One-Shot Learning of Manipulation from RGB-D Videos via Object-Centric Interaction Reasoning

Ze Fu[1,3], Pinhao Song[1,3], Yutong Hu[1,3], Renaud Detry[1,2,3]

***Abstract*—Humans can learn manipulation behaviors by observing a short demonstration video, inferring how tools and objects should move relative to one another. Enabling robots to acquire manipulation skills from visual demonstrations with similar efficiency remains a core challenge. We propose an interaction-centric framework for one-shot learning of manipulation from RGB-D videos. Our method first identifies the task-relevant interaction segment in a demonstration video to reduce noise and focus on structured interactions. On these segments, it extracts relative trajectories between entities, capturing both hand–object interactions and object–object interactions, and predicts the relative motion using a model conditioned on object geometry and language instruction. By focusing on task-relevant interactions, the approach does not require robot embodiment data and generalizes across object poses and scene configurations. We validate our method on a physical robot, showing that a single demonstration suffices to robustly execute a variety of relational manipulation tasks.**

***Index Terms*—Learning from video, Interaction-centric policy, One-shot learning**

I. Introduction

Humans can learn manipulation behaviors by simply watching a demonstration video. From a short observation, one can infer how a tool should move relative to an object and reproduce the behavior with hands. Enabling robots to acquire manipulation skills from visual demonstrations in a similarly efficient manner remains a central challenge in robot learning. Most existing manipulation policies map visual observations directly to robot actions [1], [2] (Fig. 1 (a)). Such action-centric formulations typically rely on large numbers of robot demonstrations collected across different object poses, and require robot embodiment data that videos cannot provide. This limits their ability to leverage the vast amount of manipulation knowledge available in videos.

A key observation is that in many manipulation tasks, the task outcome is determined primarily by a short interaction phase during which the motion between entities is strongly constrained [3], [4]. For example, when pouring water from a teapot into a cup, success depends mainly on the relative motion between the two objects during the pouring, while the preceding approaching can vary substantially. This suggests that the essential information in a manipulation video may lie in the *interaction trajectory* between entities. Importantly, such interaction trajectories can be directly inferred from visual observations without access to robot action data. This motivates the following question: *can robots learn manipulation behaviors from videos by modeling interaction trajectories, rather than memorizing actions?*

(a) Action-Centric Policy (b) Interaction-Centric Policy

Fig. 1. Comparison between action-centric policies and our interaction-centric policy. (a) Action-centric policies learn a direct mapping from observations to end-effector motions and therefore require many demonstrations to cover the high-dimensional robot state space for a given task. (b) our method focuses exclusively on interaction segments within demonstrations and learns to predict the relative motion between task-relevant entities.

In this work, we propose an interaction-centric framework for **one-shot learning of manipulation from video** (Fig. 1 (b)). Instead of imitating full robot trajectories, our approach extracts and models interaction trajectories between entities, representing each interaction as the relative motion between two point clouds. This unified formulation captures both hand–object interactions (grasping) and object–object interactions (manipulation), removing the need for robot embodiment data while isolating task-relevant structure directly observable from videos. By operating on relative interaction trajectories rather than full action sequences, the resulting policy can be learned from a single demonstration and generalizes robustly across varying object poses and scene configurations.

Our contributions are threefold:

- We introduce an interaction-centric formulation for learning manipulation from video demonstrations, representing both grasping and manipulation as relative motion between entities.
- We propose an automatic pipeline that extracts interaction intervals and relative motion trajectories from RGB-D videos using hand pose and object pose estimation.
- We demonstrate that the resulting policy can learn manipulation behaviors from a single video demonstration and generalize across different scene configurations.

II. Related Work

Learning from Human Videos. Recent works have explored learning manipulation skills from human videos, which

[1]KU Leuven, Dept. Mechanical Engineering, Research unit Robotics, Automation and Mechatronics firstname.lastname@kuleuven.be

[2]KU Leuven, Dept. Electrical Engineering, Research unit Processing Speech and Images

[3]Flanders Make@KU Leuven

This work was supported by Interne Fondsen KU Leuven/Internal Funds KU Leuven (C2E/24/034).

can be broadly categorized into two lines of research. The first leverages large-scale in-the-wild videos to learn general-purpose representations [5] or reward functions [6]. While these approaches benefit from diverse data, the learned representations are often difficult to transfer to robotic manipulation due to domain variability and the embodiment gap. The second line introduces explicit priors to bridge this gap. A prominent direction focuses on learning object-centric motion representations, such as object flow [7]–[9]. However, flow-based representations are sensitive to camera motion and prone to drift. Other approaches attempt to reduce the embodiment gap by synthesizing robot demonstrations from human videos, for example by replacing human hands with robot grippers via inpainting [10]. In contrast, we adopt a simple pose-based representation and focus on interaction segments, improving robustness and learning efficiency.

One-shot Imitation Learning. Learning manipulation policies from a single demonstration is highly desirable due to the high cost of collecting robot data [11]. A common approach leverages transferable priors, such as cross-task perceptual features [12]–[15] or large-scale synthetic pretraining [16], [17], to enable adaptation under limited supervision. However, their performance often depends on the alignment between pretraining data and target tasks. Another direction improves generalization via data augmentation. For example, DemoGen [3] synthesizes trajectory variants to increase diversity, but its effectiveness is bounded by the augmentation strategy. In contrast, our method achieves one-shot generalization through structural invariance, enabling direct transfer of a single demonstration to novel object poses and scene configurations without additional data or augmentation.

III. Method

A. Problem Formulation

We consider the problem of learning manipulation behaviors from a single demonstration video. Given an RGB-D video $\mathcal{V}$ of a manipulation task, our goal is to extract task-relevant interaction information and learn a policy that can reproduce the observed behavior. From the video $\mathcal{V}$, we extract a set of samples of the form $(P_A, P_B, \tau_{AB}, \ell)$, where P_A and P_B denote the point clouds of two interacting entities, ℓ is the language instruction, and $\tau_{AB} = \{T_{AB}^{(k)}\}_{k=1}^{n}$ is a sequence of relative poses describing the motion of entity A with respect to entity B during an interaction interval. This formulation captures two types of interactions in a unified manner. For **hand–object interactions**, P_A corresponds to the gripper (or hand) and P_B corresponds to the manipulated object. For **object–object interactions**, P_A and P_B correspond to two task-relevant objects involved in the manipulation process.

Our objective is to learn a policy $f : (P_A, P_B, \ell) \rightarrow \tau_{AB}$. At execution time, the predicted interaction trajectory is instantiated in the world frame to generate executable robot motions. Given the estimated pose of the reference entity B in the world frame, $\hat{T}_B$, the desired trajectory of entity A is obtained by:

$$T_A^{(k)} = T_{AB}^{(k)} \hat{T}_B. \tag{1}$$

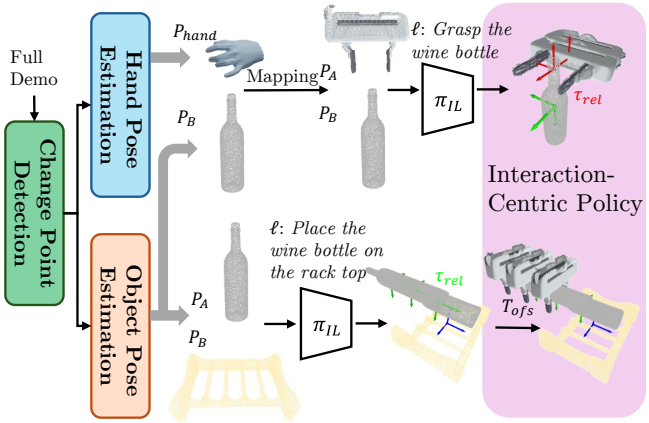

Fig. 2. Overview of our method. Given a raw demonstration, we first apply change-point detection to identify task-relevant interaction segments (e.g., pick or place) from t_s to t_g within the full horizon $[0, T]$. Next, hand and object pose estimation is performed to extract entity poses, which are then transformed into a relative reference frame. An interaction-centric model is trained to predict the corresponding relative interaction trajectory. The final end-effector trajectory is recovered by: (a) *grasping*, composing the predicted relative poses with the estimated target object pose; and (b) *manipulation*, composing the predicted interaction with the grasp offset.

B. Video Preprocessing

Given an RGB-D demonstration video $\mathcal{V}$, our goal is to extract geometric and motion cues required for interaction discovery and learning. We first estimate the hand pose trajectory from RGB-D observations using an off-the-shelf hand pose estimator [18]. Following [19], we approximate the gripper pose from the hand by aligning a parallel-jaw gripper model to key hand joints. Specifically, the gripper finger tips are aligned with the index and thumb tips, while its orientation is determined by the axis connecting these two points. This enbales the extraction of manipulation actions in the absence of robot states.

We then estimate the poses of task-relevant objects over time using [20], and extract their corresponding point clouds from depth observations. Based on the estimated poses, we compute the relative motion between interacting entities. Specifically, given the poses of two entities T_A and T_B, we express the motion of entity A in the coordinate frame of entity B:

$$T_{AB} = T_B^{-1} T_A. \tag{2}$$

This object-centric formulation provides a consistent representation of interactions across different scene layouts.

C. Interaction Segment Discovery

Learning from videos is challenging due to significant observation noise and variability. To improve learning efficiency, we focus on identifying the *interaction segment* from a demonstration video, where task-relevant dynamics are concentrated and more structured.

Based on the estimated hand poses and object poses over time, we construct a set of temporal signals that capture interaction dynamics, including (1) hand motion magnitude and

(2) spatial proximity between interacting entities. Intuitively, the transition from free motion to interaction is characterized by reduced motion variability and close spatial contact. We aggregate these signals into a multivariate time series:

$$S(t) = [v(t), \omega(t), d(t)], \tag{3}$$

where $v(t)$ and $\omega(t)$ denote the linear and angular velocities of the hand, and $d(t)$ denotes the distance between interacting entities.

To identify the interaction segment, we apply change-point detection on $S(t)$ to locate transitions in motion patterns. Specifically, we use the PELT algorithm [21] to detect a set of change points $\{\tau_i\}$. For each interaction episode, we define the interaction interval $[t_s, t_g]$ as the segment preceding the goal frame t_g (e.g., contact completion or grasp event):

$$t_s = \max\{\tau_i \mid \tau_i < t_g\}. \tag{4}$$

This procedure automatically extracts task-relevant interaction segments from raw videos without manual annotation. In practice, it yields two types of interactions: (i) *grasping*, characterized by hand–object relative motion, and (ii) *manipulation*, characterized by object–object relative motion. Both are represented in a unified form as T_{AB}, enabling a consistent learning framework.

D. Object-Centric Interaction Prediction

Given the processed interaction segment, our goal is to predict an object-centric interaction trajectory from visual observations. Specifically, we model the conditional mapping: $f : (P_A, P_B, \ell) \rightarrow \tau_{AB}$.

Geometric Encoding. We first encode each point cloud P_A and P_B into a set of feature tokens using a shared point cloud encoder. To improve robustness to partial observations, we adopt a masked encoding strategy [22], where a proportion ϕ of tokens is randomly dropped during training. This yields two sets of geometric tokens, F^A_{pcd} and F^B_{pcd}. To capture interaction-specific geometry, we apply cross-attention between the two token sets, allowing each entity to attend to the other's local structure.

Multi-modal Fusion. We fuse geometric features with the language instruction ℓ, which is encoded into a token F_ℓ using a frozen CLIP text encoder [23]. All tokens are concatenated into a sequence $F = [F^A_{pcd}, F^B_{pcd}, F_\ell]$, and processed by a Transformer encoder to produce a fused representation H.

Trajectory Prediction. Conditioned on H, a trajectory decoder directly predicts a sequence of n relative poses $\tau_{AB} = \{T^{(k)}_{AB}\}^n_{k=1}$, representing the interaction in an object-centric coordinate frame. The model is trained to regress the ground-truth interaction trajectory extracted from the interaction segment. The loss is defined as:

$$\mathcal{L} = \sum_{k=1}^{n} \left(\|\tilde{\mathbf{t}}_k - \mathbf{t}^{gt}_k\| + \lambda \|\tilde{\mathbf{r}}_k - \mathbf{r}^{gt}_k\| \right), \tag{5}$$

where $\mathbf{t}$ and $\mathbf{r}$ denote the translation and rotation components of the relative pose.

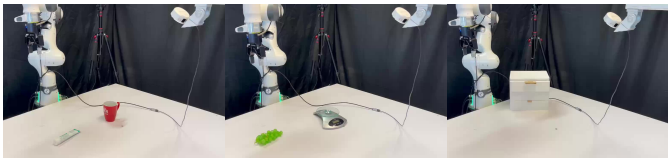

Fig. 3. Evaluation environment and manipulation tasks. From left to right: *Insert-Tube*, *Scale-Grape*, and *Open-Drawer*.

By predicting object-centric interaction trajectories, the model focuses on the functional interaction between entities, while remaining invariant to task-irrelevant variations such as viewpoint and transportation motion.

IV. Experiments

We validate our approach on a real robot platform. A single demonstration is provided for each task to train a multi-task policy, evaluating its real-world generalization under minimal supervision.

Settings. Experiments are conducted on a Franka Emika Panda robot equipped with a Robotiq parallel-jaw gripper. A RealSense L515 RGB-D camera provides a front-facing view of a $75\,\mathrm{cm} \times 50\,\mathrm{cm}$ tabletop workspace. Demonstrations are given as RGB-D videos, from which hand and object poses are estimated using the pipeline described in Section III-B. Task-relevant objects are segmented using GroundedSAM [24] and tracked with XMem++ [25]. During execution, predicted interaction trajectories are converted into robot motions via the cuRobo planner [26].

Tasks. We evaluate on three manipulation tasks involving different types of interactions (Fig. 3): ***Scale-Grape*:** grasping a plastic grape and placing it onto a scale. ***Insert-Tube*:** inserting a toothpaste tube into a mug with precise alignment. ***Open-Drawer*:** opening a hinged drawer. These tasks cover both hand–object (grasping) and object–object (manipulation) interactions, as well as varying levels of geometric precision.

Baseline and Evaluation To distinguish our method from naive trajectory replay, we implement a simple baseline, termed *Replay Policy*. During grasping, the baseline uses an off-the-shelf grasping planner [27]. During post-grasp manipulation, it applies Generalized ICP [28] to align the demonstrated object point clouds (P_A and P_B) with the current scene, and transfers the recorded gripper poses accordingly to compute robot actions. This baseline represents a straightforward replication of the motions observed in the demonstration video. Both methods are provided with a single demonstration per task. Each task is evaluated over 15 trials.

Results. We report task success rates in Table I. Despite being trained from a single demonstration, our method achieves consistent performance across tasks and remains robust under large configuration changes. The performance gap between our method and the baseline is particularly pronounced in *Open-Drawer* (47% vs. 13%), where the baseline often fails due to unreliable grasp pose prediction on the drawer handle and inaccurate ICP registration under partial observations.

TABLE I
QUANTITATIVE EVALUATION OF REAL-WORLD TASK SUCCESS GIVEN A SINGLE RGB-D DEMONSTRATION.

Method	scale_grape	insert_tube	open_drawer
Ours	Pick: 12/15 Place: 10/12 **Task: 10/15 (67%)**	Pick: 10/15 Place: 7/10 **Task: 7/15 (47%)**	Pick: 9/15 Place: 7/9 **Task: 7/15 (47%)**
Replay Policy	Pick: 8/15 Place: 5/8 Task: 5/15 (33%)	Pick: 6/15 Place: 3/6 Task: 3/15 (20%)	Pick: 2/15 Place: 2/2 Task: 2/15 (13%)

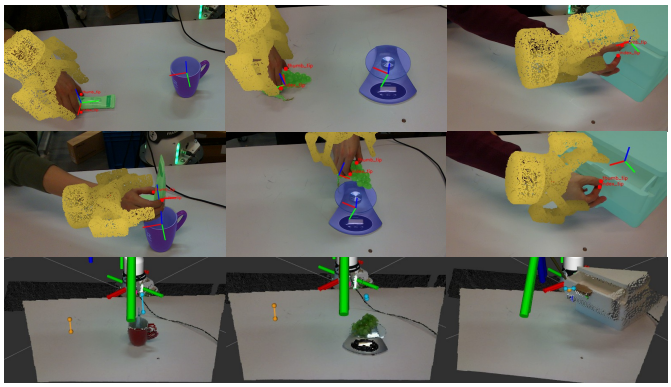

Fig. 4. Visualization of experimental results. Top two rows: processing of an RGB-D demonstration, including object segmentation, pose estimation, and hand pose extraction. The thumb and index fingertips are used to align and replace the gripper point cloud. Bottom row: predicted gripper trajectories during execution, where yellow indicates grasping and cyan indicates object–object manipulation.

Notably, the model performs reliably on precision-sensitive tasks such as *Insert-Tube*, suggesting that the learned object-centric interaction representation effectively captures the geometric constraints required for alignment. Across all tasks, we observe that most failures occur during the grasping stage, while successful grasps typically lead to correct downstream manipulation. This indicates that perception quality, particularly hand pose estimation, remains a key bottleneck in the overall pipeline.

We also visualize the extracted interaction segments and the predicted gripper trajectories during execution in Fig. 4. Despite noise in pose estimation, the proposed interaction segment extraction and discrete trajectory prediction help filter out high-frequency errors, leading to stable manipulation behaviors.

V. CONCLUSION

In this work, we presented an interaction-centric framework for one-shot manipulation from RGB-D videos. By focusing on task-relevant interaction segments and modeling object-centric relative motion between entities, our approach removes the need for robot embodiment data and enables generalization across object poses and scene configurations. Real-world experiments show that a single demonstration is sufficient to acquire and robustly execute diverse manipulation skills.

Limitations. Despite these promising results, several limitations remain. First, the performance of the proposed pipeline is constrained by the accuracy of pose estimation. Errors in hand pose or object pose estimation can propagate to interaction trajectory extraction and policy prediction. We expect this limitation to be alleviated with ongoing advances in more accurate hand pose tracking and zero-shot object pose estimation methods. Second, our current formulation represents object interactions purely in terms of relative poses, which is most suitable for rigid or semi-rigid objects. This representation is less expressive for deformable objects. Future work will explore more flexible object-centric representations, such as motion flows [29], to better capture a wider range of interaction dynamics. Overall, this work highlights the effectiveness of focusing on object-centric interactions for efficient manipulation learning from videos, and suggests a promising direction toward scalable, data-efficient robot learning from visual demonstrations.

REFERENCES

[1] A. Goyal, J. Xu, Y. Guo, V. Blukis, Y.-W. Chao, and D. Fox, "Rvt: Robotic view transformer for 3d object manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 694–710.

[2] T.-W. Ke, N. Gkanatsios, and K. Fragkiadaki, "3d diffuser actor: Policy diffusion with 3d scene representations," *arXiv preprint arXiv:2402.10885*, 2024.

[3] Z. Xue, S. Deng, Z. Chen, Y. Wang, Z. Yuan, and H. Xu, "Demogen: Synthetic demonstration generation for data-efficient visuomotor policy learning," *arXiv preprint arXiv:2502.16932*, 2025.

[4] Y. Yang, S. Cheng, Y. Fang, H. Bharadhwaj, M. Ding, G. Bertasius, and D. Szafir, "Lilo-vla: Compositional long-horizon manipulation via linked object-centric policies," *arXiv preprint arXiv:2602.21531*, 2026.

[5] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3m: A universal visual representation for robot manipulation," *arXiv preprint arXiv:2203.12601*, 2022.

[6] A. S. Chen, S. Nair, and C. Finn, "Learning generalizable robotic reward functions from" in-the-wild" human videos," *arXiv preprint arXiv:2103.16817*, 2021.

[7] C. Wen, X. Lin, J. So, K. Chen, Q. Dou, Y. Gao, and P. Abbeel, "Any-point trajectory modeling for policy learning," *arXiv preprint arXiv:2401.00025*, 2023.

[8] H. Bharadhwaj, R. Mottaghi, A. Gupta, and S. Tulsiani, "Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation," in *European Conference on Computer Vision*. Springer, 2024, pp. 306–324.

[9] J. Ren, P. Sundaresan, D. Sadigh, S. Choudhury, and J. Bohg, "Motion tracks: A unified representation for human-robot transfer in few-shot imitation learning," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 8802–8810.

[10] M. Lepert, J. Fang, and J. Bohg, "Phantom: Training robots without robots using only human videos," *arXiv preprint arXiv:2503.00779*, 2025.

[11] K. Dreczkowski, P. Vitiello, V. Vosylius, and E. Johns, "Learning a thousand tasks in a day," *Science Robotics*, vol. 10, no. 108, p. eadv7594, 2025. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.adv7594

[12] C. Pan, B. Okorn, H. Zhang, B. Eisner, and D. Held, "Tax-pose: Task-specific cross-pose estimation for robot manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 1783–1792.

[13] F. Qin, T. Hou, S. Lin, K. Wang, M. C. Yip, and S. Yu, "Anyokp: One-shot and instance-aware object keypoint extraction with pretrained vit," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 12 397–12 403.

[14] H. Ryu, J. Kim, H. An, J. Chang, J. Seo, T. Kim, Y. Kim, C. Hwang, J. Choi, and R. Horowitz, "Diffusion-edfs: Bi-equivariant denoising generative modeling on se (3) for visual robotic manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 007–18 018.

[15] C. Tang, A. Xiao, Y. Deng, T. Hu, W. Dong, H. Zhang, D. Hsu, and H. Zhang, "Functo: Function-centric one-shot imitation learning for tool manipulation," *arXiv preprint arXiv:2502.11744*, 2025.
[16] V. Vosylius and E. Johns, "Few-shot in-context imitation learning via implicit graph alignment," *arXiv preprint arXiv:2310.12238*, 2023.
[17] ——, "Instant policy: In-context imitation learning via graph diffusion," *arXiv preprint arXiv:2411.12633*, 2024.
[18] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "Mediapipe hands: On-device real-time hand tracking," *arXiv preprint arXiv:2006.10214*, 2020.
[19] G. Papagiannis, N. Di Palo, P. Vitiello, and E. Johns, "R+ x: Retrieval and execution from everyday human videos," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 8284–8290.
[20] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "Foundationpose: Unified 6d pose estimation and tracking of novel objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 868–17 879.
[21] C. Truong, L. Oudre, and N. Vayatis, "Selective review of offline change point detection methods," *Signal Processing*, vol. 167, p. 107299, 2020.
[22] Y. Pang, W. Wang, F. E. Tay, W. Liu, Y. Tian, and L. Yuan, "Masked autoencoders for point cloud self-supervised learning," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*. Springer, 2022, pp. 604–621.
[23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.
[24] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan *et al.*, "Grounded sam: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, 2024.
[25] M. Bekuzarov, A. Bermudez, J.-Y. Lee, and H. Li, "Xmem++: Production-level video segmentation from few annotated frames," 2023.
[26] B. Sundaralingam, S. K. S. Hari, A. Fishman, C. Garrett, K. V. Wyk, V. Blukis, A. Millane, H. Oleynikova, A. Handa, F. Ramos, N. Ratliff, and D. Fox, "curobo: Parallelized collision-free minimum-jerk robot motion generation," 2023. [Online]. Available: https://arxiv.org/abs/2310.17274
[27] H.-S. Fang, M. Gou, C. Wang, and C. Lu, "Robust grasping across diverse sensor qualities: The graspnet-1billion dataset," *The International Journal of Robotics Research*, 2023.
[28] A. Segal, D. Haehnel, and S. Thrun, "Generalized-icp," in *Proceedings of Robotics: Science and Systems*, Seattle, USA, June 2009.
[29] M. Xu, Z. Xu, Y. Xu, C. Chi, G. Wetzstein, M. Veloso, and S. Song, "Flow as the cross-domain manipulation interface," *arXiv preprint arXiv:2407.15208*, 2024.