [10], [11], [12], most work uses simple end-effectors such as parallel-jaw grippers, rods [13], [14], or spatulas [15]. In contrast, multi-fingered hands with high degrees-of-freedom (DOF) such as the Allegro Hand or LEAP Hand [16] enable contact patterns that can be especially useful for stabilizing complex, awkward, or top-heavy objects, or for coordinating contact across multiple objects, compensating for the challenges of modeling nonprehensile manipulation dynamics. However, despite their promise and recent progress [17] , leveraging high-DOF hands for nonprehensile manipulation remains relatively underexplored due to the challenges of modeling hand-object relationships and planning feasible contact-rich motions.

In this paper, we study pushing and pulling objects using the 4-finger, 16-DOF Allegro and LEAP Hands. We select pushing and pulling as representative tasks of nonprehensile manipulation because they are more commonly used for manipulating general daily objects and are more amenable to scaling. Our insight is to recast this problem into one of synthesizing effective pre-contact hand poses, an approach inspired by recent success in generating large-scale datasets for dexterous manipulation [18], [19], [20], [21], [22], [23].

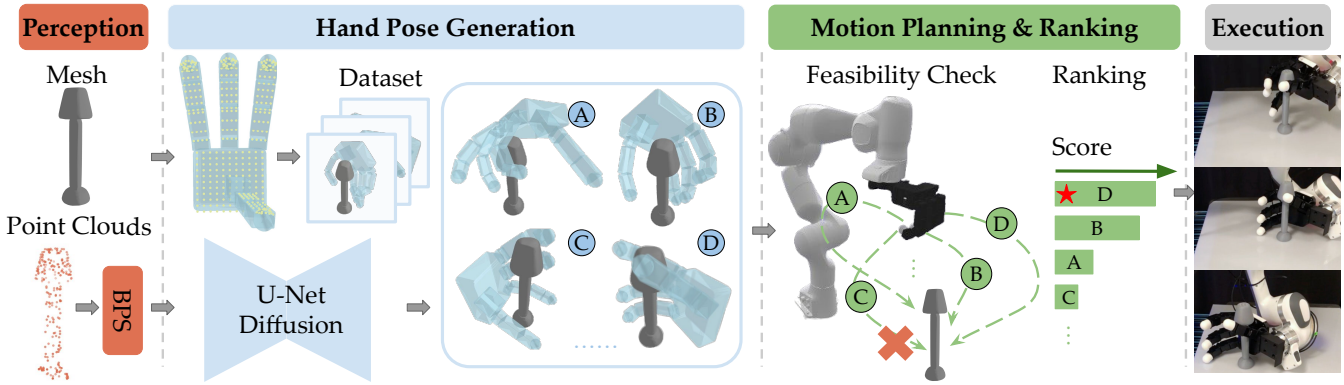


Fig. 2: Overview of GD2P. We present a large-scale dataset of hand poses specifically for pushing or pulling, and leverage it to train a diffusion model. During execution time, given an object, we obtain its basis point set representation [1] and pass that to our trained diffusion model, which uses the architecture from [2]. This model synthesizes diverse floating pre-contact hand poses formed from our large-scale data generation pipeline (Sec. III-A). Given these hand poses, we then check their feasibility in a physics simulator by adding the arm back in and performing motion planning [3]. We rank the feasible hand poses (e.g., "C" is infeasible in the example here since the motion planner detects an unavoidable collision between the arm and the table.) and select the best performing one (e.g., "D" in our example with no collision detected, while optimally facilitating the pushing direction.) and execute it in the real world.

We propose a scalable pipeline for generating hand poses for pushing and pulling objects. This involves contact-guided optimization and validation via GPU-accelerated physics simulation with IsaacGym [24]. These filtered hand poses are then used to train a generative diffusion model conditioned on object geometry, represented using basis point sets [1].

At test time, we use visual data to reconstruct an object mesh in physics simulation. The trained diffusion model uses this mesh to generate diverse hand poses for pushing or pulling. We then validate the resulting hand poses in simulation, and execute the best-performing action in the real world. We call this pipeline **G**eometry-aware **D**exterous **P**ushing and **P**ulling (GD2P) with multi-fingered hands. Fig. 1 shows several real-world examples where the hand pose differs depending on object geometry. Overall, our experimental results across diverse daily objects demonstrate that GD2P is a promising approach for generalizable object pushing and pulling. It outperforms alternative methods such as querying the nearest hand pose in our data or using a fixed spatula-like hand pose, highlighting the need for a diffusion model to generate diverse hand poses.

To summarize, the contributions of this paper include:

- A scalable pipeline for generating and filtering dexterous hand poses for nonprehensile pushing and pulling.
- A diffusion model for geometry-conditioned hand pose prediction for nonprehensile pushing and pulling.
- A motion planning framework to execute these poses in the real world, with results across 840 trials showing that GD2P outperforms alternative methods.
- A dataset of 1.3 million hand poses for nonprehensile pushing and pulling across 2.3k objects with corresponding canonical point cloud observations.

II. Problem Statement and Assumptions

We study nonprehensile object pushing and pulling on a flat surface using a single-arm robot with a high-DOF multi-finger dexterous hand (e.g., the Allegro Hand). By "nonprehensile," we emphasize the distinction from "prehensile pushing [25]." We assume that there exists one object O on the surface with configuration $S_{\text{obj}} \in SE(3)$, and that the surface's friction properties facilitate object pushing. We use P to indicate the object's point cloud sampled from its surface. Let $\mathcal{H}$ be the space of possible pushing and pulling hand poses, where $H \in \mathcal{H}$ is defined as $H = (\theta, T)$. Here, $\theta \in \mathbb{R}^d$ is the joint configuration of the d-DOF robot hand, and $T \in SE(3)$ is the end-effector pose of the robot's wrist consisting of translation and orientation. A *trial* is an instance of pushing or pulling, defined by a given direction $u_{\text{dir}} \in \mathbb{R}^3$ (with z-component of 0) resulting in the target object position as $u_{\text{targ}} \in \mathbb{R}^3$. The objective is to generate a hand pose H such that, if a motion planner moves the hand to H and then translates it along u_{dir}, the object moves closer to the target u_{targ}. The object's distance to u_{targ} must be below a threshold for a trial to be considered a success.

III. Method

GD2P consists of the following steps. First, we generate a large dataset of hand poses for pushing and pulling (Sec. III-A). Second, we use this data to train a diffusion model to synthesize hand poses conditioned on object geometry (Sec. III-B). Third, during deployment, we generate hand poses and perform motion planning (Sec. III-C).

A. Dataset Generation for Dexterous Pushing and Pulling

We first generate diverse hand poses for pushing and pulling various objects in simulation. To do this, we take inspiration from prior work on generating diverse hand poses for *grasping* [18], [19], [20], [21], [23], [26] by casting the hand synthesis problem as minimizing an energy function via optimization [27]. Unlike those works, our focus is on pushing and pulling instead of grasping. To enable hand pose optimization, we define a set of candidate contact points sampled across the hand surface. Different regions of the hand have different candidate points to encourage broad contact across the palm and fingers. For the palm and finger (excluding fingertips), we sample points uniformly over the

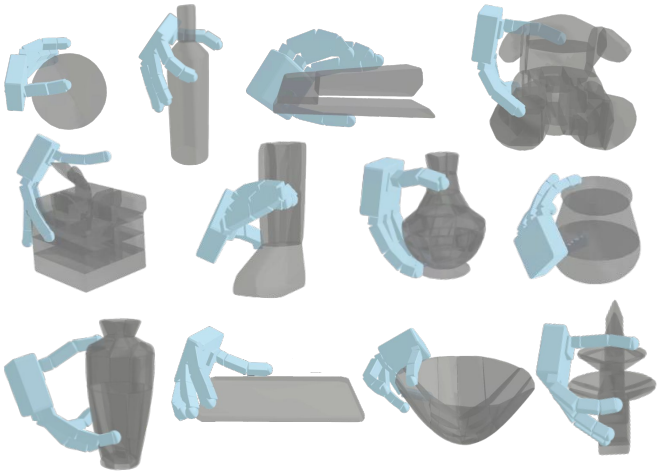


Fig. 3: Examples of pushing and pulling hand poses from optimizing our energy function (Eq. 1). These have all been validated in IsaacGym simulation. In all examples, the intended object pushing direction is to the right. These data points are used to train our diffusion model (see Sec. III-B).

rigid body surface. For the fingertips, we sample from a denser set of points uniformly on the unit hemisphere for each tip. See the website for candidate contact points details.

With the sampled contact point candidates, we run an optimization algorithm following the sampling strategy from [18], [19] that iteratively minimizes an energy function E to generate hand poses. We adapt the energy function from [18] to suit our manipulation tasks, resulting in:

$$E = E_{\text{fc}} + w_{\text{dis}} E_{\text{dis}} + w_{\text{j}} E_{\text{j}} + w_{\text{pen}} E_{\text{pen}} + w_{\text{dir}} E_{\text{dir}} + w_{\text{arm}} E_{\text{arm}} \tag{1}$$

where E_{fc} is a force closure estimator [27], E_{dis} penalizes hand-to-object distance (thus encouraging proximity), E_{j} penalizes joint violations, and E_{pen} penalizes penetration between hand-object, hand-table and hand self-collision contacts. See [18], [19] for further details. The w terms are all scalar coefficients; we adopt the values from prior work and tune the weights (see the Appendix on our website) for the following two new terms. To adapt the energy from Eq. 1 to pushing or pulling in a particular direction $u_{\text{dir}} \in \mathbb{R}^3$, we introduce E_{dir} and E_{arm}, which use the normal vector of the palm $v_{\text{palm}} \in \mathbb{R}^3$. The E_{dir} term encourages v_{palm} to align with u_{dir}, and E_{arm} encourages hand poses that are kinematically feasible when attached to the robot arm. Formally, we define E_{dir} and E_{arm} as:

$$E_{\text{dir}} = -\frac{u_{\text{dir}}^{\mathsf{T}} v_{\text{palm}}}{\|u_{\text{dir}}\|_2 \|v_{\text{palm}}\|_2}; \; E_{\text{arm}} = \max\big(0, (v_{\text{palm}})_z\big) \tag{2}$$

where $(v_{\text{palm}})_z$ is the z-component of the palm's normal vector (in the world frame). Intuitively, aligning u_{dir} and v_{palm} promotes more stable object-palm directional contact. Furthermore, if the palm faces upwards, then the rest of the arm must be below it. Thus, it is likely to lead to an infeasible robot configuration due to robot-table intersections, so E_{arm} is nonzero (i.e., worse). To inject randomness (and thus diversity) in the sampling process, we randomly resample a subset of the contact point indices from the set of valid candidates when generating a new hand pose. We use RMSProp [28] to update translation, rotation and joint angles with step size decay, then minimize the energy function with Simulated Annealing [29] to adjust parameters.

Hand Pose Validation in Simulation. After optimizing contact points to generate candidate hand poses, we must *validate* whether they can lead to successful pushing or pulling. To do this, we use IsaacGym [24], a GPU-accelerated physics simulator that has been used in prior work for filtering grasp poses [18], [19]. We define a push or a pull as successful if, after executing a 20 cm translation, the object's center is within 3 cm of the target position *and* the object's orientation changes by no more than 45 degrees relative to its original configuration. The optimization process has a low success rate because it does not account for the full dynamics of pushing and pulling. Thus, we augment successful hand poses by adding slight noise to the pose parameters. From extensive parallel experiments, we generate a dataset containing 2.3k objects with 1.3 million successful hand poses. See the Appendix on our website for more details.

B. Training a Diffusion Model to Predict Hand Poses

To generate hand poses, we adapt a conditional U-Net [30] from the diffusion policy architecture [14], and train it with the Denoising Diffusion Probabilistic Models (DDPM) objective [31]. Diffusion models are well-suited for this task as they can learn complex, high-dimensional distributions. The forward process gradually adds Gaussian noise to the hand configuration H, while the reverse process reconstructs the original pose H by iteratively denoising conditioned on the object's geometry. The model is trained to minimize denoising error. To represent the observation, we use a 4096-dimensional Basis Point Set (BPS) [1] representation $B \in \mathbb{R}^{4096}$ based on the object's point cloud P. For 3D-printed objects, we use their known meshes to directly compute their BPS representation. For the other objects, we follow the pipeline proposed in [32] to obtain real-world object point clouds (and thus, the BPS). Specifically, we reconstruct object meshes by using Nerfstudio [33] and we use Stable Normal [34] to generate normal maps. Then, we employ 2D Gaussian Splatting [35] to obtain the point clouds. This representation, which is also used in [2], [18], encodes each object as a fixed-length vector of shortest distances between canonical basis points and the points in P. BPS captures geometric properties in a compact manner and simplifies the design of the diffusion model. Given this trained diffusion model, at test time it can be used to generate diverse hand poses which we can select for motion planning. See Fig. 2 and Appendix on our website for more information.

C. Arm-Hand Motion Planning and Evaluation

During deployment, the diffusion model generates candidate hand poses. We then integrate the robotic arm into full arm-hand motion planning to select hand poses which are kinematically feasible and avoid environment collisions, such as arm-table intersections (which are not considered in Sec. III-A). See Fig. 2 (right half) for an overview. Each hand

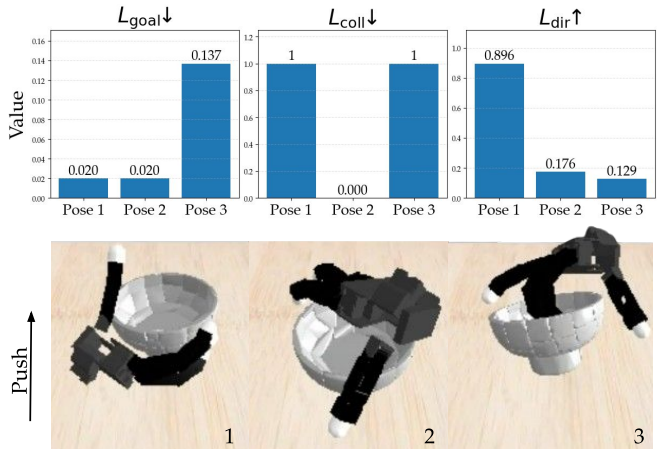


Fig. 4: Visualization of $L_{\text{goal}}, L_{\text{coll}}$, and L_{dir} values in $V(H)$ from Eq. 3 on three simulated hand poses. See Sec. III-C and Sec. VI-B.3 for more details.

pose $H = (\theta, T)$ is initially expressed in the object frame. We use the object's initial configuration S_{obj} and intended direction u_{dir} to transform H to the world frame, and supply that to the cuRobo planner [3] to generate a complete motion plan for the robotic arm. In this process, we discard infeasible trajectories (and thus, the associated hand poses) to only keep the feasible arm-hand trajectories. To select which of the feasible trajectories to execute, we associate each with a custom analytical score V, defined as:

$$V(H = (\theta, T)) = \alpha L_{\text{goal}} + \beta L_{\text{coll}} + \gamma L_{\text{dir}}, \qquad (3)$$

where L_{goal} measures the Euclidean distance between the object's final position and the target position, L_{coll} indicates whether a collision occurred during execution (1 if a collision occurs, 0 otherwise), and L_{dir} encourages the palm's orientation to align with the pushing direction. For L_{dir}, we set it equal to the E_{dir} term from the energy function (Eq. 1). The α, β, and γ are hyperparameters tuned empirically.

To provide additional context, Fig. 4 presents three candidate hand poses along with their corresponding scores on L_{goal}, L_{coll} (where lower values are preferable), and L_{dir} (where higher values are preferable). Hand pose 1 achieves the highest score on L_{dir}, indicating effective alignment with the pushing direction, yet it results in an avoidable hand–table collision. Hand pose 3 exhibits a low score on L_{goal}, suggesting limited effectiveness in pushing the object toward the target. Balancing all three objectives, the motion planner would select hand pose 2 for execution in the real world. We tune these weight terms according to their influence on pushing feasibility and overall task success.

Multi-step Planning. While we mainly study GD2P for single open-loop pushes (or pulls) to targets, our framework naturally extends to multi-step planning. In scenarios with obstacles, we first compute a collision-free global path using RRT* [36]. Then, we sequentially plan hand poses to reach each intermediate waypoint. Given an object, the same hand pose may be feasible only in certain pushing or pulling directions due to robot and hand kinematics. The waypoints from RRT* may require planning pushes across challenging directions, highlighting the importance of generating diverse hand poses for varying object positions and directions.

Fig. 5: The objects we use in our real-world experiments, including 3D printed and daily objects. See Sec. VI-B.2 for more details.

IV. Experiments

Through simulation and real-world experiments, we aim to investigate the following questions: (1) Can we learn feasible and effective hand poses from our large-scale dataset? (2) Is GD2P more effective and robust for dexterous pushing and pulling compared to baselines and ablations? (3) Can GD2P serve as a reliable module for downstream manipulation tasks such as multi-step pushing around obstacles? (4) Does GD2P perform consistently across different hand morphologies? Due to the space limitation, we include all the experiments in Section VI-B in Appendix.

V. Limitations and Conclusion

While promising, the GD2P approach has some limitations that motivate exciting directions for future work. First, our study considers only pushing and pulling as examples of nonprehensile manipulation and thus does not exhaustively characterize all possible nonprehensile manipulation procedures; we aim to expand to a broader repertoire (e.g., tilting, rolling) that leverages environmental interactions with dexterous hands to enhance nonprehensile dexterity. Second, we aim to develop a closed-loop nonprehensile manipulation policy that adapts in real time to different object physical properties and environmental changes.

To conclude, we propose GD2P, a dataset and method for nonprehensile object pushing and pulling using high-DOF robotic hands, such as the Allegro and LEAP Hands. Our extensive real-world results show that GD2P enables diverse and effective pre-contact hand poses for different combinations of objects and pushing directions. We also demonstrate its usage for multi-step planning. We hope this inspires future work on dexterous nonprehensile manipulation.

References

[1] S. Prokudin, C. Lassner, and J. Romero, "Efficient learning on point clouds with basis point sets," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[2] Z. Weng, H. Lu, D. Kragic, and J. Lundell, "Dexdiffuser: Generating dexterous grasps with diffusion models," in *IEEE Robotics and Automation Letters (RA-L)*, 2024.

[3] B. Sundaralingam et al., "Curobo: Parallelized collision-free minimum-jerk robot motion generation," *arXiv preprint arXiv:2310.17274*, 2023.

[4] M. T. Mason, "Mechanics and Planning of Manipulator Pushing Operations," in *International Journal of Robotics Research*, 1986.

[5] K. Lynch, "Nonprehensile robotic manipulation: Controllability and planning," Ph.D. dissertation, Carnegie Mellon University, 1996.

[6] M. T. Mason, "Progress in Nonprehensile Manipulation," in *International Journal of Robotics Research (IJRR)*, 1999.

[7] K. M. Lynch and M. T. Mason, "Dynamic nonprehensile manipulation: Controllability, planning, and experiments," in *International Journal of Robotics Research (IJRR)*, 1999.

[8] W. Zhou and D. Held, "Learning to Grasp the Ungraspable with Emergent Extrinsic Dexterity," in *Conference on Robot Learning (CoRL)*, 2022.

[9] W. Zhou, B. Jiang, F. Yang, C. Paxton, and D. Held, "HACMan: Learning Hybrid Actor-Critic Maps for 6D Non-Prehensile Manipulation," in *Conference on Robot Learning (CoRL)*, 2023.

[10] B. Jiang, Y. Wu, W. Zhou, C. Paxton, and D. Held, "Hacman++: Spatially-grounded motion primitives for manipulation," in *Robotics: Science and Systems (RSS)*, 2024.

[11] Y. Cho, J. Han, Y. Cho, and B. Kim, "Corn: Contact-based object representation for nonprehensile manipulation of general unseen objects," in *International Conference on Learning Representations (ICLR)*, 2024.

[12] J. Lyu, Z. Li, X. Shi, C. Xu, Y. Wang, and H. Wang, "Dywa: Dynamics-adaptive world action model for generalizable non-prehensile manipulation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[13] K. Zhang, B. Li, K. Hauser, and Y. Li, "Adaptigraph: Material-adaptive graph-based neural dynamics for robotic manipulation," in *Robotics: Science and Systems (RSS)*, 2024.

[14] C. Chi et al., "Diffusion Policy: Visuomotor Policy Learning via Action Diffusion," in *Robotics: Science and Systems (RSS)*, 2023.

[15] Y. Wang, Y. Li, K. Driggs-Campbell, L. Fei-Fei, and J. Wu, "Dynamic-resolution model learning for object pile manipulation," in *Robotics: Science and Systems (RSS)*, 2023.

[16] K. Shaw, A. Agarwal, and D. Pathak, "LEAP Hand: Low-Cost, Efficient, and Anthropomorphic Hand for Robot Learning," in *Robotics: Science and Systems (RSS)*, 2023.

[17] Y. Wang, Y. Li, Y. Yang, and Y. Chen, "Dexterous non-prehensile manipulation for ungraspable object via extrinsic dexterity," *arXiv preprint arXiv:2503.23120*, 2025.

[18] T. G. W. Lum et al., "Get a Grip: Multi-Finger Grasp Evaluation at Scale Enables Robust Sim-to-Real Transfer," in *Conference on Robot Learning (CoRL)*, 2024.

[19] R. Wang et al., "DexGraspNet: A large-scale robotic dexterous grasp dataset for general objects based on simulation," in *International Conference on Robotics and Automation (ICRA)*, 2023.

[20] J. Zhang et al., "DexGraspNet 2.0: Learning Generative Dexterous Grasping in Large-scale Synthetic Cluttered Scenes," in *Conference on Robot Learning (CoRL)*, 2024.

[21] Y. Xu et al., "UniDexGrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[22] W. Wan et al., "UniDexGrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[23] Y. Li et al., "Grasp multiple objects with one hand," in *IEEE Robotics and Automation Letters (RA-L)*, 2024.

[24] V. Makoviychuk et al., "Isaac Gym: High Performance GPU-Based Physics Simulation For Robot Learning," *arXiv preprint arXiv:2108.10470*, 2021.

[25] P. Perugini, J. Lundell, K. Friedl, and D. Kragic, "Pushing everything everywhere all at once: Probabilistic prehensile pushing," *IEEE Robotics and Automation Letters*, 2025.

[26] S. He et al., "Sequential multi-object grasping with one dexterous hand," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

[27] T. Liu, Z. Liu, Z. Jiao, Y. Zhu, and S.-C. Zhu, "Synthesizing diverse and physically stable grasps with arbitrary hand structures using differentiable force closure estimator," in *IEEE Robotics and Automation Letters (RA-L)*, 2022.

[28] T. Tieleman and G. Hinton, *Lecture 6.5—rmsprop: Divide the gradient by a running average of its recent magnitude*, COURSERA: Neural Networks for Machine Learning, 2012.

[29] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, 1983.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *MICCAI*, 2015.

[31] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Neural Information Processing Systems*, 2020.

[32] H. Lou et al., "Robo-gs: A physics consistent spatial-temporal model for robotic arm with hybrid representation," *arXiv preprint arXiv:2408.14873*, 2024.

[33] M. Tancik et al., "Nerfstudio: A modular framework for neural radiance field development," in *SIGGRAPH*, 2023.

[34] C. Ye et al., "Stablenormal: Reducing diffusion variance for stable and sharp normal," *arXiv preprint arXiv:2406.16864*, 2024.

[35] B. Huang, Z. Yu, A. Chen, A. Geiger, and S. Gao, "2D Gaussian Splatting for Geometrically Accurate Radiance Fields," in *SIGGRAPH*, ACM, 2024.

[36] S. Karaman and E. Frazzoli, "Sampling-based algorithms for optimal motion planning," *International Journal of Robotics Research (IJRR)*, 2011.

[37] N. Chavan-Dafle et al., "Extrinsic dexterity: In-hand manipulation with external forces," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2014.

[38] W. Yang and M. Posa, "Dynamic on-palm manipulation via controlled sliding," in *Robotics: Science and Systems (RSS)*, 2024.

[39] S. Chen, A. Wu, and C. K. Liu, "Synthesizing dexterous nonprehensile pregrasp for ungraspable objects," in *ACM SIGGRAPH*, 2023.

[40] A. Wu, R. Wang, S. Chen, C. Eppner, and C. K. Liu, "One-Shot Transfer of Long-Horizon Extrinsic Manipulation Through Contact Retargeting," *arXiv preprint arXiv:2404.07468*, 2024.

[41] T. Yonemaru, W. Wan, T. Nishimura, and K. Harada, "Learning to Group and Grasp Multiple Objects," *arXiv preprint arXiv:2502.08452*, 2025.

[42] J. J. Liu, Y. Li, K. Shaw, T. Tao, R. Salakhutdinov, and D. Pathak, "Factr: Force-attending curriculum training for contact-rich policy learning," in *Robotics: Science and Systems (RSS)*, 2025.

[43] M. Liu, Z. Pan, K. Xu, K. Ganguly, and D. Manocha, "Deep differentiable grasp planner for high-dof grippers," in *Robotics: Science and Systems (RSS)*, 2020.

[44] A. T. Miller and P. K. Allen, "Graspit! a versatile simulator for robotic grasping," *IEEE Robotics & Automation Magazine*, 2004.

[45] H. Jiang, S. Liu, J. Wang, and X. Wang, "Hand-object contact consistency reasoning for human grasps generation," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[46] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *International Conference on Learning Representations*, 2014.

[47] D. Turpin et al., "Grasp'd: Differentiable contact-rich grasp synthesis for multi-fingered hands," in *European Conference on Computer Vision (ECCV)*, 2022.

[48] D. Turpin et al., "Fast-grasp'd: Dexterous multi-finger grasp generation through differentiable simulation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[49] J. He et al., "Dexvlg: Dexterous vision-language-grasp model at scale," *arXiv preprint arXiv:2507.02747*, 2025.

[50] J. Ye et al., "Dex1b: Learning with 1b demonstrations for dexterous manipulation," in *Robotics: Science and Systems (RSS)*, 2025.

[51] O. Kroemer, S. Niekum, and G. Konidaris, "A Review of Robot Learning for Manipulation: Challenges, Representations, and Algorithms," in *Journal of Machine Learning Research (JMLR)*, 2021.

[52] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware," in *Robotics: Science and Systems (RSS)*, 2023.

[53] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low-Cost Whole-Body Teleoperation," in *Conference on Robot Learning (CoRL)*, 2024.

[54] H. Qi, A. Kumar, R. Calandra, Y. Ma, and J. Malik, "In-Hand Object Rotation via Rapid Motor Adaptation," in *Conference on Robot Learning (CoRL)*, 2022.

[55] J. Wang et al., "Lessons from Learning to Spin "Pens"," in *Conference on Robot Learning (CoRL)*, 2024.

[56] OpenAI et al., "Solving rubik's cube with a robot hand," *arXiv preprint arXiv:1910.07113*, 2019.

[57] H. Jiang, Y. Wang, H. Zhou, and D. Seita, "Learning to Singulate Objects in Packed Environments using a Dexterous Hand," in *International Symposium on Robotics Research (ISRR)*, 2024.

[58] L. Xu et al., "Dexsingrasp: Learning a unified policy for dexterous object singulation and grasping in cluttered environments," *arXiv preprint arXiv:2504.04516*, 2025.

[59] T. Chen, E. Cousineau, N. Kuppuswamy, and P. Agrawal, "Vegetable Peeling: A Case Study in Constrained Dexterous Manipulation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[60] T. Lin, Z.-H. Yin, H. Qi, P. Abbeel, and J. Malik, "Twisting Lids Off with Two Hands," in *Conference on Robot Learning (CoRL)*, 2024.

[61] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," in *Robotics: Science and Systems (RSS)*, 2024.

[62] T. G. W. Lum, O. Y. Lee, C. K. Liu, and J. Bohg, "Crossing the Human-Robot Embodiment Gap with Sim-to-Real RL using One Human Demonstration," in *Conference on Robot Learning (CoRL)*, 2025.

[63] J. Chen, Y. Chen, J. Zhang, and H. Wang, "Task-oriented dexterous hand pose synthesis using differentiable grasp wrench boundary estimator," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.

[64] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis," in *European Conference on Computer Vision (ECCV)*, 2020.

[65] S. Liu et al., "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[66] N. Ravi et al., "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[67] T. Ren et al., *Grounded sam: Assembling open-world models for diverse visual tasks*, 2024. arXiv: `2401.14159 [cs.CV]`.

[68] A. Kirillov et al., "Segment Anything," *arXiv preprint arXiv:2304.02643*, 2023.

[69] Q. Jiang, F. Li, Z. Zeng, T. Ren, S. Liu, and L. Zhang, *T-rex2: Towards generic object detection via text-visual prompt synergy*, 2024. arXiv: `2403.14610 [cs.CV]`.

VI. Additional Details of GD2P

A. Related Work

Nonprehensile Robot Manipulation. Classical nonprehensile manipulation includes pushing, sliding, rolling, and tilting, and has a long history in robotics [4], [5], [6], [7]. Planning methods for nonprehensile manipulation often assume access to object models or priors [37], [38]. Another recent planning-based method explores nonprehensile interaction with high-DOF hands in simulation by analyzing contact reasoning and wrench closure [39]. In contrast, our work targets real-world pushing and pulling using a high-DOF hand applied to diverse and geometrically complex objects. Applying traditional planning methods in this setting is highly challenging due to the complexity of modeling the diverse contact interactions between the hand and objects. Recent learning-based methods have extended nonprehensile manipulation beyond classical planning, including extrinsic dexterity systems [8], [40] and those based on predicting object dynamics such as HACMan [9], [10], CORN [11], and DyWA [12]. Other works approach pushing as a precursor to grasping, often in planar settings with parallel-jaw grippers for multi-object manipulation [41], or use bimanual systems for nonprehensile tasks using multi-link tools [42]. None of these works study learning for single-hand pushing and pulling with dexterous hands. Furthermore, many prior benchmarks focus on pushing single flat objects on a surface, such as a T-shape object [14], or use spatulas to move small cubes and granular media [13], [15]. Our work directly targets larger and more complex objects, including those that might topple or require coordinated multi-surface contact.

Dexterous Grasping Synthesis and Datasets. A substantial body of research focuses on generating and evaluating grasp poses for multi-fingered hands. Pioneering efforts such as [43] create a dataset of 6.9K grasps using the GraspIt! [44] software tool, while [45] synthesize human hand poses by using a conditional Variational Autoencoder [46]. More recent efforts significantly scale grasp generation with tools such as differentiable contact simulation [47], [48] or optimization over an energy function based on Differentiable Force Closure (DFC) [27]. Our work falls in the latter category, which has facilitated the generation of diverse grasping datasets such as DexGraspNet [19] with 1.32M grasps followed by DexGraspNet 2.0 [20] with 427M grasps, with recent work [49] further incorporating language into the dataset. These pipelines generate hand poses by optimization over an energy function, filter them using physics simulators, train generative diffusion models for grasp synthesis, and typically include some fine-tuning or evaluation modules [2], [18]. Most recently, Dex1B [50] scales to one billion demonstrations by integrating optimization-based control with generative modeling. While our pipeline also uses energy-based pose optimization and filtering, we focus on synthesizing hand poses for nonprehensile manipulation, specifically pushing and pulling, demonstrating the effectiveness of a tailored pipeline beyond grasping.

Learning-Based Dexterous Manipulation. Learning-based approaches for robotic grasping and manipulation have rapidly expanded in recent years [51]. While some recent work emphasizes fine-grained bimanual manipulation using parallel-jaw grippers [52], [53], our focus is on learning single-arm manipulation with high-DOF dexterous hands such as the LEAP [16], Allegro, and Shadow Hands. These hands have been applied to a variety of tasks, such as in-hand object rotation [54], [55], [56], object singulation [57], [58], multi-object manipulation [23], [26], [41], and bimanual systems [59], [60]. While showing the versatility of dexterous hardware, these works focus on largely prehensile interactions. Prior learning-based systems with high-DOF hands for nonprehensile behaviors demonstrate tasks such as rolling objects or picking up plates as examples of learning from 3D data [61] or human videos [62]. Recently, [63] synthesize task-oriented dexterous hand poses for certain nonprehensile tasks such as pulling drawers. However, none of these methods directly study pushing or pulling as their primary manipulation mode.

B. Experiments Results

1) Simulation Experiments and Results: We investigate the first question in simulation. We evaluate the quality of the hand pose generation pipeline using IsaacGym [24]. To quantify the effectiveness of our trained model and dataset, we report the number of successfully pushed objects as a function of training data size. We train our diffusion model on varying subsets of the full dataset (of 1.3M hand poses) and evaluate on 300 unseen objects from the test set. The objects we use for training and evaluation come from [18], consisting of diverse daily objects. For each test object, we sample 200 candidate hand poses. An object is considered "successful" if at least one feasible hand pose results in success. Table I reports results over 3 different seeds, which shows that our model generates feasible pushing poses more reliably with larger training sets, which validates large-scale supervision. The growth is not strictly linear, suggesting room for improvement via better model tuning or data strategies. Qualitatively, our generated hand poses are diverse across object geometries and exhibit pushing intent (see the Appendix on our website for more discussion). A common failure mode in the preliminary experiments is that some generated poses still collide with the object, which motivates the later inclusion of the collision term in Eq. 3 for motion planning to further avoid collisions before executing pushing and pulling motions in the real world.

Data Size	# of Objects
2%	41.67 $\pm$ 10.21
20%	102.67 $\pm$ 5.85
50%	110.33 $\pm$ 29.67
100%	169.33 $\pm$ 15.18

TABLE I: Number of objects with ≥ 1 feasible pushing hand pose out of 300.

2) Real-World Experiments Setup: We evaluate GD2P on a real robot to check if our pushing and pulling hand poses successfully transfer to reality. Our hardware setup consists of a Franka Panda arm equipped with a four-finger, 16-DOF

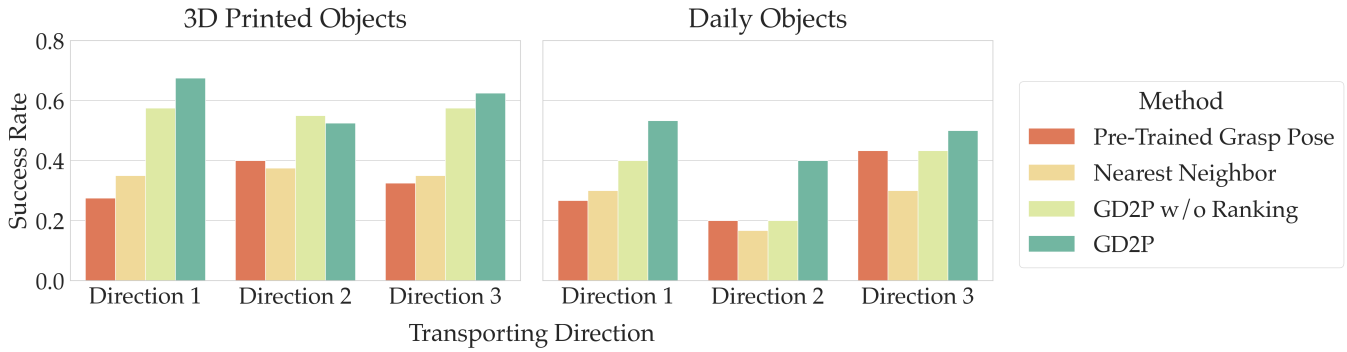


Fig. 6: Pushing and pulling success rates from GD2P and baselines, across different 3D printed (left) and daily objects (right), and with three directions evaluated. Each bar aggregates success rates from 40 trials (left bar plot) and 30 trials (right bar plot). See Sec. VI-B.2 and VI-B.3 for more details.

Allegro Hand (see the Appendix on our website for detailed setup visualization). It operates over a tabletop cutting board with dimensions 60 cm×60 cm. We use a mix of objects, including 3D-printed and common daily items (shown in Fig. 5). All evaluation objects are unseen during training. We obtain the object BPS representation as stated in III-B. While this pipeline introduces some noise, it is sufficient for GD2P to predict effective hand poses. In contrast, we empirically observed that optimization-based methods are more sensitive to mesh quality and often fail under these conditions.

Baselines and Ablations. We compare GD2P with:

- **Pre-Trained Grasp Pose**: We use a pre-trained grasp synthesis model from [18] using NeRF [64]. For each object, we train a NeRF representation, then query their pre-trained model for a grasp. This evaluates how well a grasping-centric model generalizes to nonprehensile tasks.
- **Nearest Neighbor (NN)**: Given a test object, we find the training object with the most similar BPS representation (in terms of Euclidean distance) and retrieve its associated hand poses. We then do the same motion planning pipeline as in GD2P. This tests out-of-distribution generalization with a retrieval-only approach compared to our proposed generative model.
- **GD2P w/o Ranking**: An ablation that excludes analytical ranking of hand poses (ignores Eq. 3) and executes a random feasible pose. This tests the usefulness of Eq. 3 in selecting poses.

Experiment Protocol and Evaluation. For each object, we test three pushing and pulling directions uniformly distributed around a circle. We consider "pulling" to involve applying force towards the body to move an object. This aligns with "Direction 2" in our real-world experiments. Along each direction, the robot executes the hand pose and planned motion five times, all with a fixed push length of 20 cm. A human manually places the object in a relatively consistent pose between trials. A trial is successful if the object's center is within 3 cm of the target position, the hand maintains contact throughout, and it does not lead to task failure modes such as toppling or loss of control. For NN and GD2P w/o Ranking, we randomly sample hand poses among the feasible planned actions. For Pre-trained Grasp Pose, we execute the best actions from its output. For our method, we execute the one with the highest score from Eq. 3.

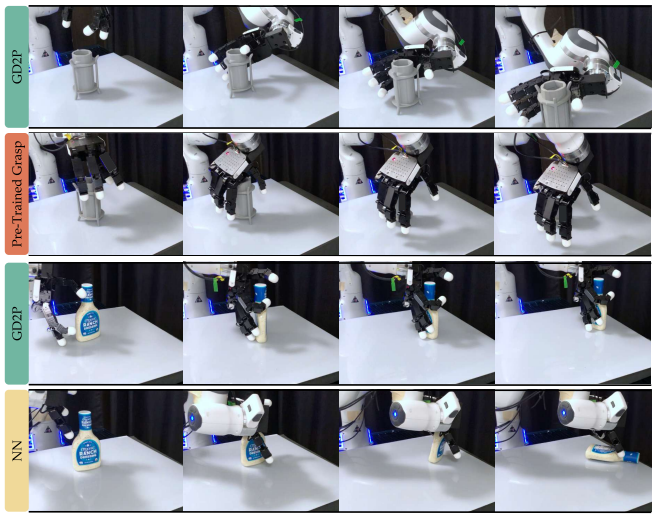


Fig. 7: Comparison between GD2P and baselines. The first two rows show GD2P (success) and Pre-Trained Grasp (failure) while pushing a 3D-printed vase forward (i.e., away from the robot). The last two rows show GD2P (success) and NN (failure) while pushing a ranch bottle to the right.

3) Is GD2P Effective for Diverse Objects and Directions?: We summarize quantitative results in Fig. 6, which shows that GD2P outperforms or matches alternative methods for both object categories. As shown in Fig. 7, the **Pre-Trained Grasp Pose** baseline suffers from two major issues. First, the hand pose is not conditioned on the pushing direction, which means during the push, the object is likely to slide off the hand due to limited support (Fig. 7, second row). Second, some objects are unsuitable for grasping with a single hand due to their geometry or awkward aspect ratios (e.g., a flat box with limited area for enclosure). Additionally, the similarity-based **Nearest Neighbor** baseline struggles due to limited granularity in object geometry matching, motivating the need for our geometry-conditioned generative model. For **GD2P w/o ranking**, we observe that its hand poses are more likely to collide with the table or objects. To further investigate this ablation, Fig. 4 shows three different hand

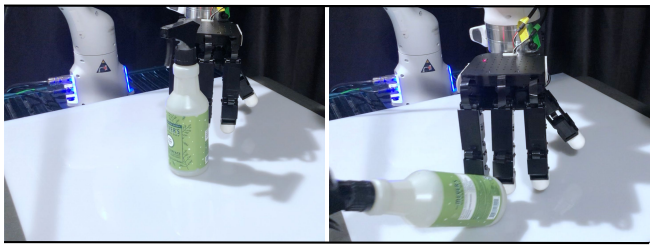


Fig. 8: Example of a typical failure case using the Fixed Hand Pose strategy (see Sec. VI-B.3 for details), which topples the spray.

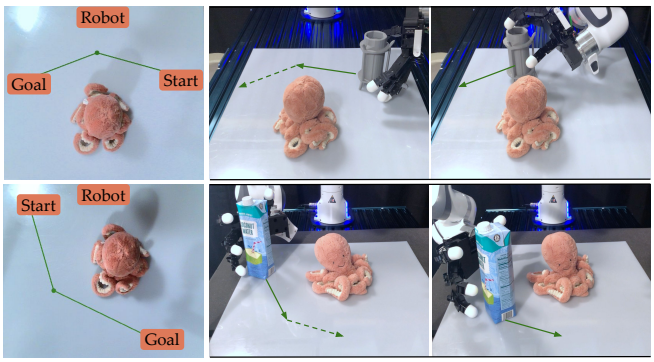


Fig. 9: Path planning using RRT* for multi-step planning. The first column shows the visualization of path planning results. The second and third columns show two consecutive hand poses for pushing the object along the path.

poses. The first one has a low collision score because it is easy to collide with the table, while the third collides with the objects and scores low on the palm direction. The second hand pose leads to a successful push in real-world experiments. This suggests the importance of our ranking system via Eq. 3. **GD2P** outperforms baselines in all directions tested in Fig. 6, demonstrating the robustness of its generated hand poses for pushing and pulling objects. Fig. 7 (first row) demonstrates using the palm and thumb to provide strong support moving the object forward, and the third row shows using the thumb and index finger to form a circular shape support for the thinner upper parts of the object while providing force at the bottom, aiding stable movement. For more rollouts, see the Appendix and the website.

Fixed Hand Pose: Inspired by prior pushing work [15], we manually define a "spatula" hand pose with the fingers spread flat (see Fig. 8) to assess whether simple flat-hand strategies suffice for diverse objects. We perform a case study on the 6 objects in Fig. 5 that are taller than 20 cm. We push each object 10 times and get a relatively low 18/60 success rate, suggesting insufficient object support.

4) Can GD2P Support Multi-step Manipulation Tasks?: **GD2P can serve as a reliable module for multi-step pushing.** Selecting a kinematically feasible hand pose for a given object state S_{obj} and direction u_{dir} is challenging in multi-step planning, as different waypoints may require different hand poses. Our method resolves this by identifying valid poses across object configurations and coupling pose selection with kinematic feasibility (see Sec. III-C). By doing so, GD2P can be used to perform multiple pushes. Fig. 9 shows a multi-step pushing sequence using GD2P. The robot uses two different hand poses to push the 3D-printed vase, as the first hand pose may not be ideal for the second hand pose, which shows the benefit of re-planning.

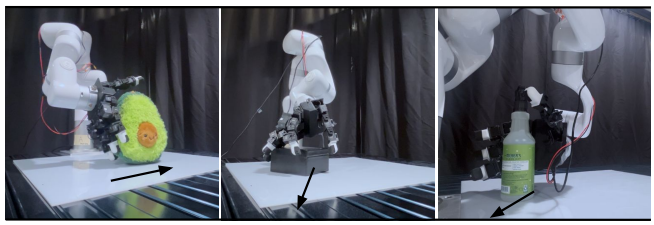


Fig. 10: Pushing and pulling experiments with a LEAP Hand, following the same protocol and evaluations from Section VI-B.2. We evaluates using the same 14 objects across three directions. The first two images illustrate pushing motions and the third illustrates a pulling motion of the spray bottle toward the robot base.

5) Does GD2P Apply to Different Hand Morphologies?: **GD2P achieves consistent performance across various robotic hand designs.** To evaluate the robustness of GD2P across different robotic hands, we applied the same pipeline on a four-finger, 16-DOF LEAP Hand [16] with an xArm7 arm and conducted experiments following the procedure described in Section VI-B.2 (see Fig. 10). Specifically, we tested on the same set of 14 objects over three directions in 210 new trials (not part of the 840 trials in Sec. VI-B.3) and achieved a comparable success rate of 68.1%. These results highlight the potential of GD2P for generating effective dexterous pushing and pulling policies across different robotic hands.

C. Dataset Generation and Statistical Analysis

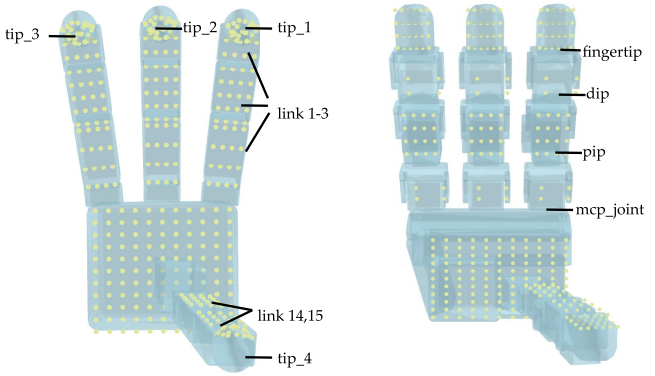


Fig. 11: Contact candidates on the Allegro Hand and LEAP Hand. Refer to Table II and Table III for the number of contacts on each link.

During dataset generation, we specify the contact candidates according to Figure 11 and Table II&III, and we set the weight parameters (from Eq. 1) according to values listed in Table IV. For the optimization we discussed in Sec. III-A, the detailed hyperparameters are in Table V.

In the original hand pose generation procedure, we mainly consider the object geometry and encourage contact between selected contact candidates all over the hand and the object surface. However, it is crucial to test pushing to validate the quality of the nonprehensile hand poses. Initially, we obtain a

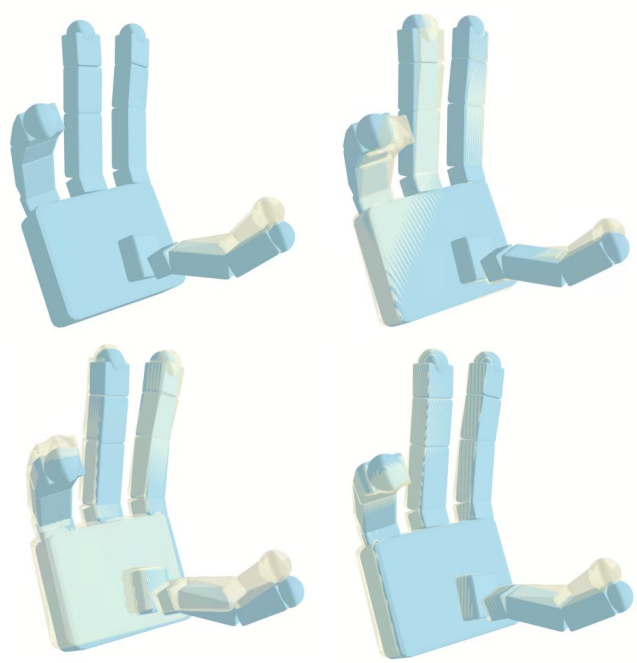


Fig. 12: A visualization of an example of augmentations. *Lightyellow* indicates the hand pose with the perturbation, and *lightblue* is the original one.

low success rate of all generated hand poses, so we augment each successful hand pose 10 times. These perturbations involve small changes in rotation (max 2.5 deg), translation (max 0.005 m) and joint pose (0.05 rad) using a Halton sequence. Figure 12 shows an example of a random original hand pose (lightblue color) and 4 different perturbed hand poses (lightyellow color). By doing so, we get a large dataset of only successful hand poses, which we use for training the diffusion model.

Figure 13 shows the distribution of joint angle values across our dataset. Most joints span the full range between their lower and upper bounds, and tend to have one or several modes. Those modes may lead to "general" stable hand poses for pushing motions. Other joint values may vary depending on particular object geometries. Figure 14 shows a breakdown of object categories and the frequency of the top 20 objects in our dataset.

Parameter	Value
w_{fc}	0.5
w_{dis}	500
w_{pen}	300.0
w_{spen}	100.0
w_{joints}	1.0
w_{ff}	3.0
w_{fp}	0.0
w_{tpen}	100.0
$w_{\text{direction}}$	200.0
$w_{\text{kinematics}}$	100.0

TABLE IV: Weight parameters.

Parameter	Value
Switch Possibility	0.5
μ	0.98
Step Size	0.005
Stepsize Period	50
Starting Temp.	18
Annealing Period	30
Temp. Decay	0.95

TABLE V: Optimization hyperparameters.

D. Training Details

We train our model with one NVIDIA 4090 GPU on a desktop. Detailed training and model parameters are shown in Table VI. We also show the training curves with training loss and validation loss on different scales of the dataset in Figure 15, which is relevant to our experiments in Sec. VI-B.1.

VII. Additional Details of Experiments

A. Experiment Details

Our physical experiment setup consists of a Franka Panda manipulator equipped with an Allegro Hand, as shown in Figure 16. We also place an L515 RealSense camera above the table, which is *only* used for path planning in multi-step planning experiments in Sec. VI-B.3 and Sec. VII-D. The surface we use for all experiments is a commercially available product purchased from Amazon (product_link). Since our focus is on nonprehensile hand pose generation, we assume that the surface's friction properties are sufficient to support pushing interactions. We leave a more detailed

Embodiment Part	Finger Tip	Finger Link	Palm
Link No.	tip_1, tip_2, tip_3, tip_4	1,2,3,5,6,7,9,10,11,14,15	palm_link
Number of Contact Candidates / each	96	16	128

TABLE II: Number of contact candidates on different parts of the Allegro Hand. We specify potential contacts all over the hand to encourage whole-hand (especially palm) nonprehensile manipulation on the object. See Table III for that of the LEAP Hand.

Embodiment Part	Finger Tip	mcp_joint & dip	dip	thumb_pip & thumb_dip	Palm
Link No.	tip_1, tip_2, tip_3, thumb_tip	1,2,3	1,2,3	–	palm_link
Number of Contact Candidates / each	24	16	4	16	128

TABLE III: Number of contact candidates on different parts of the LEAP hand. See Table II for the Allegro Hand.

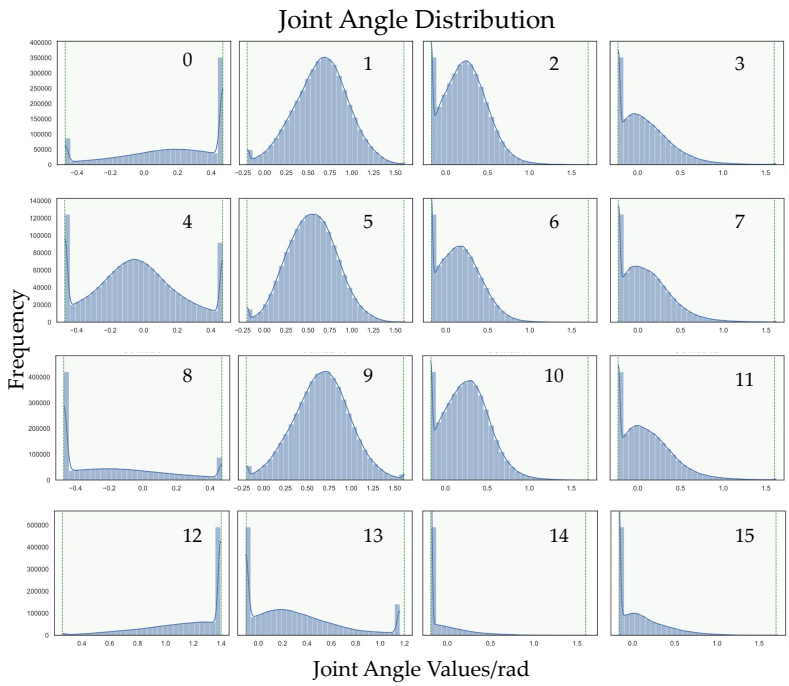


Fig. 13: Visualization of the distribution of joint angle values in our proposed dataset, demonstrating the diversity of our generated hand poses. The number on the top right corner of each subfigure indicates the joint index. The *green dashed lines* on the edge of x-axis indicate the lower/upper bounds of each joint angle values.

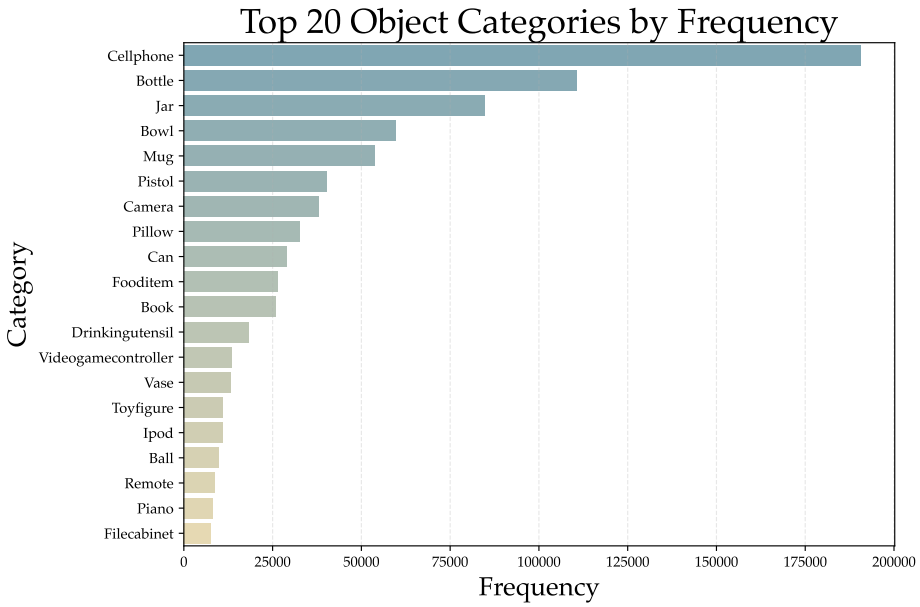


Fig. 14: Visualization of the top 20 objects in terms of pushing hand poses frequency in our proposed dataset.

investigation of how physical properties influence dexterous nonprehensile manipulation as future work.

We select 8 3D-printed objects and 6 real-world objects, covering flat, volumetric, and tall objects, as shown in Figure 17. Each object presents unique challenges for pushing. For example, when the robot hand approaches flat objects (e.g., Cake, Cookie Box) it may risk colliding with the table. In addition, tall objects (e.g., Lamp, Spray) frequently topple during pushing due to a high center of mass. While our method also suffers from these failure modes (particularly object toppling), it outperforms baselines, which topple objects more frequently. This motivates our case study on using a fixed hand pose to push objects taller than 20 cm. While fixed hand poses can reliably work for objects with simple geometries, they frequently fail on these taller objects. As discussed in Sec. VI-B.3, our results highlight the need

Component	Parameter	Default / value
Data Config	observation_dim	4096
	pushingpose_dim	25
Model Config	name	ConditionalUnet1D
	input_dim	25
	global_cond_dim	4096
DDPM Scheduler	beta_schedule	squaredcos_cap_v2
	clip_sample	True
	num_diffusion_timesteps	100
	prediction_type	epsilon
Training Config	batch_size	16
	n_epochs	200
	print_freq	10
	snapshot_freq	25
Optim Config	optimizer	Adam
	lr	1×10^{-4}
	weight_decay	1×10^{-6}
	beta1	0.9
	amsgrad	False
	eps	1×10^{-8}
	grad_clip	1.0
lr Scheduler	name	cosine
	num_warmup_steps	500
EMAModel	power	0.75

TABLE VI: Configuration and training hyperparameters of the diffusion model.

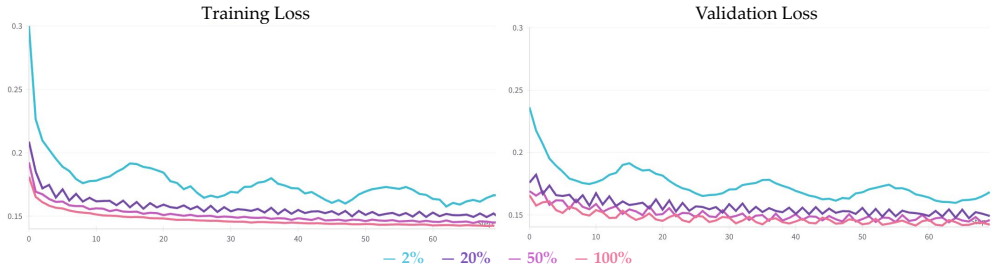


Fig. 15: Training curves on different scales of the dataset. See Sec.VI-B.1 for more discussion.

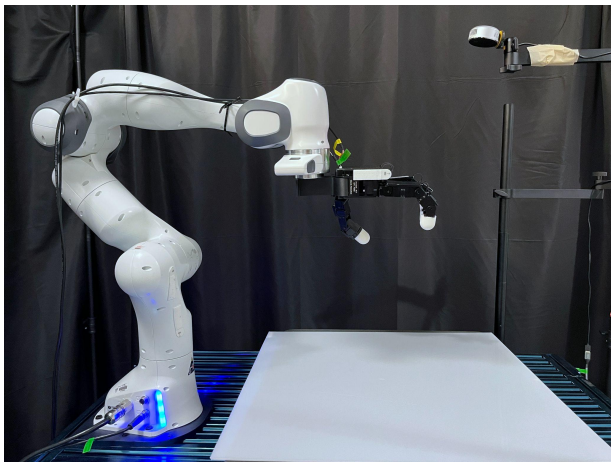


Fig. 16: Our physical experiment setup including a Frank Panda robot with an attached Allegro Hand. The camera is only used for high-level path planning.

for hand poses that provide more stable object support for transporting.

We list the number of successful trials out of 5 for each method and direction in Table VII. A blank entry (-) indicates that the robot could not execute the motion due to kinematic infeasibility. While GD2P has marginally more infeasible trials than the baselines, this is expected because GD2P generates diverse hand orientations beyond top-down poses. All methods execute pushes for 20 cm, which is relatively long within the robot's workspace, and this can be infeasible for many hand poses. In contrast, the Pre-Trained Grasp Pose baseline tends to result in consistently top-down hand poses, which are generally easier to execute due to reachability and kinematic constraints. Despite counting all kinematically infeasible trials as failures, GD2P outperforms the baseline methods, demonstrating its robustness on pushing or pulling tasks.

B. *More Successful Rollouts*

We provide additional example visualizations of successful rollouts of GD2P in Figure 18. For videos, please refer to

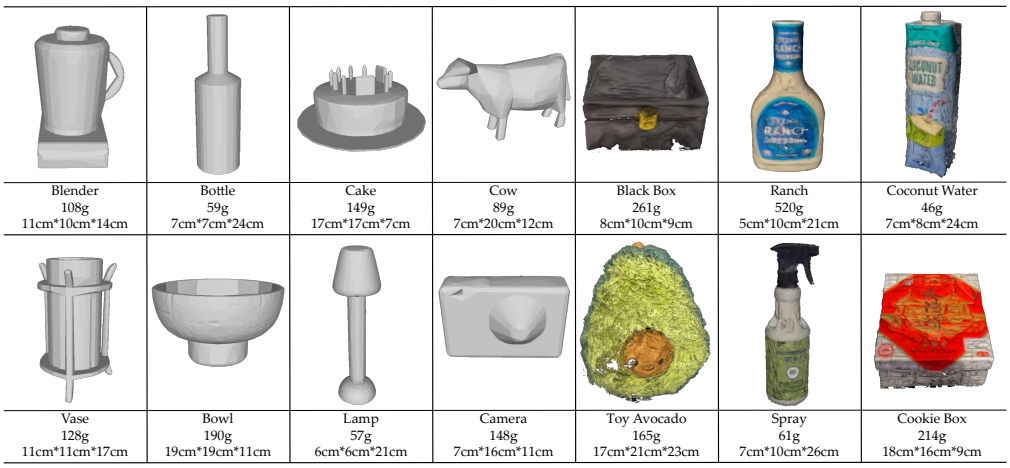


Fig. 17: 3D meshes, mass and physical dimensions of all objects tested in real-world experiments. Dimensions are listed as (x, y, z).

	GD2P			**GD2P w/o Ranking**			**Nearest Neighbor**			**Pre-Trained Grasp Pose**		
	Dir.1	Dir.2	Dir.3	Dir.1	Dir.2	Dir.3	Dir.1	Dir.2	Dir.3	Dir.1	Dir.2	Dir.3
Blender	5/5	4/5	4/5	3/5	3/5	5/5	2/5	2/5	2/5	1/5	1/5	1/5
Vase	5/5	3/5	4/5	2/5	4/5	4/5	4/5	4/5	3/5	2/5	3/5	2/5
Bottle	4/5	4/5	5/5	3/5	3/5	3/5	0/5	4/5	3/5	3/5	2/5	2/5
Bowl	4/5	1/5	-	4/5	1/5	-	2/5	2/5	1/5	3/5	2/5	2/5
Cake	4/5	3/5	4/5	4/5	4/5	3/5	3/5	1/5	1/5	1/5	0/5	1/5
Lamp	1/5	1/5	1/5	2/5	2/5	2/5	1/5	0/5	0/5	0/5	1/5	1/5
Cow	5/5	3/5	3/5	3/5	2/5	3/5	1/5	1/5	1/5	0/5	3/5	2/5
Camera	2/5	2/5	4/5	2/5	3/5	3/5	1/5	1/5	3/5	1/5	4/5	2/5
3D Avg./ %	67.5	52.5	62.5	57.5	55.0	57.5	35.0	37.5	35.0	27.5	40.0	32.5
Black Box	4/5	4/5	3/5	3/5	1/5	2/5	1/5	1/5	2/5	3/5	3/5	2/5
Toy Avocado	4/5	-	1/5	3/5	-	2/5	-	-	1/5	3/5	0/5	4/5
Ranch	3/5	2/5	3/5	4/5	1/5	2/5	3/5	1/5	4/5	1/5	-	2/5
Spray	3/5	-	1/5	0/5	-	1/5	2/5	-	2/5	0/5	0/5	2/5
Coconut Water	2/5	3/5	4/5	2/5	2/5	1/5	2/5	1/5	2/5	0/5	0/5	0/5
Cookie Box	-	5/5	3/5	-	2/5	5/5	-	3/5	2/5	2/5	2/5	1/5
DO Avg./ %	53.3	40.0	50.0	40.0	20.0	43.3	30.0	16.7	30.0	26.7	20.0	43.3
All Avg./ %	61.4	47.1	57.1	50.0	40.0	51.4	32.9	28.6	32.9	27.1	31.4	37.1

TABLE VII: Detailed experiment results for each object and direction combination. "3D Avg." refers to the average success rate over all 3D-printed objects, "DO Avg." is that of daily objects and "All Avg." is that of all 14 test objects. These results correspond to the bar charts in Figure 6.

our website: geodex2p.github.io.

C. Results and Analysis of Baseline Methods

We visualize 3 examples of the nearest neighbor (NN) retrieval results and the trained NeRF representation in Figure 19. The retrieved NN objects are similar in shape and scale of the query object (left 3 columns in Figure 19). However, their coarse geometry granularity is insufficient to generate robust hand poses. For example, with the Toy Avocado, our method selects a hand pose that pushes from the bottom to avoid sliding or toppling. In contrast, the NN method retrieves a vase-like object, where pushes from the middle make more sense. The irregular geometric shape at the bottom of the vase-like object could potentially cause more collisions and may increase the difficulty of solving the kinematics. The right 3 columns in Figure 19 visualize the NeRF input to the Pre-Trained Grasp Pose method, since we use their pre-trained model taking in NeRF representations. Though a common failure mode of the pre-trained grasp pose is that the object slips from the hand because the palm is oriented at an improper angle, we observe notable visual noise in the NeRF representation, which may also deteriorate performance of this baseline. For more discussions of baseline performance, see Sec. VI-B.3.

D. Multi-step Planning

Here, we provide more information and context on top of the *Multi-step Planning* section in Sec. VI-B.3. These experiments explore the potential for GD2P's hand poses to support long-horizon planning. As shown in Figure 9, an Intel RealSense L515 camera captures a top-down view of the scene (see Figure 9). A toy placed in the scene serves as an obstacle. We extract its segmentation mask using Grounded SAM 2 [65], [66], [67], [68], [69], define the toy's position at its (estimated) center, and set a fixed 20 cm radius for path planning. The start and goal positions are manually assigned. We use RRT* as a high-level planner to compute a collision-free path in the 2D image space. Through camera calibration, we convert the 2D waypoints into 3D coordinates in the robot frame. For each edge along the planned path,

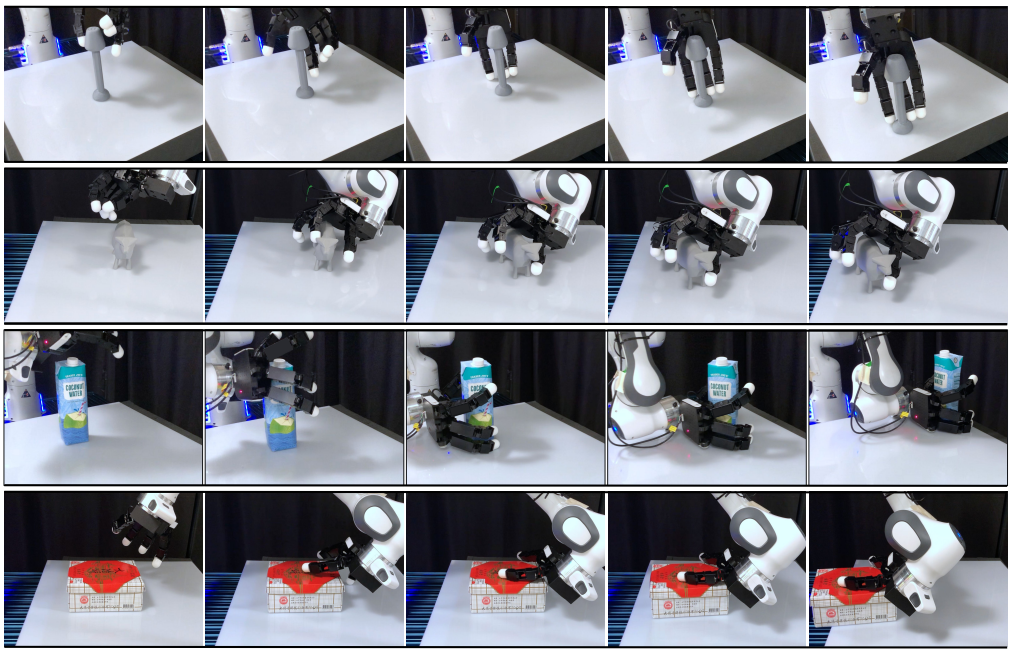
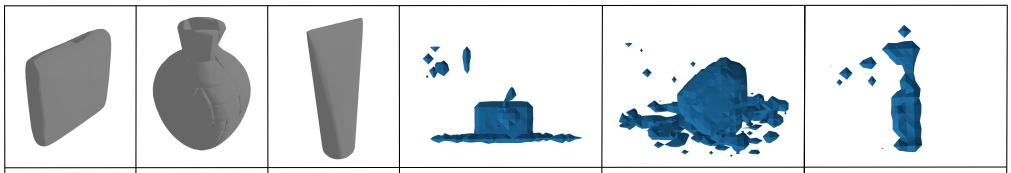


Fig. 18: Successful rollouts of GD2P, one per row.



NN of Camera 6cm*17cm*8cm	NN of Toy Avocado 21cm*21cm*23cm	NN of Spray 11cm*7cm*29cm	Trained NeRF of Camera	Trained NeRF of Toy Avocado	Trained NeRF of Spray

Fig. 19: Nearest Neighbor retrieval results of three test objects (left three columns). The retrieved objects share geometric similarities with the query object: a box-shaped match for the camera, a spherical-shaped vase for the toy avocado, and a tall, slender shape for the spray. The pre-trained grasp pose baseline takes in NeRF as input(right three columns). We trained the NeRF for our objects following their instructions and pipelines [18].

GD2P generates a corresponding hand pose, and the robot pushes the object towards the next waypoint.

We test with two episodes that cover more pushing directions. The key insight in these experiments is that hand poses should be considered and evaluated while considering the kinematics of the arm as the motion becomes more complex. In the second row of Figure 9, a similar hand pose is able to finish the two-step pushing tasks while avoiding the obstacle. However, the first row of Figure 9 shows the need to change hand poses to better fit the object pose and the intended pushing direction. This motivates our use of motion planning and pose ranking to facilitate stable and smooth multi-step pushing motions.