Vi-TacMan: Articulated Object Manipulation via Vision and Touch

Leiyao Cui[1,2,3,4,6] Zihang Zhao[3,4,5,6,7,†] Sirui Xie[3] Wenhuan Zhang[3] Zhi Han[2] Yixin Zhu[4,3,5,6,8,†]

https://vi-tacman.github.io

***Abstract*— Autonomous manipulation of articulated objects remains a fundamental challenge for robots in human environments. Vision-based methods can infer hidden kinematics but can yield imprecise estimates on unfamiliar objects. Tactile approaches achieve robust control through contact feedback but require accurate initialization. This suggests a natural synergy: vision for global guidance, touch for local precision. Yet no framework systematically exploits this complementarity for generalized articulated manipulation. Here we present Vi-TacMan, which uses vision to propose grasps and coarse directions that seed a tactile controller for precise execution. By incorporating surface normals as geometric priors and modeling directions via von Mises-Fisher (vMF) distributions, our approach achieves significant gains over baselines (all p<0.0001). Critically, manipulation succeeds without explicit kinematic models—the tactile controller refines coarse visual estimates through real-time contact regulation. Tests on more than 50,000 simulated and diverse real-world objects confirm robust cross-category generalization. This work establishes that coarse visual cues suffice for reliable manipulation when coupled with tactile feedback, offering a scalable paradigm for autonomous systems in unstructured environments.**

I. Introduction

Household robots must master articulated object manipulation to function effectively in human environments, yet face enormous diversity in object appearance, geometry, and kinematics [1–4]. Unlike structured industrial settings where objects are standardized, everyday articulated structures—cabinets, refrigerators, ovens—exhibit vast variability that renders precise a priori modeling impractical [5]. This variability poses a fundamental challenge: reliable manipulation requires both accurate localization of interaction points and precise execution of kinematically-constrained motions. The question then becomes: which sensory modality is best suited to address each aspect of this challenge?

L. Cui, Z. Zhao, S. Xie, and W. Zhang contributed equally to this work. † Corresponding authors. Emails: zhaozihang@stu.pku.edu.cn and yixin.zhu@pku.edu.cn.

[1] Shenyang Institute of Automation, Chinese Academy of Sciences [2] University of Chinese Academy of Sciences [3] Institute for Artificial Intelligence, Peking University [4] School of Psychological and Cognitive Sciences, Peking University [5] State Key Lab of General AI, Peking University [6] Beijing Key Laboratory of Behavior and Mental Health, Peking University [7] LeapZenith AI Research [8] Embodied Intelligence Lab, PKU-Wuhan Institute for Artificial Intelligence

This work is supported in part by the Brain Science and Brain-like Intelligence Technology–National Science and Technology Major Project (2025ZD0219400), the National Natural Science Foundation of China (62376009), the State Key Lab of General AI at Peking University, the PKU-BingJi Joint Laboratory for Artificial Intelligence, the Wuhan Major Scientific and Technological Special Program (2025060902020304), the Hubei Embodied Intelligence Foundation Model Research and Development Program, and the National Comprehensive Experimental Base for Governance of Intelligent Society, Wuhan East Lake High-Tech Development Zone.

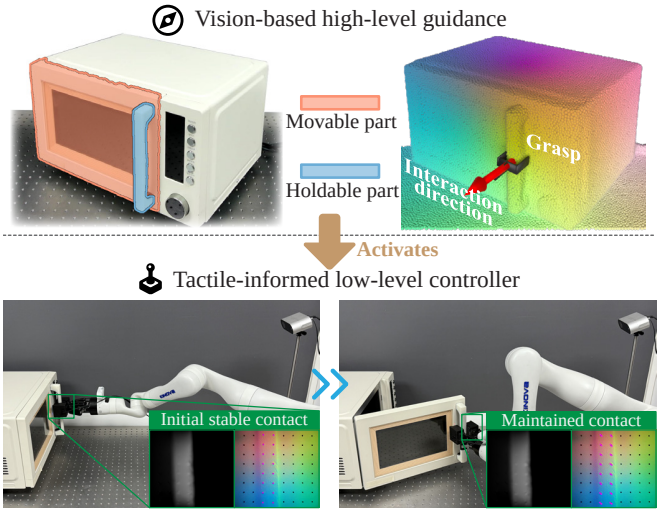


Fig. 1: Overview of Vi-TacMan. Vi-TacMan exploits the complementary strengths of vision and touch for manipulating unseen articulated objects. Vision provides global context to propose grasps and estimate coarse interaction directions, which initialize a tactile controller that leverages local contact feedback for precise and robust execution.

Dominant approaches rely on vision to reconstruct object kinematics for manipulation planning [6–13]. Vision's global receptive field makes it well-suited for identifying interaction points across the entire object. However, articulation mechanisms are typically hidden within object interiors, forcing vision systems to infer kinematics from limited surface observations. This inverse problem proves brittle on unfamiliar objects: even state-of-the-art methods trained on large-scale datasets [1, 2, 4] produce imprecise kinematic estimates that fail during execution—particularly problematic in safety-critical home environments where reliability is paramount.

Recent tactile methods offer an alternative paradigm [14, 15]: rather than recovering precise kinematics, they maintain successful manipulation through continuous contact regulation. By directly sensing contact geometry, tactile feedback provides rich local information that vision cannot access. Critically, these approaches demonstrate that stable contact feedback enables reliable execution given only coarse initial conditions—a feasible grasp and approximate motion direction. This insight reframes the vision problem: precise kinematic recovery is unnecessary if vision provides sufficient cues to initialize tactile control. The natural division of labor emerges: vision for global, coarse guidance; touch for local, precise execution.

We present Vi-TacMan, a systematic framework exploiting this complementarity. Vision detects movable and holdable

parts, proposes grasps, and estimates coarse interaction directions; tactile feedback then refines execution through real-time contact regulation (Fig. 1). Three key technical components enable robust generalization to unseen objects. First, we incorporate surface normals as geometric priors for direction estimation, providing physical constraints that significantly improve performance (p<0.0001). Second, recognizing that multiple plausible directions may exist for unfamiliar objects, we model directional uncertainty via von Mises-Fishers (vMFs) on the unit sphere [16], enabling principled inference under ambiguity. Third, our detector achieves 0.86 mAP [17], reliably identifying interaction regions even in complex multi-part objects. Together, these components provide the sufficient initialization required by tactile control.

Our contributions are:

- We present Vi-TacMan, a vision-touch framework where coarse visual guidance activates precise tactile control for articulated manipulation.
- We develop a robust detection model achieving 0.86 mAP [17] that identifies movable and holdable parts in complex multi-component objects.
- We incorporate surface normals as geometric priors for direction estimation, yielding significant gains over baselines (all p<0.0001).
- We apply von Mises-Fisher (vMF) distributions to model directional uncertainty on the unit sphere, enabling principled inference under ambiguity.
- We validate our approach on over 50,000 simulations and diverse real objects, demonstrating reliable manipulation without explicit kinematic models.

The remainder of this paper is organized as follows: Sec. II presents our systematic approach to articulated object manipulation using vision and touch, with implementation details provided in Sec. III. The proposed approach is empirically validated in Sec. IV and concluded in Sec. V.

II. The Vi-TacMan Framework

In this section, we present Vi-TacMan, a systematic framework for manipulating articulated objects by integrating vision and touch. We first introduce the contact-regulation methods that motivate our framework in Sec. II-A. These methods require a stable grasp and a coarse direction estimate as initialization. To address these requirements, we formulate the problem as a maximum a posteriori (MAP) estimation task, decomposed into two tractable components in Sec. II-B. Finally, we describe our approach for estimating a distribution over coarse motion directions without constraining the solution to specific articulation types in Sec. II-C.

A. Background: Contact-Regulating Methods

Recent advances in articulated object manipulation demonstrate that kinematic priors are not strictly necessary if the robot regulates contact through tactile sensing [14, 15]. Given a coarse interaction direction, these methods iteratively adjust the end-effector pose by a transformation $T_\Delta \in \mathrm{SE}(3)$ such that the resulting contact returns to a stable state. Formally, the update is computed as

$$T_\Delta = \underset{T_\Delta \in \mathrm{SE}(3)}{\arg\min}\, f(\mathcal{C}_0, \mathcal{C}_{t+1}), \tag{1}$$

where $\mathcal{C}_0$ denotes the reference contact, $\mathcal{C}_{t+1}$ the contact after applying T_Δ, and $f(\cdot,\cdot)$ a metric measuring their difference. By maintaining contact stability rather than tracking kinematic models, this formulation naturally handles objects with unknown or imprecisely estimated kinematics.

This kinematic-invariant property is precisely what enables reliable manipulation across diverse objects: vision modules need not recover error-prone hidden kinematics. However, successful execution requires two (i) a proper grasp that establishes stable contact and (ii) a coarse interaction direction to trigger the controller. Our framework addresses these requirements through principled visual inference.

B. Problem Formulation

Contact-regulating methods assume the availability of an initial stable grasp and a coarse interaction direction. Given visual observation $\mathcal{V}$, our goal is to recover these prerequisites by estimating:

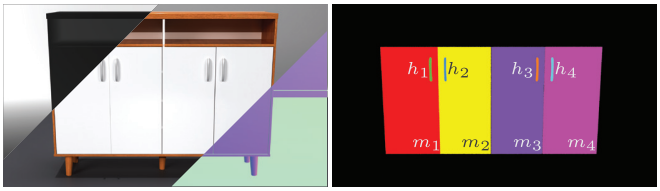
- A parallel-gripper grasp $G \in \mathrm{SE}(3) \times \mathbb{R}$, where the $\mathrm{SE}(3)$ component specifies the gripper pose and the scalar encodes the gripper width.
- An interaction direction $\boldsymbol{d} \in \mathbb{S}^2$, representing a unit vector on the 2-sphere.

Together, $(G, \boldsymbol{d})$ provide the initialization required for contact-regulation control.

In our setting, the visual observation consists of visually observable points $\mathcal{V} = \{P_i | i = 1, \ldots, n\}$, where each point P_i is represented by:

$$P = (\boldsymbol{p}, \boldsymbol{c}, \boldsymbol{n}, m, h). \tag{2}$$

As illustrated in Fig. 2, $\boldsymbol{p} \in \mathbb{R}^3$ denotes the 3D position in the camera frame, and $\boldsymbol{c} \in [0, 255]^3$ represents RGB color. The surface normal $\boldsymbol{n} \in \mathbb{S}^2$ provides geometric constraints that guide direction estimation beyond random guessing—a hypothesis we validate experimentally. The label $m \in \mathbb{N}$ specifies whether the point is movable ($m > 0$) or fixed ($m = 0$), with different positive values corresponding to distinct movable parts within a single object. Similarly, $h \in \mathbb{N}$ indicates whether the point provides a viable holdable



RGB-D with surface normals | Instance-level semantic masks

Fig. 2: Inputs to the vision module of Vi-TacMan. The vision module of Vi-TacMan processes RGB-D data from a depth sensor, surface normals computed from the depth map (visualized as a normal map), and instance-level semantic masks identifying holdable and movable parts. This representation accommodates objects with multiple interactable components. *Note:* Holdable masks are subsets of their associated movable masks; regions appear overlapped in the visualization.

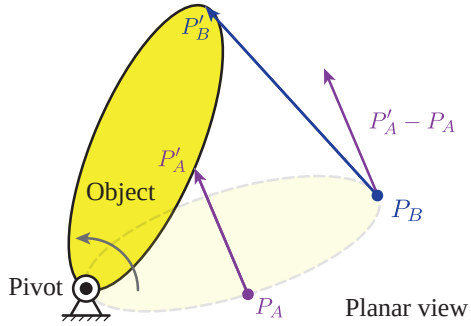


Fig. 3: Coupling between grasp point and interaction direction. The interaction direction depends on the selected grasp point even when the same rigid transformation is applied. Different point selections yield different directions under identical transformations.

location ($h > 0$) associated with a specific movable part. While position and color are obtained directly from depth sensing, the remaining attributes are inferred from them, as detailed in Sec. III-B.

Formally, we seek to obtain:

$$G^*, \boldsymbol{d}^* = \arg\max_{G,\boldsymbol{d}} p(G, \boldsymbol{d}|\mathcal{V}), \tag{3}$$

where $p(\cdot)$ represents a probability density function (PDF).

Directly modeling the joint density $p(G, \boldsymbol{d}|\mathcal{V})$ is challenging, yet treating G and $\boldsymbol{d}$ as conditionally independent is not justified. As illustrated in Fig. 3, the interaction direction depends on the grasp point $\boldsymbol{g} \in \mathbb{R}^3$ determined by G: even under the same rigid transformation, different grasp locations yield different directions.

We make the problem tractable by modeling the rigid transformation $T = [R \in \mathrm{SO}(3)|\boldsymbol{t} \in \mathbb{R}^3] \in \mathrm{SE}(3)$, which is independent of the specific grasping point. We then recover the interaction direction from T and point position $\boldsymbol{p}$ via:

$$\boldsymbol{d} = \frac{(R - I)\boldsymbol{p} + \boldsymbol{t}}{\|(R - I)\boldsymbol{p} + \boldsymbol{t}\|_2}, \tag{4}$$

where I is the 3×3 identity matrix. Then Eq. (3) can be reformulated as:

$$G^*, T^* = \arg\max_{G,T} p(G, T|\mathcal{V}) \tag{5}$$

$$= \underbrace{\arg\max_{G} p(G|\mathcal{V})}_{\text{grasp}} \underbrace{\arg\max_{T} p(T|\mathcal{V})}_{\text{direction}}, \tag{6}$$

where $p(G|\mathcal{V})$ and $p(T|\mathcal{V})$ separately model grasp selection and transformation estimation. Since parallel-jaw grasping is well-studied and does not affect T estimation, we defer implementation details to Sec. III-C. The remainder of this section focuses on estimating the transformation distribution $p(T|\mathcal{V})$, which determines the interaction direction.

C. Vision-Based Direction Estimation

We detail our method for estimating the rigid transformation of a movable part from visual inputs. Unlike prior methods restricted to specific joint types such as revolute or prismatic joints, our approach makes no such assumption. Real-world articulated objects often deviate from these idealized models [14], and although current datasets underrepresent such complexity, our method is designed to accommodate it.

Without assuming a predefined kinematic structure, we adopt a numerical approach to infer the rigid transformation. We introduce small perturbations to the movable part and analyze the resulting displacement patterns of associated points $\boldsymbol{p}_i$ between consecutive frames. Each point acquires a displacement vector $\boldsymbol{q}_i \in \mathbb{R}^3$ determined by T:

$$\boldsymbol{q}_i = T \begin{bmatrix} \boldsymbol{p}_i \\ 1 \end{bmatrix} - \boldsymbol{p}_i. \tag{7}$$

With sufficient point-displacement pairs $(\boldsymbol{p}_i, \boldsymbol{q}_i)$, we efficiently solve for T using the Kabsch algorithm [18].

Under rigid-body assumption, every sub-part within the movable component undergoes the same transformation. Evaluating Eq. (7) with different point combinations therefore provides insight into the conditional probability distribution $p(T|\mathcal{V})$. In an idealized scenario with perfect observations and strictly rigid motion, this distribution would collapse to a Dirac delta at the true transformation. Real-world conditions—noise, partial visibility, object complexity—introduce ambiguities that yield multiple plausible motion directions. This approach thus captures and represents uncertainties inherent in the vision-based model $p(T|\mathcal{V})$.

With grasp point $\boldsymbol{g}$ chosen to maximize the first term in Eq. (6), we map each candidate transformation T to its corresponding interaction direction $\boldsymbol{d}$ deterministically via Eq. (4). This mapping induces a distribution over directions $p(\boldsymbol{d}|\mathcal{V})$ from the underlying $p(T|\mathcal{V})$. To model this distribution on the unit sphere $\mathbb{S}^2$, we fit a vMF distribution to the sampled directions $\{\boldsymbol{d}_i\}_{i=1}^n$. The vMF distribution is formulated as:

$$p(\boldsymbol{d}|\mathcal{V}) = \frac{1}{c(\kappa, \boldsymbol{\mu})} \exp\left(\kappa \boldsymbol{\mu}^\mathsf{T} \boldsymbol{d}\right), \quad \boldsymbol{d} \in \mathbb{S}^2. \tag{8}$$

Analogous to a Gaussian distribution in Euclidean space, the vMF distribution employs two parameters: a mean direction $\boldsymbol{\mu} \in \mathbb{S}^2$ specifying the central location, and a concentration parameter $\kappa \in \mathbb{R}_{>0}$ controlling how tightly the distribution clusters around $\boldsymbol{\mu}$. The normalizing constant $c(\kappa, \boldsymbol{\mu})$ ensures that $p(\boldsymbol{d}|\mathcal{V})$ integrates to one over $\mathbb{S}^2$.

Since the normalizing constant and κ do not affect the maximizer, we obtain:

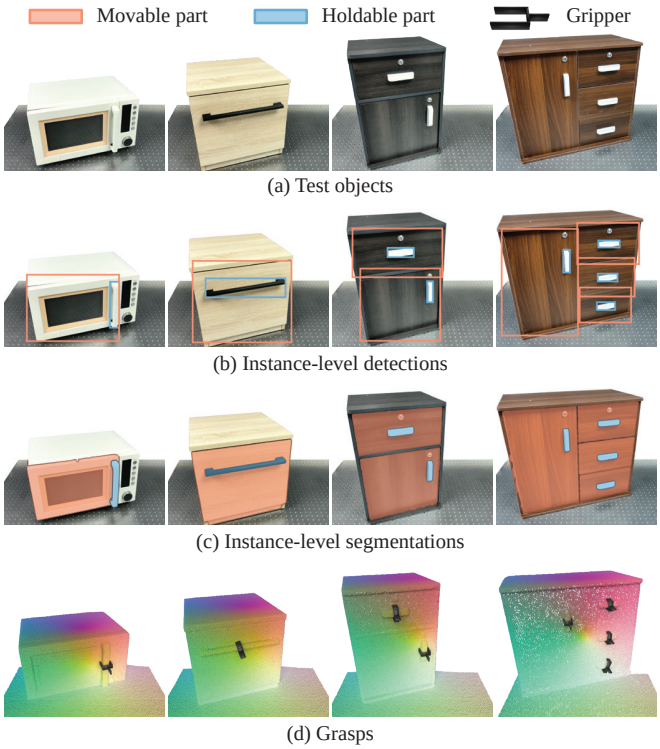
$$\arg\max_{\boldsymbol{d} \in \mathbb{S}^2} p(\boldsymbol{d}|\mathcal{V}) = \arg\max_{\boldsymbol{d} \in \mathbb{S}^2} \exp\left(\boldsymbol{\mu}^\mathsf{T} \boldsymbol{d}\right). \tag{9}$$

The density is maximized when $\boldsymbol{d}$ aligns with $\boldsymbol{\mu}$. We estimate $\boldsymbol{\mu}$ by computing the Fréchet mean of sampled directions under the geodesic metric (arc length) on the sphere, yielding an unbiased estimator:

$$\boldsymbol{d}^* = \hat{\boldsymbol{\mu}} = \arg\min_{\boldsymbol{\mu} \in \mathbb{S}^2} \sum_{i=1}^{n} \left|\arccos\left(\boldsymbol{\mu}^\mathsf{T} \boldsymbol{d}_i\right)\right|^2. \tag{10}$$

III. Implementation

In this section, we describe the implementation details of Vi-TacMan. We first introduce the dataset in Sec. III-A, which enables detection of movable and holdable parts in Sec. III-B. We then explain how to leverage sampling-based models for stable grasping, followed by learning-based acquisition of point displacements using the established dataset in Sec. III-D, which is critical for recovering interaction



(a) Test objects

(b) Instance-level detections

(c) Instance-level segmentations

(d) Grasps

Fig. 4: Real-world articulated objects and processing pipeline. (a) We evaluate Vi-TacMan on real-world objects spanning diverse configurations: prismatic to revolute joints, and single-part to multi-part structures. (b) Our trained detector reliably identifies movable and holdable parts, even in complex multi-part cases. (c) These detections provide prompts for the segmentation model, enabling fine-grained part segmentation. (d) Based on segmented parts, suitable grasps are generated at grasping points $\boldsymbol{g}$. These results provide the necessary information for inferring interaction directions.

directions. Finally, we present the tactile control policy in Sec. III-E.

A. Dataset Preparation

We construct a dataset to support learning-based extraction of movable and holdable features and direction estimation. We select 385 articulated objects spanning eight categories from the PartNet-Mobility dataset [1] and import them into the SAPIEN simulator, rendering them in ray tracing mode from up to 72 viewpoints. This process captures the color and positional information defined in Eq. (2). Surface normals are estimated by computing the cross product of vectors formed from each point and its neighbors to the right and below in image space. Movable and holdable instance labels m and h are obtained from GAPartNet annotations [19].

The dataset is divided at the category level: microwaves, refrigerators, storage furniture, and trash cans are assigned to the training set, while dishwashers, doors, ovens, and tables are reserved for testing. Within the training portion, we split data into training and validation subsets using an 8:2 ratio, yielding 39,524 training samples, 9,881 validation samples, and 5,836 test samples.

To evaluate performance beyond simulation, we collect four real-world examples, each captured from five viewpoints using a Femto Bolt depth sensor. One view is illustrated in Fig. 4(a); additional results appear in the supplementary video. These examples capture real-world diversity, including objects with single and multiple movable parts, and are reserved strictly for testing [20]. To improve depth quality, we first estimate a relative depth map using a depth foundation model [20]. Since this estimate lacks an absolute scale, we recover the correct scale by fitting a linear model between estimated disparities and ground-truth sensor measurements using RANSAC for robustness. The enhancement is illustrated in Fig. 5.

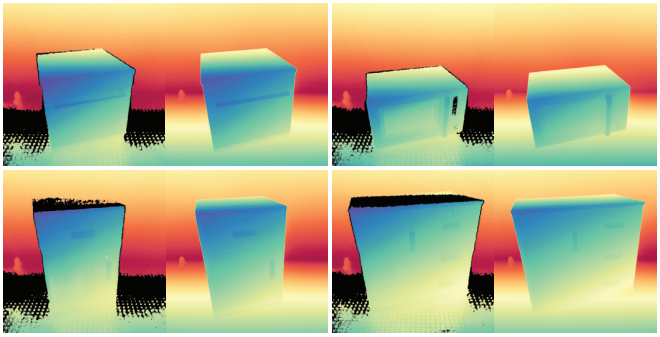


Fig. 5: Depth refinement using foundation models. We leverage a depth foundation model [20] to refine raw depth measurements from the image sensor. Left: raw depth. Right: refined depth. Both visualizations use the same colorbar range for comparability.

B. Movable and Holdable Part Segmentation

Using the prepared data from Sec. III-A, we derive the movable and holdable masks defined in Eq. (2), which serve as key inputs to our vision module. We train an object detector with a DINOv3 backbone and transformer-based head to detect movable and holdable parts [21, 22]. The model is trained using the AdamW optimizer on a single H100 GPU with batch size 2 and learning rate 6×10^{-6}. Following the protocol suggested by Lin *et al.* [17], we report mean Average Precision across IoU thresholds from 0.50 to 0.95 (mAP@[0.50:0.95]). The model attains 0.86 mAP on the test set; detailed breakdowns appear in Tab. I. Since mAP above 0.6 in multi-class settings is typically considered practically useful [17] and detection is not our primary contribution, we provide the model and checkpoints in code rather than extensive baseline comparisons.

TABLE I: Detection performance on the test set.

mAP	AP(50)	AP(75)	AP(S)	AP(M)	AP(L)
0.86	0.97	0.94	0.66	0.86	0.94

Detector outputs are passed to SAM2 [23] to produce final movable and holdable masks on an RTX 3090 GPU. We associate each holdable part with its corresponding movable part by selecting the pair whose mask intersection has the largest area. Real-world results are presented in Fig. 4(b)–(c) for illustration.

C. Grasp Selection

With movable and holdable masks defined, we establish a stable grasp on the handle. Recent advances demonstrate the effectiveness of parallel grippers for object grasping,

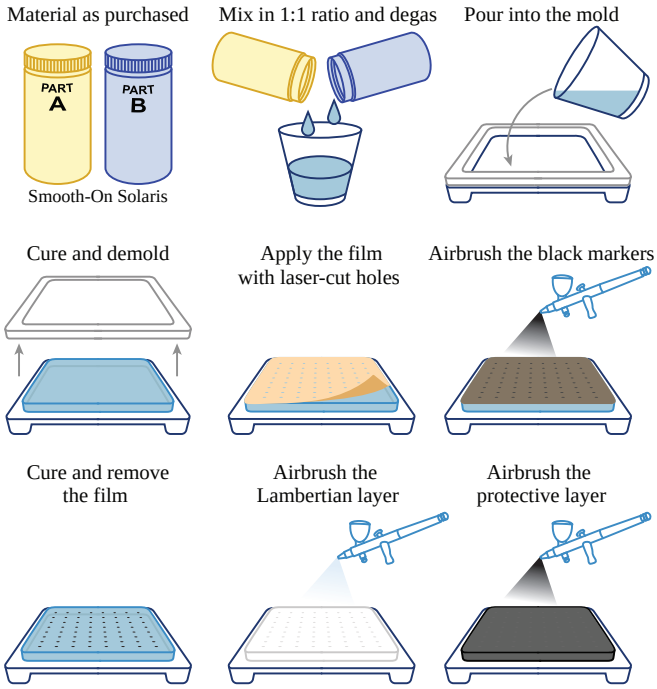


Fig. 6: Fabrication process for GelSight-style tactile sensor elastomer. We use Smooth-On Solaris silicone as the base elastomer. Marker placement is standardized using a laser-cut stencil to ensure uniform spacing and geometry. The Lambertian coating and protective topcoat are applied via airbrush.

even in cluttered environments [24–26]. The handle-grasping problem is largely simplified in our setting. We adopt a sampling-based method similar to Ten *et al*. [27], restricting the grasp region to the holdable area. The grasping point $\boldsymbol{g}$ is defined as the centroid of this region, which determines the gripper translation. We sample gripper rotations to identify one yielding a collision-free grasp with minimal gripper width. Considering the symmetry of the parallel gripper, we select the pose closest to the robot's home position [28]. Qualitative examples appear in Fig. 4(d).

D. Vision-Based Displacement Estimation

We estimate the displacement flow from visual inputs defined in Eq. (7) using a neural network based on PointNet++ [29]. The network takes point coordinates as input and augments them with surface normals in the movable region, along with movable masks as additional features.

Training uses the loss:

$$\mathcal{L} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\frac{\|\hat{\boldsymbol{q}}_i - \boldsymbol{q}_i\|_1}{\|\boldsymbol{q}_i\|_1}}_{\text{magnitude}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{\hat{\boldsymbol{q}}_i^{\mathsf{T}}\boldsymbol{q}_i}{\|\hat{\boldsymbol{q}}_i\|_2\|\boldsymbol{q}_i\|_2}\right)}_{\text{direction}}, \quad (11)$$

where n is the number of points and $\hat{\boldsymbol{q}}_i$ is the network's estimate. The first term penalizes magnitude error using relative ℓ_1 loss, which stabilizes optimization across a wide dynamic range and drives equal absolute errors toward zero regardless of scale. This is important because small displacements arise both outside masks and within masked regions near rotation axes. The second term aligns predicted and target directions via cosine similarity, ensuring accurate orientation

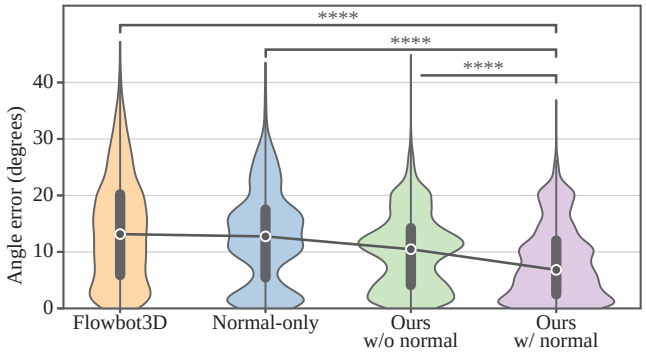


Fig. 7: Quantitative results of direction estimation on unseen object categories. Prediction errors from four methods over 5,836 test samples drawn from categories not seen during training. Vi-TacMan, which uses surface normals as an inductive bias, achieves significant performance gains over baselines. The violin plots show error distributions: the outer shape is the kernel density estimate (KDE); the white dot is the median; the thick bar denotes the interquartile range (IQR); and the whiskers extend to 1.5× IQR beyond the quartiles. Note: **** indicates $p < 0.0001$.

even when magnitudes are small. The model is trained using the AdamW optimizer on a single H100 GPU with a batch size of 32 and a learning rate of 1×10^{-3}. Inference is performed on an RTX 3090 GPU for all experiments.

E. Tactile Manipulation Policy

Following initialization by the estimated grasp and interaction direction, all subsequent manipulation is governed by a tactile-based policy. We employ a tactile controller to manipulate articulated objects, building on the work of Zhao *et al*. [14, 15], which utilizes a GelSight-style tactile sensor [30] to provide contact feedback. This approach extracts tactile features from the positions of activated markers—defined as those whose normal deformation exceeds a predefined threshold. By tracking marker-wise position changes, the controller computes pose updates (Eq. (1)) via a point registration algorithm operating at 50 Hz. For complete algorithmic details, we refer readers to the original work due to space constraints.

While GelSight-style sensors are widely adopted and their mechanical design and calibration are well documented [30–33], fabrication of the core component—the elastomer with Lambertian coating—appears to remain lab-specific. To improve reproducibility, we detail one practical fabrication procedure used in this study in Fig. 6. The airbrushable silicone pigment is prepared by mixing silicone pigment with Smooth-On Psycho Paint (a platinum-silicone paint base) and thinning the mixture using Smooth-On NOVOCS Matte solvent. This enables uniform spray application and consistent elastomer finishes suitable for tactile imaging.

IV. Experiments

This section evaluates Vi-TacMan through comprehensive experiments. We begin with large-scale tests on synthetic objects in Sec. IV-A to assess generalization across unseen categories. We then validate Vi-TacMan in the real world (Sec. IV-B), demonstrating the complete pipeline for manipulating unknown articulated objects via vision and touch.

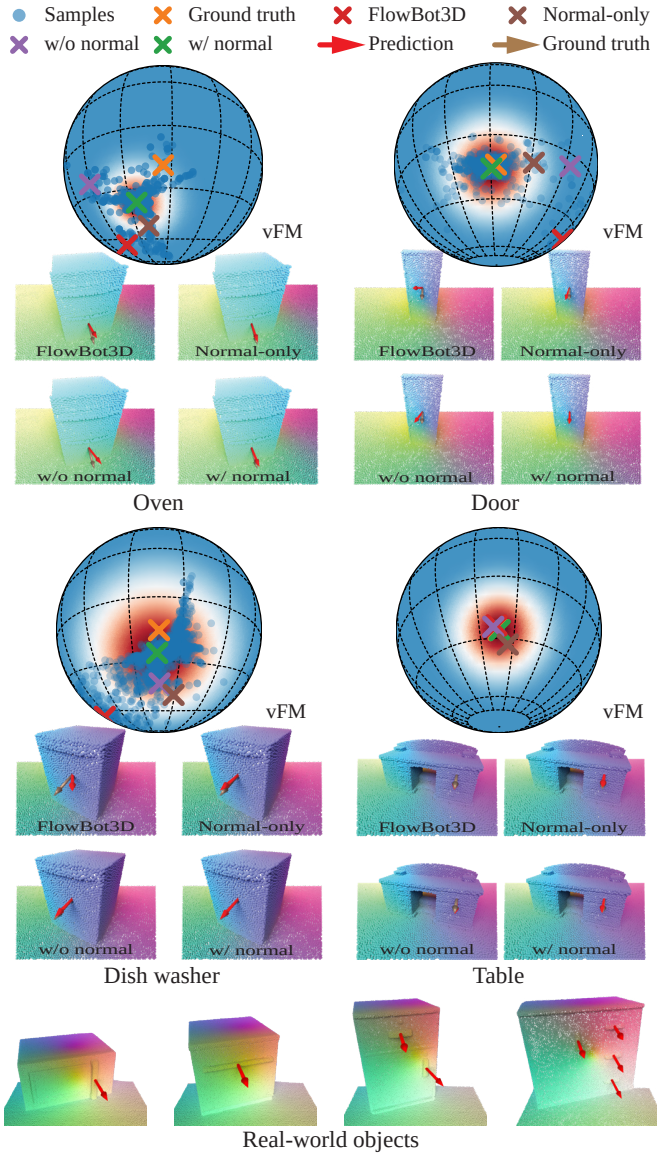


Fig. 8: Qualitative results of direction estimation on unseen object categories. We illustrate the approach using four representative objects, one from each test category. For each object, we show the obtained samples, the fitted vMF distribution, the ground truth, and predictions from the three baseline methods. By fitting the distribution and incorporating surface normals as an inductive bias, Vi-TacMan demonstrates greater robustness to high uncertainty when encountering previously unseen objects. The bottom row presents results on real-world examples using the grasping points shown in Fig. 4(d), demonstrating successful transfer from simulation to real-world settings.

A. Simulation Studies

We evaluate interaction-direction estimation using 5,836 test samples from categories unseen during training, as introduced in Sec. III-A. This setup allows us to assess generalization to previously unknown articulated objects. We compare Vi-TacMan, which leverages surface normals as an important inductive bias, against three baselines:

- **FlowBot3D:** A recent method for articulated object manipulation that employs point-displacement modeling similar to ours [34]. It selects the interaction direction that maximizes articulation movement without modeling the full direction distribution.
- **Normal-only:** A simple, learning-free baseline that computes the Fréchet mean (see Eq. (10)) of surface normals within the moving region.
- **Without-normal:** An ablation that trains our model without surface-normal inputs while keeping all other components unchanged, isolating the contribution of this feature.

For fair comparison, we set the grasping point as the movable region's centroid for both the Vi-TacMan and the Without-normal baseline. For FlowBot3D and Normal-only, we translate predictions to grasping points for comparison.

Quantitative results are shown in Fig. 7, where prediction

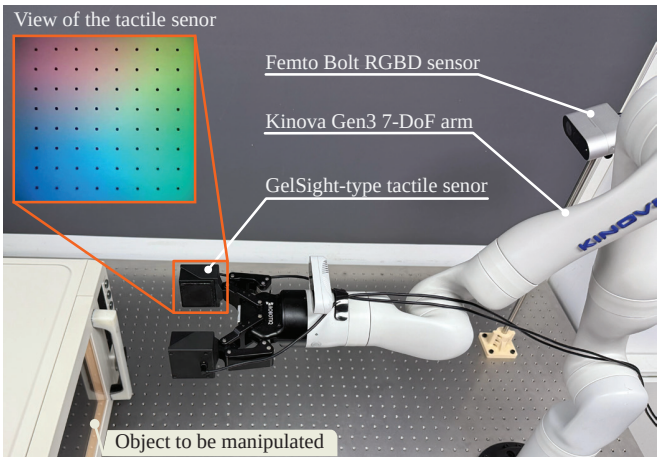


Fig. 9: Experimental platform for real-world validation.

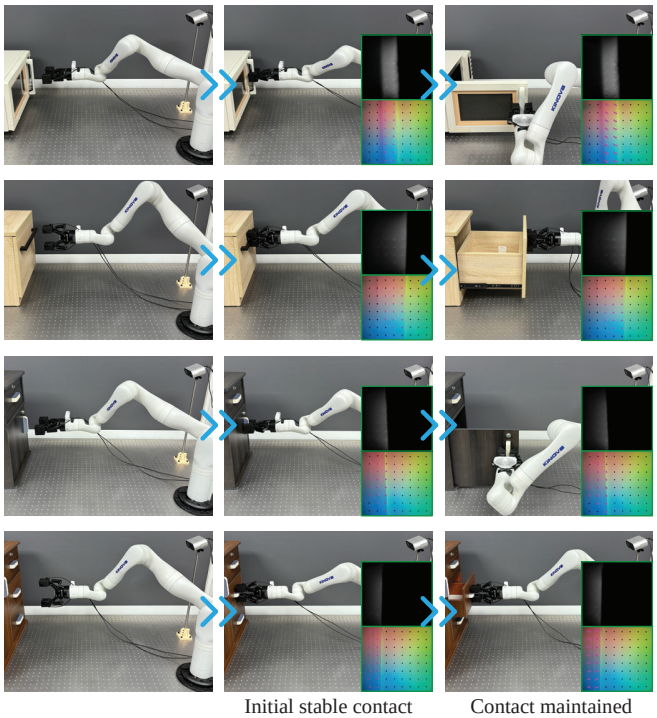


Fig. 10: Real-world validation of Vi-TacMan. Leveraging visual cues, the robot automatically establishes stable contact with the handle of the articulated object. Following the estimated interaction direction, the low-level, tactile-informed controller reliably completes the manipulation.

error is measured as the angle between predicted and ground-truth directions. All four methods achieve median errors around 10°, highlighting the challenge of recovering precise motion directions on unfamiliar geometries. FlowBot3D, without modeling the distribution of point displacements, shows greater sensitivity to unseen categories. The Normal-only baseline, despite its simplicity, achieves competitive performance. Vi-TacMan reduces uncertainty by modeling the distribution of fitted normals, and explicitly incorporating surface normals further improves performance. One-sided paired t-tests confirm statistically significant improvements over all three baselines ($p < 0.0001$).

For qualitative illustration, Fig. 8 shows four representative objects from the test categories, visualizing sample directions alongside the corresponding fitted vMF distributions.

B. Real-World Experiments

To assess the gap between synthetic objects and real-world scenarios, we evaluate our model on physical objects captured in the real world, as shown in Fig. 4(a). Using the selected grasping points in Fig. 4(d), we present four representative examples in Fig. 8, demonstrating that Vi-TacMan generates plausible interaction direction estimates.

To further assess whether visual cues alone can drive complete manipulation of articulated objects, we implement the full pipeline in the real world, from vision-based high-level guidance to tactile-informed low-level control. We use a Kinova Gen3 7-DoF arm equipped with GelSight-type tactile sensors in place of its default gripper pads, as described in Sec. III-E. The integrated system is illustrated in Fig. 9.

As shown in Fig. 10, Vi-TacMan guides the robot to reliably establish valid grasps on real objects and follow the estimated interaction direction. By leveraging tactile feedback, the system adapts its motions in real time (50 Hz), achieving consistent and robust manipulation across all articulated objects. The complete manipulation process and additional experimental results are provided in the supplementary materials and on our website.

V. Conclusion and Future Work

We introduced Vi-TacMan, a framework for articulated object manipulation that leverages the complementary strengths of vision and touch. Rather than inferring precise but unreliable kinematics from vision alone, Vi-TacMan uses vision for coarse guidance—grasp proposals and interaction direction estimates—while relying on tactile feedback for robust execution. By incorporating surface normals as a geometric prior and modeling interaction directions with a vMF distribution, Vi-TacMan generalizes to unseen objects and outperforms existing baselines. Our evaluations demonstrate that Vi-TacMan enables autonomous manipulation of diverse articulated objects without explicit kinematic models, highlighting the value of integrating visual guidance with tactile control.

Interpretability through hierarchical design: A key advantage of Vi-TacMan's hierarchical architecture—which separates visual intention from tactile execution—is its inherent interpretability. Unlike end-to-end policies that map pixels directly to actions, our system explicitly generates a coarse interaction direction d and grasp G before contact is made. This intermediate representation serves as a communicable "intention" that could be exposed to human users in future iterations. For instance, augmented reality projections of the intended trajectory or verbal announcements (*e.g.*, "Opening cabinet") prior to execution would foster safer, more predictable human-robot interaction and simplify debugging for safety certification in unstructured domestic environments.

Extending to non-rigid and multi-modal scenarios: Our current formulation leverages the Kabsch algorithm under a strict rigid-body assumption to estimate displacement. However, domestic objects often exhibit compliance or multi-stage articulation (*e.g.*, flexible handles or nested joints). Future work will explore extending our displacement estimation to handle non-rigid deformations, potentially through deformable object tracking or sequential state estimation. Additionally, while our vMF-based modeling captures directional uncertainty, objects with multiple distinct valid interaction directions (*e.g.*, a lever that can toggle both up and down) may require multi-modal distribution modeling or conditioning on high-level user commands.

Handling grasp failures: Our method currently assumes the initial grasp remains stable throughout manipulation. To mitigate failures from grasp slippage, we plan to integrate tactile slip detection with dynamic re-grasping primitives, enabling the system to recover from execution errors and improve overall robustness.

References

[1] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, *et al.*, "Sapien: A simulated part-based interactive environment," in *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[2] L. Liu, W. Xu, H. Fu, S. Qian, Q. Yu, Y. Han, and C. Lu, "Akb-48: A real-world articulated object knowledge base," in *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[3] W. Wang, Z. Zhao, Z. Jiao, Y. Zhu, S.-C. Zhu, and H. Liu, "Rearrange indoor scenes for human-robot co-activity," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[4] Z. Jin, Z. Che, Z. Zhao, K. Wu, Y. Zhang, Y. Zhao, Z. Liu, Q. Zhang, X. Ju, J. Tian, *et al.*, "Artvip: Articulated digital assets of visual realism, modular interaction, and physical fidelity for robot learning," *arXiv preprint arXiv:2506.04941*, 2025.

[5] C. C. Kemp, A. Edsinger, and E. Torres-Jara, "Challenges for robot manipulation in human environments [grand challenges of robotics]," *IEEE Robotics and Automation Magazine (RA-M)*, vol. 14, no. 1, pp. 20–29, 2007.

[6] K. Mo, L. Guibas, M. Mukadam, A. Gupta, and S. Tulsiani, "Where2act: From pixels to actions for articulated 3d objects," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2021.

[7] A. Jain, R. Lioutikov, C. Chuck, and S. Niekum, "Screwnet: Category-independent articulation model estimation from depth images using screw theory," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[8] Z. Zeng, T.-E. Lee, J. Liang, and O. Kroemer, "Visual identification of articulated object parts," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021.

[9] B. Eisner, H. Zhang, and D. Held, "FlowBot3D: Learning 3D articulation flow to manipulate articulated objects," in *Robotics: Science and Systems (RSS)*, 2022.

[10] M. Mittal, D. Hoeller, F. Farshidian, M. Hutter, and A. Garg, "Articulated object interaction in unknown scenes with whole-body mobile manipulation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022.

[11] Q. Yu, J. Wang, W. Liu, C. Hao, L. Liu, L. Shao, W. Wang, and C. Lu, "Gamma: Generalizable articulation modeling and manipulation for articulated objects," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[12] J. Wang, W. Liu, Q. Yu, Y. You, L. Liu, W. Wang, and C. Lu, "Rpmart: Towards robust perception and manipulation for articulated objects," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.

[13] Y. Wang, X. Zhang, R. Wu, Y. Li, Y. Shen, M. Wu, Z. He, Y. Wang, and H. Dong, "Adamanip: Adaptive articulated object manipulation environments and policy learning," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2025.

[14] Z. Zhao, Y. Li, W. Li, Z. Qi, L. Ruan, Y. Zhu, and K. Althoefer, "Tac-Man: Tactile-informed prior-free manipulation of articulated objects," *IEEE Transactions on Robotics (T-RO)*, vol. 41, pp. 538–557, 2025.

[15] Z. Zhao, Z. Qi, Y. Li, L. Cui, Z. Han, L. Ruan, and Y. Zhu, "Tacman-turbo: Proactive tactile control for robust and efficient articulated object manipulation," *arXiv preprint arXiv:2508.02204*, 2025.

[16] K. V. Mardia and P. E. Jupp, *Directional statistics*. John Wiley & Sons, 2009.

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2014.

[18] W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, vol. 32, no. 5, pp. 922–923, 1976.

[19] H. Geng, H. Xu, C. Zhao, C. Xu, L. Yi, S. Huang, and H. Wang, "Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts," in *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[20] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[21] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, *et al.*, "Dinov3," *arXiv preprint arXiv:2508.10104*, 2025.

[22] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2020.

[23] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, *et al.*, "Sam 2: Segment anything in images and videos," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2025.

[24] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg, "Learning ambidextrous robot grasping policies," *Science Robotics*, vol. 4, no. 26, p. eaau4984, 2019.

[25] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[26] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE Transactions on Robotics (T-RO)*, vol. 39, no. 5, pp. 3929–3945, 2023.

[27] A. Ten Pas, M. Gualtieri, K. Saenko, and R. Platt, "Grasp pose detection in point clouds," *International Journal of Robotics Research (IJRR)*, vol. 36, no. 13-14, pp. 1455–1473, 2017.

[28] Z. Zhao, L. Cui, S. Xie, S. Zhang, Z. Han, L. Ruan, and Y. Zhu, "B*: Efficient and optimal base placement for fixed-base manipulators," *IEEE Robotics and Automation Letters (RA-L)*, vol. 10, no. 10, pp. 10634–10641, 2025.

[29] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[30] 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.

[31] W. Li, Z. Zhao, L. Cui, W. Zhang, H. Liu, L.-A. Li, and Y. Zhu, "Minitac: An ultra-compact 8 mm vision-based tactile sensor for enhanced palpation in robot-assisted minimally invasive surgery," *IEEE Robotics and Automation Letters (RA-L)*, vol. 9, no. 12, pp. 11170–11177, 2024.

[32] Z. Zhao, W. Li, Y. Li, T. Liu, B. Li, M. Wang, K. Du, H. Liu, Y. Zhu, Q. Wang, *et al.*, "Embedding high-resolution touch across robotic hands enables adaptive human-like grasping," *Nature Machine Intelligence*, vol. 7, no. 6, pp. 889–900, 2025.

[33] Y. Li, W. Du, C. Yu, P. Li, Z. Zhao, T. Liu, C. Jiang, Y. Zhu, and S. Huang, "Taccel: Scaling up vision-based tactile robotics via high-performance gpu simulation," in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[34] B. Ebner, A. Fischer, R. E. Gaunt, B. Picker, and Y. Swan, "Stein's method of moments," *Scandinavian Journal of Statistics*, 2025.