

Robot Learning with Visual Predicted Force

Haonan Chen^{1,*}, Feiyang Wu^{1,*}, Yuxiang Ma², Mustafa Mete¹, Pengfei Ye²,
Junxuan Shen², Cheng Zhu³, Aurora Ruggeri¹, Kelvin Cheung¹,
Jiayuan Mao³, Edward Adelson², Jiajun Wu⁴, Robert D. Howe¹, Yilun Du¹

Abstract—Force-aware manipulation typically relies on specialized force or tactile sensors. We show that force-aware manipulation can instead be achieved through visual force prediction from the deformation of a compliant Fin Ray gripper. Our approach trains two models. First, we train a visual force estimator on calibration data and use it to annotate task demonstrations with force estimates. Second, we train an action–force proposal policy on these force-augmented demonstrations to jointly generate candidate robot actions and their associated forces. At test time, we sample candidate actions and the forces they are expected to produce, then execute the action whose predicted force is closest to a target from the demonstrations. We evaluate our approach on berry picking, empty-can grasping, in-hand reorientation, and plug insertion. Our results show that visual force prediction can guide inference-time action selection for contact-rich manipulation without requiring force or tactile sensors at deployment.

I. INTRODUCTION

Controlling interaction forces is important for contact-rich manipulation, where excessive or insufficient force can damage objects or cause task failure. This is particularly important when grasping soft objects (e.g., fruits, soft packages), handling deformable objects, or performing precise insertion [1], [2]. Existing robotic systems obtain such feedback using wrist-mounted force/torque sensors [3], flexible printed circuit board (FPCB)-based tactile sensors [4], [5], or optical tactile sensors such as GelSight Mini [6]. However, these sensing solutions introduce hardware and integration overhead and can require changes to the manipulator’s form factor and mechanical structure. Commercial six-axis force/torque sensors from manufacturers such as ATI and Bota can cost several thousand dollars per unit, excluding robot-specific mounting hardware and system integration [7], [8]. FPCB-based tactile sensors often require custom fabrication, wiring, and calibration, while optical tactile sensors require embedded cameras, illumination, and compliant sensing surfaces that may wear under repeated contact. These limitations motivate a natural question: Can a robot estimate force directly from visual observations and use these estimates to guide contact-rich manipulation?

In this work, we investigate learning force-aware manipulation using visual observations of a compliant Fin Ray gripper. As illustrated in Fig. 1, contact forces produce visible deformation in the gripper, providing visual cues for

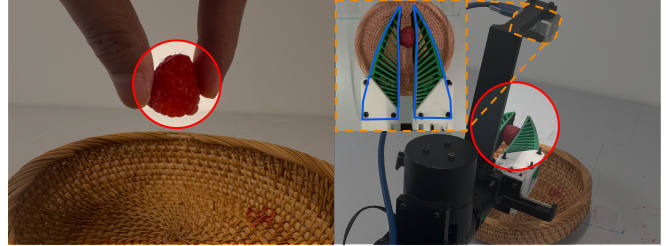


Fig. 1. **Force-aware manipulation from visual gripper deformation.** Humans naturally regulate force when handling fragile objects (left). We use wrist-camera observations of a compliant gripper (right, inset) to infer interaction force from visible deformation and enable gentle robotic manipulation without dedicated force or tactile sensing at deployment.

force estimation. Prior studies have explored this relationship through simulation-based force prediction and experimental analysis of Fin Ray deformation [9], [10]. The gripper is inexpensive and can be fabricated by 3D printing with TPU. We train a visual force estimator to infer force from wrist-camera images of the gripper during contact.

To use visual force estimates for manipulation, we first train a visual force estimator on calibration data and use it to label task demonstrations. We then train an action–force proposal policy on these demonstrations to jointly generate candidate actions and the associated forces values from the demonstration distribution. At test time, the policy jointly samples multiple action–force candidates. We execute the action associated with the force value closest to a task-specific target derived from the demonstration force estimates. We evaluate the method on berry picking, empty-can grasping, in-hand reorientation, and plug insertion. Our results show that visual force estimates can guide action selection for contact-rich manipulation without requiring force or tactile sensors at deployment.

Our main contributions are as follows:

- We introduce a visual force estimation method that infers force from the visible deformation of a compliant Fin Ray gripper.
- We develop an action–force proposal policy and an inference-time selection procedure that uses target forces from demonstrations to guide action selection.
- We demonstrate the effectiveness of the approach on four contact-rich manipulation tasks (berry picking, empty-can grasping, in-hand reorientation, and plug insertion) without requiring force or tactile sensors at deployment.

*Equal contribution.

¹Harvard University, Cambridge, MA, USA.

²Massachusetts Institute of Technology, Cambridge, MA, USA.

³University of Pennsylvania, Philadelphia, PA, USA.

⁴Stanford University, Stanford, CA, USA.

II. RELATED WORKS

A. Hardware-Based Force and Tactile Sensing

Dedicated force and tactile sensors obtain contact information through several sensing mechanisms. Force-torque sensors provide accurate measurements but require mounting, wiring, and calibration, while piezoelectric tactile sensors capture high-frequency stimuli [11]. Optical tactile sensors instead image a compliant sensing surface [12]. GelSight recovers contact geometry and force-related signals [13], [14], with subsequent designs exploring compact integration and biomimetic structures [15], [16], [17]. Related approaches reconstruct contact geometry and estimate forces from tactile images [18], [19], [20], while GelSight Fin Ray variants integrate optical tactile sensing into compliant fingers [21], [22]. Another direction observes the gripper structure itself. De Barrie et al. [9] use finite-element simulation data to train a model that predicts contact forces and stress distributions from finger deformation images. Kitamura et al. [10] relate camera-observed finger deformation to contact forces measured by 16 sensors embedded in the grasped object. Stuhne et al. [23] introduce SofIA, which uses a depth camera to observe the state of its compliant fingers. Forces for Free estimates contact force from visual observations of a tendon-driven underactuated hand optimized for force observability [24]. Building on visual observation of compliant structures, our work uses estimated forces to label demonstrations and train a policy whose generated force values guide action selection.

B. Force and Tactile Feedback for Robotic Manipulation

Force and tactile information support both contact perception and manipulation control. For perception, prior work uses touch to assess and improve grasps [25], [1], [3], estimate object pose from contact [26], and track object shape and pose during manipulation [27]. These contact cues also support grasp adjustment, sliding, pivoting, and rolling [28], [29], [30]. Policies combining vision and touch extend these capabilities to learned manipulation and in-hand rotation [31], [2], with further work incorporating audio [32] or adapting vision-only policies to force/torque sensing [33]. For control, tactile feedback supports cable following and needle threading [34], [35], while tactile-informed dynamics support predictive control for packing [36]. Hierarchical and reactive approaches integrate contact feedback into broader manipulation systems [37], [38], [39]. Other methods explicitly learn force or compliance control: RETAF separates tactile grasp-force control from arm pose prediction [40], Adaptive Compliance Policy learns variable compliance from demonstrations [41], and UMI-FT uses force-labeled demonstrations collected with a handheld interface to learn motion, grasp force, and stiffness [42]. Our method uses visual force estimates as training labels and policy-generated forces for inference-time action selection, without dedicated contact-sensing hardware at deployment.

C. Test-Time Searching

Prior work has explored inference-time search over sampled robot actions using learned interaction models [43] and diffusion-based planning [44], [45], [46], [47]. Frozen generative robot policies can be improved at test time by ranking action samples with auxiliary objectives. Value-Guided Policy Steering (V-GPS) [48] ranks policy samples using an offline-learned value function. Bidirectional Decoding [49] selects action chunks using consistency with previous decisions and contrast among future plans. Generative Predictive Control (GPC) [50] ranks and refines policy samples using a learned world model. DynaGuide [51] guides generative policies with learned dynamics, VLS [52] uses VLM-based trajectory rewards, and TouchGuide [53] steers pretrained generative visuomotor policies [54], [55] by using a tactile-informed Contact Physical Model to refine early visual-policy samples toward physically feasible contact. Our method uses visual force estimates to label demonstrations for policy training. At test time, the policy jointly generates candidate actions and force values. We select the action whose policy-generated force is closest to the target. The visual force estimator does not score candidates at test time.

III. METHOD

Our method consists of three stages. First, we collect paired images of gripper deformation and reference force measurements to train a visual force estimator. Second, we collect task demonstrations and use the trained estimator to annotate them with forces and select a target force for each task. Third, we train an action–force proposal policy on the annotated demonstrations. At test time, we sample candidate actions and the forces they are expected to produce, then execute the action whose predicted force is closest to the task’s target. We first formulate the problem, then describe each stage. Fig. 2 illustrates visual force estimation and inference-time action selection.

A. Problem Formulation

We consider contact-rich manipulation using a robot equipped with a compliant Fin Ray gripper, a wrist-mounted RGB camera, and an external side-view RGB camera. At time t , the robot observes

$$\mathbf{o}_t = \left(\mathbf{I}_t^{\text{wrist}}, \mathbf{I}_t^{\text{side}}, \mathbf{s}_t \right),$$

where $\mathbf{I}_t^{\text{wrist}}$ captures the visible deformation of the gripper, $\mathbf{I}_t^{\text{side}}$ captures the global robot–object interaction, and $\mathbf{s}_t$ denotes the robot proprioceptive state, including the end-effector pose and gripper state.

Let $\mathbf{x}_t^{\text{eef}}$ denote an end-effector action consisting of the Cartesian position, orientation quaternion, and gripper command, and let $f_t = |F_{z,t}|$ denote the magnitude of the axial interaction force. Given a task-specified target force f_{target} , our goal is to select a task-relevant end-effector action whose associated interaction force is close to the target.

The key challenge is that the interaction force is not directly observed during deployment: the robot has no

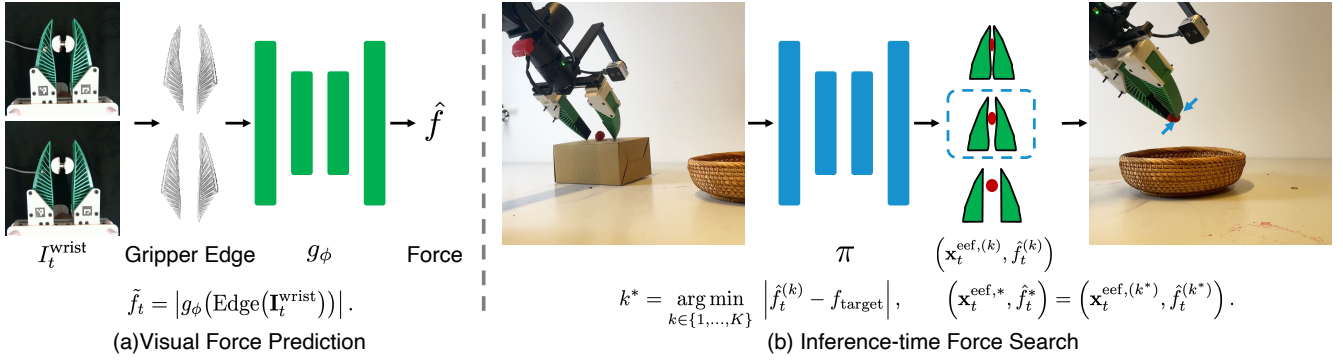


Fig. 2. **Overview of visual force estimation and inference-time force search.** (a) Visual force estimation converts a wrist-camera image into a masked Sobel edge representation and uses g_ϕ to estimate the axial interaction force from the visible deformation of the compliant Fin Ray gripper. (b) During inference-time force search, the policy jointly samples candidate end-effector actions and force values. We execute the action from the sample whose generated force value most closely matches the target force.

force/torque or tactile sensor attached to the manipulation system. It must therefore recover force information from the visible deformation of the compliant gripper and use this information to select an appropriate action.

B. Force Calibration and Visual Estimator Training

We estimate the axial interaction force from the visible deformation of the compliant Fin Ray gripper. The wrist-mounted camera provides a close-up view of the gripper fingers, whose internal stripes and overall geometry change under external loading. Our visual force estimator uses only this visual deformation representation as input.

Visual Force Data Collection. Training the visual force estimator requires paired observations of gripper deformation and ground-truth force measurements. To collect these pairs, we mount 19 3D-printed contact shells (Fig. 3(b)) on a reference force sensor. During contact, the fixtures transmit the force applied by the gripper to the reference sensor, while the wrist-mounted camera records the corresponding deformation of the Fin Ray fingers. We collect data using both a Mitsumi MMS101 resistive force sensor and an ATI six-axis force/torque (F/T) sensor as reference sensors.

We press the gripper against the fixtures at diverse contact locations and with varying force magnitudes, ranging from gentle contact to firm loading. We collect 117 episodes totaling 47.5 minutes. Images are recorded at 30 Hz and force measurements at approximately 20 Hz. We align the streams by timestamp and use only F_z to train the force estimator.

Model Training and Inference. Given a wrist-camera image $\mathbf{I}_t^{\text{wrist}}$, we first use SAM2 [56] to segment the Fin Ray gripper. We remove the background using the segmentation mask and resize the masked image to 256×256 . We then convert the masked image to grayscale and compute its Sobel edge magnitude. To suppress artificial edges introduced by the segmentation boundary, we erode the gripper mask by one pixel and apply the eroded mask to the edge image. The resulting representation highlights the internal stripes and geometry of the Fin Ray fingers while reducing task-irrelevant variation from the object and background.

The resulting single-channel edge image is processed by an encoder-only U-Net [57] with feature dimensions

(32, 64, 128, 256) and a 512-channel bottleneck. The network is trained by minimizing the mean-squared error between the predicted signed axial force and the timestamp-aligned F_z measurement.

For downstream manipulation, we use the magnitude of the predicted axial force. The complete visual force estimation pipeline is

$$\tilde{f}_t = \left| g_\phi \left(\text{Edge} \left(\mathbf{I}_t^{\text{wrist}} \right) \right) \right|, \quad (1)$$

where $\text{Edge}(\cdot)$ denotes the segmentation and edge-extraction pipeline described above, and g_ϕ denotes the learned visual force estimator. Thus, $\tilde{f}_t$ is inferred solely from the visual deformation of the gripper and requires no F/T sensor or robot proprioceptive state at deployment time.

C. Task Demonstration Collection and Force Annotation

For each manipulation task, we collect demonstrations using the robot arm, recording camera observations, robot states, and demonstrated end-effector actions. We freeze the visual force estimator g_ϕ trained in the first stage and use Eq. (1) to annotate each demonstration. At timestep t , the demonstrated end-effector action $\mathbf{x}_t^{\text{eef}}$ is paired with the corresponding visual force estimate $\tilde{f}_t$. These annotations produce force-augmented demonstrations for policy training.

For each task, we select a target force f_{target} from the estimated force values in its demonstration dataset. This task-specific target is provided to the action-selection procedure at test time.

D. Action-Force Proposal Policy and Inference-Time Selection

We use the force-augmented demonstrations to train an action-force proposal policy. At test time, we sample candidate action-force pairs from this policy and select an action using the task-specific target force.

Joint Action-Force Policy Training. We train a stochastic policy π_θ to jointly generate an end-effector action and an associated force value from the current observation:

$$\left(\mathbf{x}_t^{\text{eef}}, \hat{f}_t \right) \sim \pi_\theta(\cdot \mid \mathbf{o}_t). \quad (2)$$

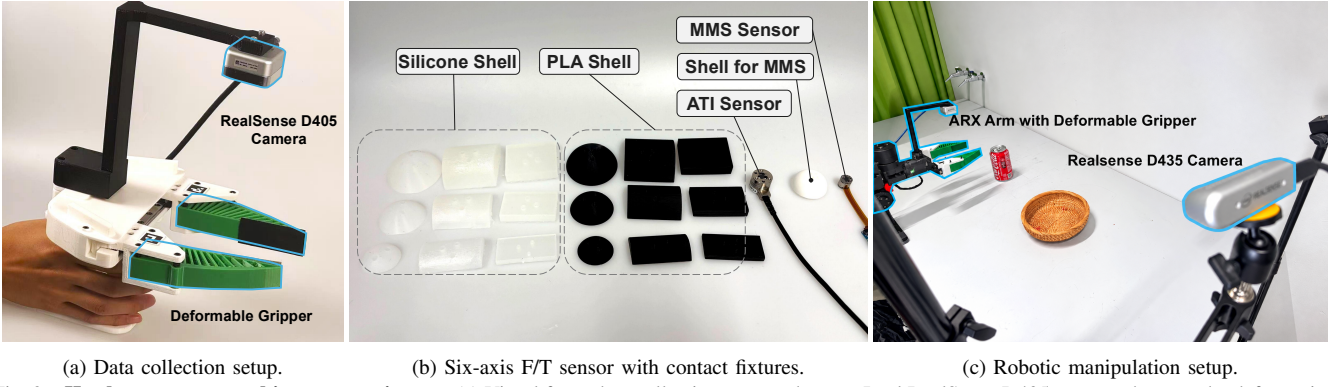


Fig. 3. **Hardware setups used in our experiments.** (a) Visual force data collection setup, where an Intel RealSense D405 camera observes the deformation of the hand-held Fin Ray gripper and a six-axis F/T sensor records the ground-truth interaction force through 3D-printed contact fixtures. (b) The six-axis F/T sensor and 3D-printed contact fixtures used to record ground-truth interaction force during data collection. (c) Robotic manipulation platform used for policy evaluation, consisting of an ARX robotic arm equipped with wrist and side-view cameras.

During training, the action component is supervised by the demonstrated end-effector action, while the force component is supervised by the visual force label $\hat{f}_t$. The policy therefore learns the correspondence between task-relevant actions and the interaction forces observed in the demonstrations.

Importantly, π_θ is not an action-conditioned force estimator. The visual force estimator g_ϕ predicts force solely from the visual deformation of the gripper, whereas π_θ jointly generates an action and a force value from the current observation. The generated force component is used to rank candidate actions during inference and is not directly sent to the robot as a control command.

Force Search. Given the current observation $\mathbf{o}_t$, we independently sample $K = 32$ candidate action–force pairs from Eq. (2):

$$\left\{ \left(\mathbf{x}_t^{\text{eef},(k)}, \hat{f}_t^{(k)} \right) \right\}_{k=1}^K \sim \pi_\theta(\cdot | \mathbf{o}_t).$$

Given a task-specified target force f_{target} , we select the candidate whose generated force component is closest to the target and execute its corresponding end-effector action:

$$k^* = \arg \min_{k \in \{1, \dots, K\}} \left| \hat{f}_t^{(k)} - f_{\text{target}} \right|, \quad \mathbf{x}_t^{\text{eef},*} = \mathbf{x}_t^{\text{eef},(k^*)}. \quad (3)$$

Only the selected end-effector action $\mathbf{x}_t^{\text{eef},*}$ is sent to the robot. The generated force value is used only for candidate selection. This procedure is repeated at each timestep, allowing the target force to guide action selection without requiring a force/torque or tactile sensor during deployment.

Algorithm 1 summarizes the inference-time search procedure.

IV. EXPERIMENT

Our experiments evaluate whether visually estimated force can improve policy performance on contact-rich manipulation tasks.

A. Experimental Setup

For visual force data collection, we use a hand-held Fin Ray gripper equipped with an Intel RealSense D405 camera, as shown in Fig. 3. Ground-truth axial-force measurements are recorded using both a Mitsumi MMS101 resistive force

Algorithm 1 Inference-Time Force Search

Require: current observation $\mathbf{o}_t$, target force f_{target} , number of candidates K , frozen policy π_θ

- 1: **for** $k = 1$ to K **do**
- 2: sample $\left(\mathbf{x}_t^{\text{eef},(k)}, \hat{f}_t^{(k)} \right) \sim \pi_\theta(\cdot | \mathbf{o}_t)$
- 3: **end for**
- 4: $k^* \leftarrow \arg \min_{k \in \{1, \dots, K\}} \left| \hat{f}_t^{(k)} - f_{\text{target}} \right|$
- 5: $\mathbf{x}_t^{\text{eef},*} \leftarrow \mathbf{x}_t^{\text{eef},(k^*)}$
- 6: execute $\mathbf{x}_t^{\text{eef},*}$ on the robot {generated force used only for selection}
- 7: advance to timestep $t + 1$ and repeat

sensor and an ATI six-axis force/torque (F/T) sensor supporting the 3D-printed contact fixtures. The reference sensor is used only to collect supervision for the visual force estimator; it is not attached to the robot during manipulation-policy deployment.

For the robot manipulation experiments, we train a stochastic policy using the Diffusion Policy architecture [54] and evaluate it on an ARX robot arm equipped with the same Fin Ray gripper design. An Intel RealSense D405 wrist camera observes the deformation of the gripper, while an Intel RealSense D435 side-view camera captures the global robot–object interaction. The policy is trained and executed on a single NVIDIA GeForce RTX 5090 GPU.

B. Force Estimation Evaluation

Evaluation Data. We collect three evaluation episodes with 442 frames in total. All three episodes are excluded from visual estimator training and current-baseline calibration. We record synchronized reference force measurements using a Nextech DFS-XM force gauge with an external Micro-SBeam load cell (Fig. 4(a)).

Compared Estimators. We compare visual and current-based force estimation. The current baseline uses $\hat{F} = aI + b$, where I is gripper motor current. We fit a and b using ten separate calibration recordings. The visual estimator uses gripper deformation as discussed previously.

Evaluation Protocol. We evaluate both estimators on these episodes. We compute mean absolute error (MAE) within

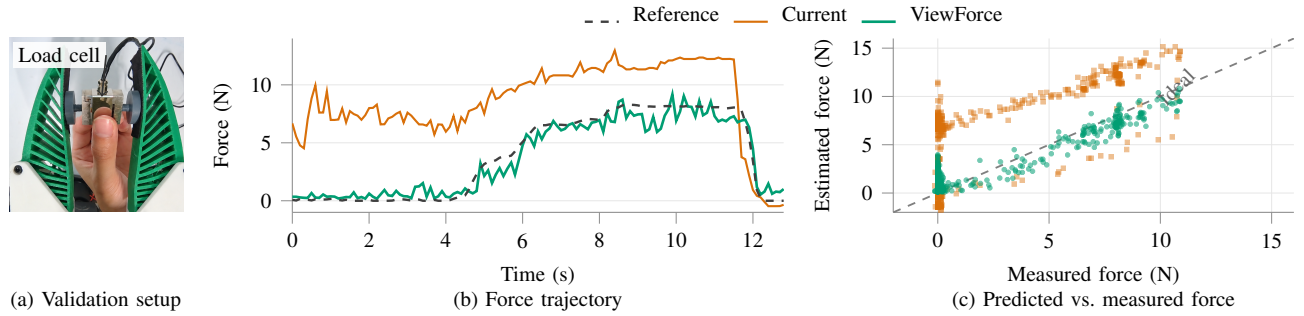


Fig. 4. **Force-estimation accuracy.** (a) Reference load cell. (b) Force trajectories. (c) Estimated versus measured force (three episodes, 442 frames); dashed line: perfect agreement. Mean episode MAE: motor current, 4.20N; ViewForce, 0.91N (equal episode weights).

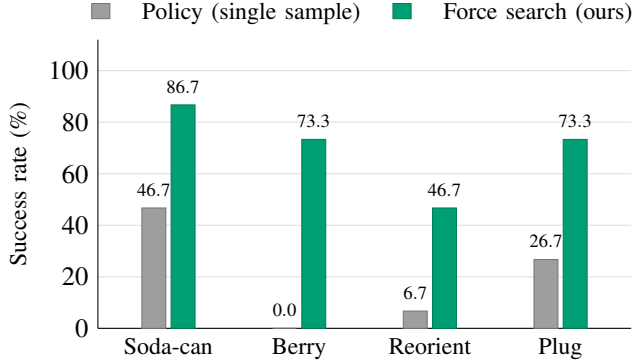


Fig. 5. **Manipulation success rates.** Success rate (%) over 15 trials per task for single-sample policy inference and our force-search method, both using the same policy checkpoint.

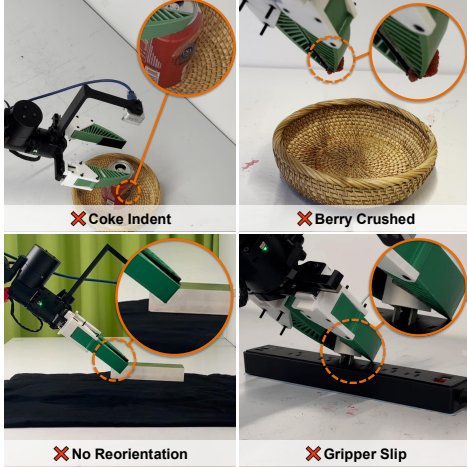


Fig. 6. **Representative failure modes under single-sample policy inference.** From top left to bottom right: a dented empty soda can, a crushed berry, unsuccessful in-hand reorientation, and plug slippage within the gripper. These examples correspond to the four tasks reported in Fig. 5.

each episode. This is the average absolute difference between estimated and measured force. We then average the three episode MAEs with equal weight.

C. Manipulation Evaluation

Experiments Design. We evaluate our method on four force-sensitive manipulation tasks. Grasping an empty soda can tests force regulation when handling a thin-walled object: too little force causes slippage, whereas too much force

permanently deforms the can. Berry picking additionally requires precise control at a much smaller scale, as the berry’s small size and compliance produce subtle deformation cues and leave only a narrow margin between a stable grasp and crushing. In-hand reorientation requires the robot to pivot a metal plate upright without slipping or lifting it from the table, while plug insertion requires sufficient gripping force to prevent the plug from slipping within the gripper during insertion. Success requires transferring the can or berry without damage, reorienting the plate upright, or fully inserting the plug.

Demonstration Data and Policy Training. For each task, we collect 40 demonstrations using the ARX robot arm. We then freeze the visual force estimator g_ϕ and use it to annotate every demonstration timestep with the estimated force magnitude $\tilde{f}_t$. The resulting action–force pairs are used to train a stochastic policy that jointly generates an end-effector action and its associated force value.

Compared Methods. Both procedures use the same trained policy checkpoint and differ only in how policy samples are selected at inference time:

- **Policy (single sample):** We draw one action–force sample from the policy, ignore its generated force component, and execute the corresponding end-effector action.
- **Ours (Force Search):** We draw $K = 32$ action–force samples from the same policy and execute the action whose generated force component is closest to the task-specific target force.

For each task, we select a target force from the estimated force values in its demonstration dataset.

Evaluation Protocol. We evaluate each method for 15 trials on every task. Both methods are evaluated under the same initial-condition distribution. We report the number and percentage of successful trials.

V. RESULTS AND ANALYSIS

A. Force-Estimation Results

We first quantify the accuracy of visually predicted forces and compare with a conventional baseline that uses gripper motor current for force estimation.

Estimation Accuracy. The visual estimator achieves an MAE of 0.91N, compared with 4.20N for the current baseline.

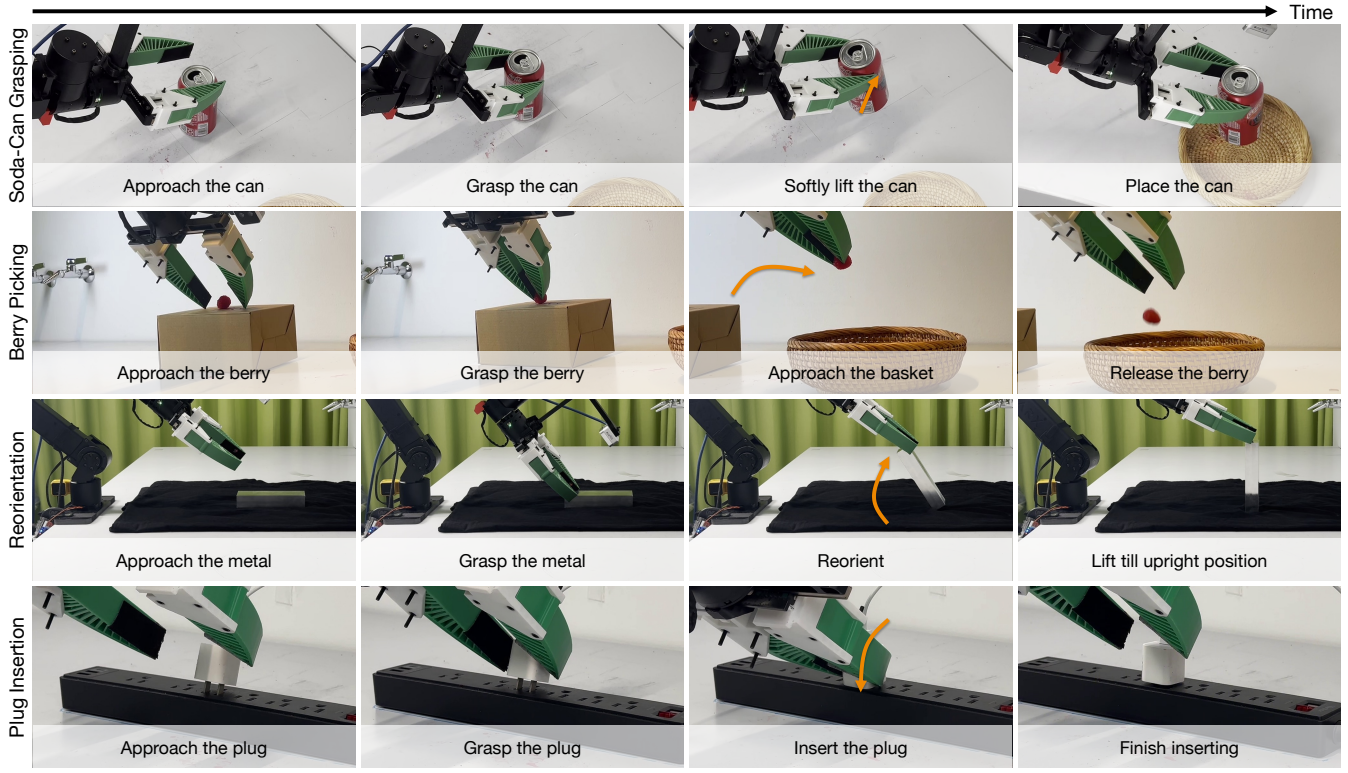


Fig. 7. **Representative successful rollouts of four force-sensitive manipulation tasks using force-based candidate selection.** Each row shows a task sequence progressing from left to right. For empty soda can grasping and berry picking, the robot applies sufficient gripping force to lift and transfer the object without crushing it. During in-hand reorientation, the robot maintains contact while allowing the metal plate to pivot into an upright configuration. During plug insertion, the robot maintains sufficient contact friction to prevent slippage and fully inserts the plug into the socket.

Temporal Tracking. Figure 4(b) compares the estimates over a representative force trajectory. The visual estimate remains close to zero during the initial low-force period. It follows the rise in measured force and returns toward zero as the force drops near the end. The current baseline reports substantial force even when the reference measurement is near zero. It also overestimates force during loading. The visual estimate fluctuates at higher forces but remains closer to the reference. Gripper deformation therefore provides a more faithful estimate of both low-force periods and changes during contact.

Prediction Accuracy. Figure 4(c) compares estimated and measured forces across all three evaluation episodes. Points on the dashed line represent perfect predictions. Most current-based estimates lie above the line, showing that the baseline overestimates force. It often predicts substantial force when the measured force is near zero. Visual estimates lie closer to the line, although some errors remain at low and high forces. Overall, visual deformation gives more accurate force estimates than motor current in the tested conditions.

B. Force-Aware Manipulation Results

We next evaluate whether force-based action selection improves success rates on four force-sensitive manipulation tasks compared with single-sample inference from the same policy.

Success Rates. Figure 5 shows that force search improves success rates across all four tasks: from 46.7% to 86.7%

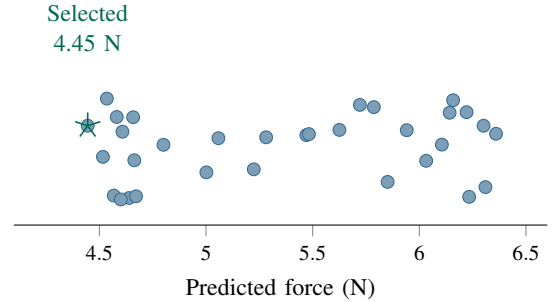


Fig. 8. **Force Search.** Distribution of 32 force predictions sampled from the policy for a held-out soda-can observation. The teal star marks the prediction selected by force search (4.45 N). Points are offset vertically for readability.

for can grasping, 0% to 73.3% for berry picking, 6.7% to 46.7% for reorientation, and 26.7% to 73.3% for plug insertion. Berry picking benefits most from force search, while reorientation remains the most challenging task, reaching a success rate of 46.7%.

Force Search. Figure 8 illustrates how force search selects among 32 action–force samples generated by the policy for a held-out soda-can observation. Predicted forces range from 4.45 to 6.36 N, with a mean of 5.36 N and a standard deviation of 0.69 N. Force search selects the candidate closest to the task-specific target force; in this example, the selected candidate also has the lowest predicted force, 4.45 N. Single-sample inference instead executes a randomly drawn action without considering its predicted force.

Qualitative Behavior. Figure 6 shows failures under single-

sample inference. The can is dented and the berry is crushed. Both indicate excessive gripping force. The plate fails to reorient, and the plug slips within the gripper. The examples highlight the need to balance a secure grasp with the contact conditions required by each task.

Figure 7 shows successful rollouts with force search. The robot lifts and transfers the can and berry without visible damage. During reorientation, the plate pivots against the table while remaining in contact with the gripper. During insertion, the plug stays secure as the robot pushes it into the socket. The rollouts illustrate how suitable gripping force supports both gentle handling and tasks that require friction.

The four tasks illustrate two roles of gripping force: balancing grasp stability against damage, and providing friction for reorientation and insertion. Together, the results show that visually estimated force can guide action selection across both types of interaction.

VI. LIMITATIONS

Our work has several limitations. First, our method requires a task-specific target force, which is obtained from successful demonstrations in our experiments. Automatically inferring and adapting the desired force based on object properties, contact state, and task phase remains an important direction for future work. Inferring object properties to parameterize adaptive grasp policies, as explored in DeliGrasp [58], offers one possible direction for selecting task-appropriate target forces. Second, our current method estimates only the normal gripping force F_z and does not recover tangential forces or torques. Extending visual force estimation to the full contact wrench would require deformable gripper with rich multi-axis deformation cues. Third, action selection relies on policy-generated force predictions without online visual force feedback, limiting correction of prediction errors and unexpected contact disturbances. Future work could incorporate this feedback into action execution and evaluate robustness under external disturbances.

VII. CONCLUSIONS

In this work, we presented a force-aware manipulation approach that estimates normal gripping force directly from the visible deformation of a soft gripper. We use these visual force estimates to supervise a policy that jointly generates end-effector actions and their associated force values. At inference time, force search samples multiple action–force candidates and selects the action whose predicted force best matches a task-specific target. This enables force-aware action selection without dedicated force or tactile sensors, policy updates, or changes to the robot control interface at deployment.

Across four force-sensitive manipulation tasks, force search consistently improves success over single-sample policy inference. These results show that visual gripper deformation can provide a useful contact signal not only for force estimation, but also for guiding policy inference.

More broadly, our results suggest that a robot’s compliant structure can serve as a visually observable sensing medium.

We hope this perspective encourages the development of low-cost manipulation systems that recover physical interaction signals directly from visual observations of robot deformation.

REFERENCES

- [1] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, and S. Levine, “More than a feeling: Learning to grasp and regrasp using vision and touch,” *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3300–3307, 2018.
- [2] H. Qi, B. Yi, S. Suresh, M. Lambeta, Y. Ma, R. Calandra, and J. Malik, “General in-hand object rotation with vision and touch,” in *Proceedings of The 7th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 229. PMLR, 2023, pp. 2549–2564.
- [3] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, “Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks,” in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 8943–8950.
- [4] H. Chen, J. Xu, H. Chen, K. Hong, B. Huang, C. Liu, J. Mao, Y. Li, Y. Du, and K. Driggs-Campbell, “Multi-modal manipulation via multi-modal policy consensus,” in *2026 IEEE International Conference on Robotics and Automation (ICRA)*, 2026.
- [5] B. Huang, Y. Wang, X. Yang, Y. Luo, and Y. Li, “3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing,” in *Proceedings of The 8th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 270. PMLR, 2024, pp. 2557–2578.
- [6] GelSight Inc., “GelSight Mini,” <https://www.gelsight.com/gelsightmini/>, accessed: 2026-07-24.
- [7] ATI Industrial Automation, “Multi-axis force/torque sensors,” accessed: Sep. 7, 2026. [Online]. Available: <https://www.atia.com/products/ft/sensors.aspx>
- [8] Bota Systems AG, “Precision 6-axis force torque sensors for robotics,” accessed: Sep. 7, 2026. [Online]. Available: <https://shop.botasys.com/shop/category/force-torque-sensors-4>
- [9] D. De Barrie, M. Pandya, H. Pandya, M. Hanheide, and K. Elgencidy, “A deep learning method for vision based force prediction of a soft Fin Ray gripper using simulation data,” *Frontiers in Robotics and AI*, vol. 8, p. 631371, 2021.
- [10] T. Kitamura, K. Matsushita, and N. Nakatani, “Analysis of contact force and shape change on grasping a square object using an actual Fin Ray soft gripper,” *Sensors*, vol. 23, no. 24, p. 9827, 2023.
- [11] J. Zhang, H. Yao, J. Mo, S. Chen, Y. Xie, S. Ma, R. Chen, T. Luo, W. Ling, L. Qin, Z. Wang, and W. Zhou, “Finger-inspired rigid-soft hybrid tactile sensor with superior sensitivity at high frequency,” *Nature Communications*, vol. 13, p. 5076, 2022.
- [12] K. Shimomura, “Tactile image sensors employing camera: A review,” *Sensors*, vol. 19, no. 18, p. 3933, 2019.
- [13] W. Yuan, S. Dong, and E. H. Adelson, “GelSight: High-resolution robot tactile sensors for estimating geometry and force,” *Sensors*, vol. 17, no. 12, p. 2762, 2017.
- [14] S. Dong, W. Yuan, and E. H. Adelson, “Improved gelsight tactile sensor for measuring geometry and slip,” in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 137–144.
- [15] E. Donlon, S. Dong, M. Liu, J. Li, E. Adelson, and A. Rodriguez, “GelSlim: A high-resolution, compact, robust, and calibrated tactile-sensing finger,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 1927–1934.
- [16] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, D. Jayaraman, and R. Calandra, “DIGIT: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation,” *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 3838–3845, 2020.
- [17] B. Ward-Cherrier, N. Pestell, L. Cramphorn, B. Winstone, M. E. Giannaccini, J. Rossiter, and N. F. Lepora, “The TacTip family: Soft optical tactile sensors with 3D-printed biomimetic morphologies,” *Soft Robotics*, vol. 5, no. 2, pp. 216–227, 2018.
- [18] C. Lin, Z. Lin, S. Wang, and H. Xu, “DTact: A vision-based tactile sensor that measures high-resolution 3d geometry directly from darkness,” 2022. [Online]. Available: <https://arxiv.org/abs/2209.13916>

- [19] C. Lin, H. Zhang, J. Xu, L. Wu, and H. Xu, "9DTact: A compact vision-based tactile sensor for accurate 3D shape reconstruction and generalizable 6D force estimation," *arXiv preprint arXiv:2308.14277*, 2023.
- [20] C. Sferazza and R. D'Andrea, "Design, motivation and evaluation of a full-resolution optical tactile sensor," *Sensors*, vol. 19, no. 4, p. 928, 2019.
- [21] S. Q. Liu and E. H. Adelson, "GelSight Fin Ray: Incorporating tactile sensing into a soft compliant robotic gripper," in *2022 IEEE 5th International Conference on Soft Robotics (RoboSoft)*, 2022, pp. 925–931.
- [22] S. Q. Liu, Y. Ma, and E. H. Adelson, "GelSight Baby Fin Ray: A compact, compliant, flexible finger with high-resolution tactile sensing," in *2023 IEEE International Conference on Soft Robotics (RoboSoft)*, 2023, pp. 1–8.
- [23] D. Stuhne, J. Vuletić, M. Car, and M. Orsag, "Design and prototyping of soft finger AI-enabled hand (SofIA)," in *2022 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, Jul. 2022, pp. 1581–1586.
- [24] Y. Zhu, M. Hao, X. Zhu, Q. Bateux, A. Wong, and A. M. Dollar, "Forces for free: Vision-based contact force estimation with a compliant hand," *Science Robotics*, vol. 10, no. 103, p. eadq5046, Jun. 2025.
- [25] R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E. H. Adelson, and S. Levine, "The feeling of success: Does touch sensing help predict grasp outcomes?" in *Proceedings of the 1st Annual Conference on Robot Learning (CoRL)*, 2017. [Online]. Available: <https://proceedings.mlr.press/v78/calandra17a/calandra17a.pdf>
- [26] M. Bauza, A. Bronars, and A. Rodriguez, "Tac2Pose: Tactile object pose estimation from the first touch," *The International Journal of Robotics Research*, vol. 42, no. 13, pp. 1185–1209, 2023.
- [27] S. Suresh, H. Qi, T. Wu, T. Fan, L. Pineda, M. Lambeta, J. Malik, M. Kalakrishnan, R. Calandra, M. Kaess, J. Ortiz, and M. Mukadam, "NeuralFeels with neural fields: Visuotactile perception for in-hand manipulation," *Science Robotics*, vol. 9, no. 96, p. ead10628, 2024.
- [28] S. Tian, F. Ebert, D. Jayaraman, M. Mudigonda, C. Finn, R. Calandra, and S. Levine, "Manipulation by feel: Touch-based control with deep predictive models," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 818–824.
- [29] F. R. Hogan, M. Bauza, O. Canal, E. Donlon, and A. Rodriguez, "Tactile regrasp: Grasp adjustments via simulated tactile transformations," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 2963–2970.
- [30] F. R. Hogan, J. Ballester, S. Dong, and A. Rodriguez, "Tactile dexterity: Manipulation primitives with tactile feedback," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 8863–8869.
- [31] Y. Chen, M. Van der Merwe, A. Sipos, and N. Fazeli, "Visuo-tactile transformers for manipulation," in *6th Annual Conference on Robot Learning*, 2022.
- [32] H. Li, Y. Zhang, J. Zhu, S. Wang, M. A. Lee, H. Xu, E. Adelson, L. Fei-Fei, R. Gao, and J. Wu, "See, hear, and feel: Smart sensory fusion for robotic manipulation," in *Proceedings of The 6th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 205. PMLR, 2023, pp. 1368–1378. [Online]. Available: <https://proceedings.mlr.press/v205/li23c.html>
- [33] J. Clark, C. Wang, Y. Gao, S. Hong, H. Choi, M. Cutkosky, Y. Hou, and S. Song, "Multisensory continual learning: Adapting pretrained visuomotor policies to force," 2026. [Online]. Available: <https://arxiv.org/abs/2606.30988>
- [34] Y. She, S. Wang, S. Dong, N. Sunil, A. Rodriguez, and E. Adelson, "Cable manipulation with a tactile-reactive gripper," *The International Journal of Robotics Research*, vol. 40, no. 12-14, pp. 1385–1401, 2021.
- [35] Z. Yu, W. Xu, S. Yao, J. Ren, T. Tang, Y. Li, G. Gu, and C. Lu, "Precise robotic needle-threading with tactile perception and reinforcement learning," in *Proceedings of The 7th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 229. PMLR, 2023, pp. 3266–3276. [Online]. Available: <https://proceedings.mlr.press/v229/yu23c.html>
- [36] B. Ai, S. Tian, H. Shi, Y. Wang, C. Tan, Y. Li, and J. Wu, "RoboPack: Learning tactile-informed dynamics models for dense packing," in *Proceedings of Robotics: Science and Systems (RSS)*, 2024. [Online]. Available: <https://arxiv.org/abs/2407.01418>
- [37] N. Fazeli, M. Oller, J. Wu, Z. Wu, J. B. Tenenbaum, and A. Rodriguez, "See, feel, act: Hierarchical learning for complex manipulation skills with multisensory fusion," *Science Robotics*, vol. 4, no. 26, p. eaav3123, 2019.
- [38] F. R. Hogan and A. Rodriguez, "Reactive planar non-prehensile manipulation with hybrid model predictive control," *The International Journal of Robotics Research*, vol. 39, no. 7, pp. 755–773, 2020.
- [39] H. Chen, Y. Ma, S. Tian, X. Han, W. Huang, F. Wu, Y. Li, J. Wu, E. H. Adelson, and Y. Du, "Costream: Composing simple behaviors for generalizable complex manipulation," 2026. [Online]. Available: <https://arxiv.org/abs/2606.26423>
- [40] X. Kang, T. Tian, S.-W. Lee, B. Huang, Y. Li, and Y.-L. Kuo, "Learning force-regulated manipulation with a low-cost tactile-force-controlled gripper," *arXiv preprint arXiv:2602.10013*, 2026. [Online]. Available: <https://arxiv.org/abs/2602.10013>
- [41] Y. Hou, Z. Liu, C. Chi, E. Cousineau, N. Kuppaswamy, S. Feng, B. Burchfiel, and S. Song, "Adaptive compliance policy: Learning approximate compliance for diffusion guided control," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 4829–4836. [Online]. Available: <https://arxiv.org/abs/2410.09309>
- [42] H. Choi, Y. Hou, C. Pan, S. Hong, A. Patel, X. Xu, M. R. Cutkosky, and S. Song, "In-the-wild compliant manipulation with UMI-FT," *arXiv preprint arXiv:2601.09988*, 2026. [Online]. Available: <https://arxiv.org/abs/2601.09988>
- [43] H. Chen, Y. Niu, K. Hong, S. Liu, Y. Wang, Y. Li, and K. R. Driggs-Campbell, "Predicting object interactions with behavior primitives: An application in stowing tasks," in *7th Annual Conference on Robot Learning*, 2023.
- [44] N. Ma, S. Tong, H. Jia, H. Hu, Y.-C. Su, M. Zhang, X. Yang, Y. Li, T. Jaakkola, X. Jia, and S. Xie, "Scaling inference time compute for diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 2523–2534.
- [45] X. Zhang, H. Lin, H. Ye, J. Zou, J. Ma, Y. Liang, and Y. Du, "Inference-time scaling of diffusion models through classical search," 2025. [Online]. Available: <https://arxiv.org/abs/2505.23614>
- [46] M. Janner, Y. Du, J. Tenenbaum, and S. Levine, "Planning with diffusion for flexible behavior synthesis," in *Proceedings of the 39th International Conference on Machine Learning*, 2022.
- [47] M. Reuss, M. Li, X. Jia, and R. Lioutikov, "Goal-conditioned imitation learning using score-based diffusion policies," in *Proceedings of Robotics: Science and Systems (RSS)*, 2023.
- [48] M. Nakamoto, O. Mees, A. Kumar, and S. Levine, "Steering your generalists: Improving robotic foundation models via value guidance," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: <https://arxiv.org/abs/2410.13816>
- [49] Y. Liu, J. I. Hamid, A. Xie, Y. Lee, M. Du, and C. Finn, "Bidirectional decoding: Improving action chunking via guided test-time sampling," *arXiv preprint arXiv:2408.17355*, 2024. [Online]. Available: <https://arxiv.org/abs/2408.17355>
- [50] H. Qi, H. Yin, A. Zhu, Y. Du, and H. Yang, "Inference-time enhancement of generative robot policies via predictive world modeling," *IEEE Robotics and Automation Letters*, 2026. [Online]. Available: <https://arxiv.org/abs/2502.00622>
- [51] M. Du and S. Song, "Dynaguide: Steering diffusion policies with active dynamic guidance," in *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS)*, 2025.
- [52] S. Liu, I. S. Singh, Y. Xu, J. Duan, and R. Krishna, "VIs: Steering pretrained robot policies via vision-language models," *arXiv preprint arXiv:2602.03973*, 2026.
- [53] Z. Zhang, J. Ma, X. Yang, X. Wen, Y. Zhang, B. Li, Y. Qin, J. Liu, C. Zhao, L. Kang, H. Hong, Z. Yin, P. Torr, H. Su, R. Zhang, and D. Ma, "Touchguide: Inference-time steering of visuomotor policies via touch guidance," 2026. [Online]. Available: <https://arxiv.org/abs/2601.20239>
- [54] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Proceedings of Robotics: Science and Systems (RSS)*, 2023.
- [55] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *The Eleventh International Conference on Learning Representations*, 2023.
- [56] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollar, and C. Feichtenhofer,

- “SAM 2: Segment anything in images and videos,” in *The Thirteenth International Conference on Learning Representations*, 2025.
- [57] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, ser. Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M. W. III, and A. F. Frangi, Eds., vol. 9351. Springer, 2015, pp. 234–241.
- [58] W. Xie, M. Valentini, J. Lavering, and N. Correll, “DeliGrasp: Inferring object properties with LLMs for adaptive grasp policies,” in *8th Annual Conference on Robot Learning*, 2024.