2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

M-VTOP: Modular Visuo-Tactile Object Pose Estimation for High-Precision Robotic Manipulation

Miquel Oller*[1,2], Qiyang Qian*[1,3], Radu Corcodel[1], and Siddarth Jain[1]

***Abstract*— Accurate object pose estimation is essential for robotic manipulation, particularly in tasks involving small or geometrically intricate objects where high precision is required. Existing vision, tactile, and hybrid-based approaches struggle with occlusion, noise, and limited generalization, often requiring extensive retraining or large annotated datasets. In this work, we present M-VTOP, a modular framework for in-hand object pose estimation that integrates vision, tactile, and contact sensing in a flexible manner, allowing robustness against noisy or missing modalities. At the core of the framework is a belief-based particle filter that fuses heterogeneous sensor observations, maintains probabilistic estimates, and continuously refines them toward high-precision convergence in closed-loop robotic control with the pose estimation feedback. A mask-based observation representation unifies visual and tactile signals into geometry-centric inputs, enhancing robustness to texture and lighting variations while supporting zero-shot generalization. The framework requires only an object's CAD model and avoids task-specific retraining. Experiments show that M-VTOP achieves sub-millimeter accuracy under complex geometries, occlusions, and tight tolerances, demonstrating its promise for high-precision robotic manipulation.**

I. INTRODUCTION

Accurate object pose estimation is essential for robotic manipulation, as it forms the basis for reliable execution of tasks such as assembly, alignment, and insertion. In both industrial and service robotics, the capability to precisely perceive and localize objects directly impacts the success of subsequent operations. This challenge is especially pronounced for small or geometrically complex objects, such as electrical connectors, where even minor pose errors can lead to misalignment, failed insertion, or component damage. Consequently, despite substantial progress, current methods for object pose estimation often struggle to meet the demands of real-world applications. Vision-based approaches [1]–[4] are vulnerable to occlusion and resolution limits. Tactile sensing methods [5]–[7] provide accurate local estimates but are prone to noise and geometric ambiguities; with contact-only information, capturing the object geometry or orientation becomes difficult. Hybrid visuo-tactile approaches [8]–[10] mitigate some limitations but typically demand extensive training, parameter tuning, and offer limited adaptability.

This paper introduces M-VTOP, a modular framework for in-hand object pose estimation, designed to meet the stringent requirements of small-object manipulation. The framework integrates observations from vision, tactile, and contact sensing modalities, with the flexibility to accommodate a variable number of inputs. Such modularity enhances robustness: when one modality is degraded by noise or occlusion, others can compensate. The central component is a belief-based particle filter that iteratively refines pose estimates by comparing simulated and real sensor observations. This probabilistic formulation mitigates overconfidence in unreliable signals, and facilitates precise convergence. Notably, the method requires only the object's CAD model and can operate in a zero-shot fashion, thereby eliminating the need for retraining or task-specific adaptation. A key innovation is the mask-based observation representation, which integrates visual and tactile inputs into geometry-focused signals. By filtering out surface texture and lighting variations, this approach enhances robustness and enables generalization. The primary contributions of this work are as follows:

- A modular, multi-sensor framework for in-hand object pose estimation that flexibly integrates heterogeneous modalities while maintaining robustness.
- A belief-based particle filter that fuses diverse sensor inputs, explicitly accounts for uncertainty, and achieves sub-millimeter accuracy without retraining.
- Experimental validation under challenging conditions: small objects, occlusions, and complex geometries.

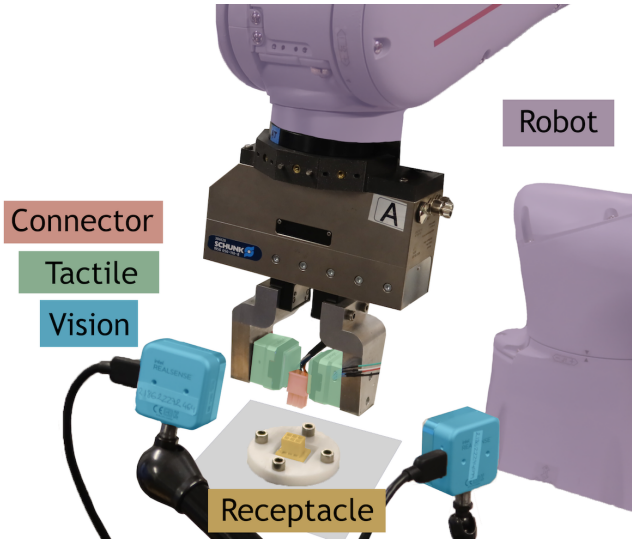

Fig. 1: Robotic setup for precise insertion tasks. The insertion piece (red) initially grasped by the robot in an unknown pose. Using both vision and tactile observations, the robot must estimate the insertion object's position and orientation, then align and guide it to fit snugly into the receptacle (yellow).

[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA. sjain@merl.com
[2]University of Michigan, Ann Arbor, MI, USA.
[3]University of California, Berkeley, CA, USA.
*Work done during the MERL internships of M. Oller and Q. Qian.

II. RELATED WORK

1) Vision-Based Pose Estimation: Many learning-based approaches estimate pose directly from images through regression [3], [11], [12]. In contrast, template-based methods align observations with synthetic or CAD-derived templates [2], [13]–[15]. More recent work incorporates foundation models to improve feature matching [14], frequently combined with iterative pose refinement to reduce rendering overhead [1], [16]–[18]. Despite these advances, both paradigms are typically instance-specific, requiring retraining for new objects, and they often struggle with small objects.

2) Tactile-Based Pose Estimation: Tactile pose estimation is often learning-based. For example, Tac2Pose [5] learns latent tactile representations from synthetic contact data, but such approaches typically require retraining and provide limited resolution. Physics-based methods instead model sensor deformation [7], though they are often computationally expensive. Other approaches reconstruct contact geometries and align them with known models [19]–[21]; however, the local nature of tactile observations can lead to ambiguity, particularly for large objects [22]. To mitigate aliasing, recent work aggregates tactile observations over time [23], [24].

3) Visuo-Tactile Pose Estimation: Joint visuo-tactile approaches leverage complementary sensing cues. Optimization-based methods estimate pose by minimizing energy functions [8], while learning-based techniques reconstruct occluded regions and improve robustness [9], [10], [25], [26]. VT2Pose [27] jointly fuses visual and tactile features; however, its reliance on end-effector cameras limits applicability in industrial environments that typically use external cameras, where pose invariance is not preserved. More broadly, many learning-based fusion methods struggle to generalize to unseen geometries without retraining. In contrast, our method is zero-shot, sensor-agnostic, and does not assume fixed camera placement. It can operate flexibly with vision only, touch only, or both modalities.

4) Contact-Based Pose Estimation: Contact signals, often obtained from wrench measurements, have been used to localize contact points [28] or reduce in-hand pose ambiguity [29], [30], but rely on highly accurate force sensing. Integrations with external vision [31] require fixed, collision-free poses. Our method differs by treating contact as a binary event with a smooth signed-distance likelihood, improving robustness to noise and compliance. Contact events also trigger recovery and reattempt strategies, allowing updated observations to refine pose estimates during tasks.

III. PROBLEM FORMULATION

In this work, we aim to estimate the 6D pose of a grasped object from vision and tactile observations. Our method assumes access to the object's CAD model, along with vision, tactile, and collision information. Additionally, we assume knowledge of the sensor parameters, i.e. intrinsics and extrinsics, as well as the environment's geometry. Our method is designed to accommodate an arbitrary number of vision and tactile sensors. We denote $N_{\mathrm{T}} \geq 0$ as the number of tactile sensors and $N_{\mathrm{V}} \geq 0$ as the number of vision sensors. The objective is to estimate the object's pose $\mathbf{x} \in \mathrm{SE}(3)$ given a set of tactile observations $\mathcal{T} = \{T_i\}_{i=1}^{N_{\mathrm{T}}}$ and vision observations $\mathcal{V} = \{V_i\}_{i=1}^{N_{\mathrm{V}}}$ where $T_i \in \mathbb{R}^{w_{ti} \times h_{ti} \times 3}$ and $V_i \in \mathbb{R}^{w_{vi} \times h_{vi} \times 3}$. Additionally, we incorporate contact information through a binary variable $c \in \{0, 1\}$, which indicates whether external contact is detected. We further assume that the object is rigid and that external collisions occur only between the grasped object and the environment.

IV. METHOD

This section presents methodology for M-VTOP that fuses vision, touch, and contact data for robust object pose estimation using a pose particle filter (PPF). Figure 2 shows the pipeline. The object pose uncertainty is represented by pose particles $\mathcal{X} = \{\mathbf{x}^{(n)} \mid \mathbf{x}^{(n)} \in \mathrm{SE}(3), n = 1, \ldots, N\}$. Observations are aggregated via a model that scores vision and tactile data using a render-and-compare strategy over segmented object masks generated by foundation models.

The subsections detail each step: Section IV-A covers mask generation; Section IV-B describes visuo-tactile matching; Section IV-C explains the PPF for pose belief estimation; and Section IV-D presents the motion model for updating the pose under robot actions and interaction uncertainty.

A. Visuo-Tactile Masking

The Visuo-Tactile Masking module identifies object pixels from each sensor observation, enabling grasped object segmentation for posterior pose estimation. Given a set of tactile observations $\mathcal{T} = \{T_i\}_{i=1}^{N_{\mathrm{T}}}$ and vision observations $\mathcal{V} = \{V_i\}_{i=1}^{N_{\mathrm{V}}}$, the module estimates the grasped object's masks for each of the sensor observations, i.e., $\mathcal{M}_T = \{M_{T,i}\}_{i=1}^{N_T}$ and $\mathcal{M}_V = \{M_{V,i}\}_{i=1}^{N_V}$. In our approach, tactile observations are treated as RGB images. We introduce a plug-and-play, backend-agnostic module that uses zero-shot foundation models for generalization and supervised models for higher precision when task-specific annotations are available. The downstream pipeline is unchanged regardless of backend. Similar to [32], the module produces candidate masks (zero-shot) or a single mask (supervised). We integrate visual and tactile observations to enforce cross-modal consistency, improving robustness against partial or inaccurate masks.

Algorithm 1 outlines our approach for mask extraction for zero-shot backends. For each tactile and vision observation T_i and V_i, under zero-shot settings, we obtain the mask proposals $\tilde{\mathcal{M}}_{T,i}$ and $\tilde{\mathcal{M}}_{V,i}$. We collect all mask proposals as $\bar{\mathcal{M}}_T = \bigcup_{i=1}^{N_T} \tilde{\mathcal{M}}_{T,i}$ and $\bar{\mathcal{M}}_V = \bigcup_{i=1}^{N_V} \tilde{\mathcal{M}}_{V,i}$. Since candidates may include extraneous scene elements or incomplete object masks, we select masks by leveraging object geometry and sensor information. To achieve this, we compare the generated mask proposals against rendered reference masks from a coarse set of object poses. Specifically, we uniformly sample pose candidates $\mathcal{X}$ and render the reference masks $\hat{\mathcal{M}}_{T,n}$ and $\hat{\mathcal{M}}_{V,n}$ for each $\boldsymbol{x}_n \in \mathcal{X}$. We then compute a matching score s_n by measuring Intersection over Union (IoU) between the rendered references and the mask proposals as defined in Algorithm 2 and illustrated in Fig. 3.

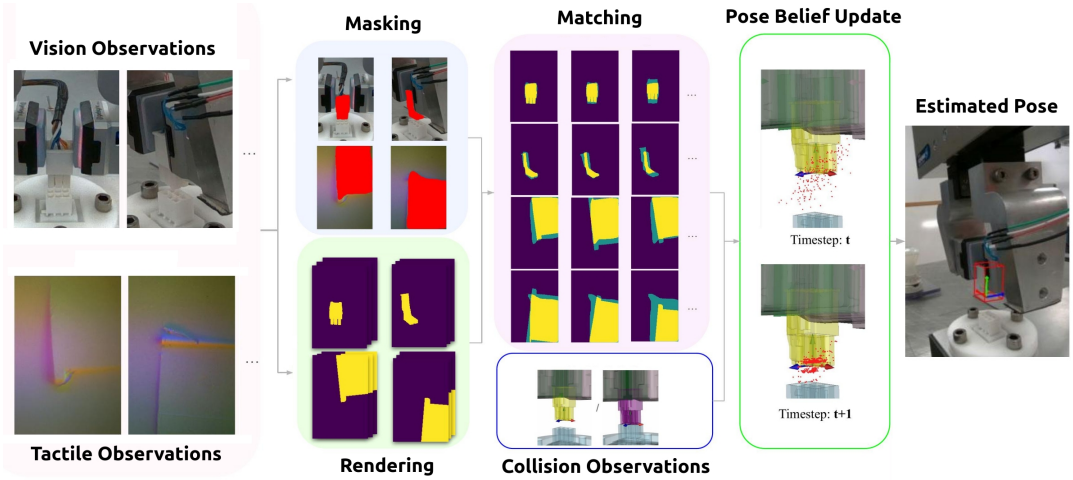

Fig. 2: **Method Overview:** Our method takes as input the visuo-tactile observations and the object CAD model. The segmentation backend produces masks $\mathcal{M}_T$ and $\mathcal{M}_V$. For each pose particle in $\mathcal{X}_t$ at time step $\mathbf{t}$, the rendering module generates binary masks $\hat{\mathcal{M}}_T^{(n)}$, $\hat{\mathcal{M}}_V^{(n)}$, which are scored against observation masks. These scores update particles $\mathcal{X}_{t+1}$.

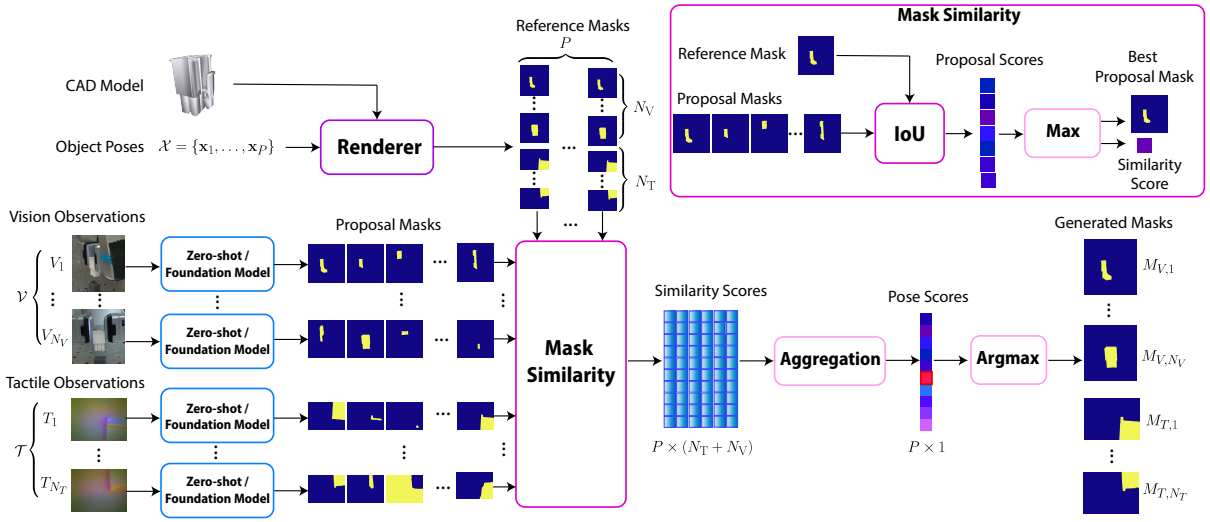

Fig. 3: **Zero-shot Visuo-tactile masking.** Zero-shot backend (SAM2/Mobile-SAM) generates mask proposals for vision and tactile observations that are matched with CAD-rendered masks via IoU; scores are aggregated to select the best masks.

For each pose, we select the best-fitting masks per observation and sum their IoUs to obtain the pose score. These scores are aggregated across the sensor modalities to compute a pose-wise score. Finally, we select the optimal pose $\boldsymbol{x}^*$ with the highest aggregated score s^*. We then take the corresponding masks $\mathcal{M}_T$ and $\mathcal{M}_V$ for that pose, ensuring consistency across both modalities.

While this method ensures accurate segmentation without requiring object pose priors, it is computationally demanding. For efficiency, we introduce an alternative strategy when pose uncertainty is low (i.e., when particle convergence is achieved). Instead of generating and comparing multiple mask candidates, promptable zero-shot backends will switch to query-point prompts using the object CAD model to reduce computational overhead. We also find that zero-shot backends often struggle with tactile observations of objects that contain fine details. To address this, we refine the tactile masks using depth information, removing undeformed (contact-free) regions prior to binarization.

B. Visuo-Tactile Matching

The visuo-tactile matching module assigns a score $s^{(n)}$ to each pose proposal $\boldsymbol{x}^{(n)}$, quantifying its alignment with the observed tactile masks $\mathcal{M}_T$ and vision masks $\mathcal{M}_V$. We compute the score as a weighted sum of IoUs between observed masks and simulated masks $\hat{\mathcal{M}}_T^{(n)}$ and $\hat{\mathcal{M}}_V^{(n)}$, which correspond to the expected observations for pose $\boldsymbol{x}^{(n)}$:

$$s^{(n)} = \frac{e^{\frac{\tilde{s}_n}{\tau}}}{\sum_{n=1}^{N} e^{\frac{\tilde{s}_n}{\tau}}} \tag{1}$$

$$\tilde{s}_n = \sum_{i=1}^{N_T} \gamma_i \ \mathrm{IoU}(M_{T,i}, \hat{M}_{T,i}^{(n)}) + \sum_{j=1}^{N_V} \nu_j \ \mathrm{IoU}(M_{V,j}, \hat{M}_{V,j}^{(n)}) \tag{2}$$

where τ is the temperature factor and γ_i and ν_i are importance weights such that $0 \leq \gamma_i, \nu_j \leq 1$, $\sum_i \gamma_i +$

Algorithm 1: Visuo-Tactile Masking

Data: $\mathcal{T}_k, \mathcal{V}_k$
Result: $\mathcal{M}_T, \mathcal{M}_V$
$\bar{\mathcal{M}}_T \leftarrow \emptyset$;
$\bar{\mathcal{M}}_V \leftarrow \emptyset$;
$\mathcal{X} \leftarrow$ InitPoseCandidates();
▷ Obtain mask proposals;
for $T_i \in \mathcal{T}$ **do**
  $\tilde{\mathcal{M}}_{T,i} \leftarrow$ GenerateMaskProposals(T_i);
  $\bar{\mathcal{M}}_T \leftarrow \bar{\mathcal{M}}_T + \langle \tilde{\mathcal{M}}_{T,i} \rangle$;
for $V_i \in \mathcal{V}$ **do**
  $\tilde{\mathcal{M}}_{V,i} \leftarrow$ GenerateMaskProposals(V_i);
  $\bar{\mathcal{M}}_V \leftarrow \bar{\mathcal{M}}_V + \langle \tilde{\mathcal{M}}_{V,i} \rangle$;
▷ Score and select mask proposals $s^* \leftarrow 0$;
for $\boldsymbol{x}_n \in \mathcal{X}$ **do**
  $\hat{\mathcal{M}}_{T,n}, \hat{\mathcal{M}}_{V,n} \leftarrow$ Render($\boldsymbol{x}_n$);
  $s_n, \mathcal{M}_T^\star, \mathcal{M}_V^\star \leftarrow$ MaskSimilarity($\hat{\mathcal{M}}_{T,n}, \hat{\mathcal{M}}_{V,n}, \bar{\mathcal{M}}_T, \bar{\mathcal{M}}_V$);
  if $s_n > s^*$ **then**
    $\mathcal{M}_T \leftarrow \mathcal{M}_T^\star$;
    $\mathcal{M}_V \leftarrow \mathcal{M}_V^\star$;
    $s^* \leftarrow s_n$;

Algorithm 2: Score Best Masks

Data: $\hat{\mathcal{M}}_T, \hat{\mathcal{M}}_V, \bar{\mathcal{M}}_T, \bar{\mathcal{M}}_V$
Result: $s, \mathcal{M}_T^\star, \mathcal{M}_V^\star$
$s \leftarrow 0$;
$\mathcal{M}_T^\star \leftarrow \emptyset$;
$\mathcal{M}_V^\star \leftarrow \emptyset$;
▷ Score tactile mask proposals
for $\hat{M}_T \in \hat{\mathcal{M}}_T$ **do**
  $s_T^* \leftarrow 0$;
  for $\tilde{M}_T \in \bar{\mathcal{M}}_T$ **do**
    $s_i \leftarrow \text{IoU}(\hat{M}_T, \tilde{M}_T)$;
    if $s_i \geq s_T^*$ **then**
      $\tilde{M}_T^* \leftarrow \tilde{M}_T$;
      $s_T^* \leftarrow s_i$;
  $s \leftarrow s + s_T^*$;
  $\mathcal{M}_T^\star \leftarrow \mathcal{M}_T^\star + \langle \tilde{M}_T^* \rangle$;
▷ Score vision mask proposals;
for $\hat{M}_V \in \hat{\mathcal{M}}_V$ **do**
  $s_V^* \leftarrow 0$;
  for $\tilde{M}_V \in \bar{\mathcal{M}}_V$ **do**
    $s_i \leftarrow \text{IoU}(\hat{M}_V, \tilde{M}_V)$;
    if $s_i \geq s_V^*$ **then**
      $\tilde{M}_V^* \leftarrow \tilde{M}_V$;
      $s_V^* \leftarrow s_i$;
  $s \leftarrow s + s_V^*$;
  $\mathcal{M}_V^\star \leftarrow \mathcal{M}_V^\star + \langle \tilde{M}_V^* \rangle$;

$\sum_j \nu_j = 1$. In our experiments, we equally weight the different sensor modalities by setting $\gamma_i = \frac{1}{N_T}$ and $\nu_j = \frac{1}{N_V}$.

The synthetic mask observations $\hat{\mathcal{M}}_T^{(n)}$ and $\hat{\mathcal{M}}_V^{(n)}$ are obtained by rendering the expected tactile and vision observations for a given pose $\boldsymbol{x}^{(n)}$ using their respective camera models. For this work, we employ `pyrender` [33] to synthesize realistic sensor measurements. To improve mask accuracy, we explicitly include robot and environment geometry so that occlusions are rendered correctly.

Algorithm 3: Pose Particle Filter

Data: $\mathcal{X}_{k-1}, \mathcal{T}_k, \mathcal{V}_k, c_k, u_k$
Result: $\mathcal{X}_k = \{\mathbf{x}_k^{(1)}, \ldots, \mathbf{x}_k^{(N)}\}, \quad \mathbf{x}_k^{(n)} \in SE(3)$
if $\mathcal{X}_{k-1} = \emptyset$ **then**
  $\mathcal{X}_k^{'} \leftarrow \mathcal{X}_{\text{init}}$;
else
  $\mathcal{X}_k^{'} \leftarrow$ MotionModel($\mathcal{X}_{k-1}, u_k, c_k$) ;
$\bar{\mathcal{X}} \leftarrow \emptyset$;
for $\mathbf{x}_k^{(n)} \in \mathcal{X}_k^{'}$ **do**
  $w_k^{(n)} \leftarrow p(\mathcal{V}_k, \mathcal{T}_k | \mathbf{x}_k^{(n)}) p(c_k | \mathbf{x}_k^{(n)})$;
  $\bar{\mathcal{X}} \leftarrow \bar{\mathcal{X}} + \langle \mathbf{x}_k^{(n)}, w_k^{(n)} \rangle$;
$\mathcal{X}_k =$ ImportanceResample($\bar{\mathcal{X}}$);

C. Pose Particle Filter

Our goal is to estimate the probabilistic belief of an object's pose across a sequence of observations and robot motions. Using a particle filter, we integrate vision, tactile, and contact data over time while maintaining a distribution over possible poses (see Algorithm 3). Given these observations and motions, the pose belief is expressed as:

$$\text{bel}(\boldsymbol{x}_k) = p(\mathbf{x}_k | \mathcal{V}_{1:k}, \mathcal{T}_{1:k}, c_{1:k}, u_{1:k}) \tag{3}$$

We approximate this distribution with a set of particles:

$$p(\mathbf{x}_k | \mathcal{V}_{1:k}, \mathcal{T}_{1:k}, c_{1:k}, u_{1:k}) \approx \sum_{n=1}^{N} w_k^{(n)} \delta(\mathbf{x}_k^{(n)}) \tag{4}$$

Assuming independence between visuo-tactile and contact observations, the posterior belief can be decomposed as:

$$p(\mathbf{x}_k | \mathcal{V}_{1:k}, \mathcal{T}_{1:k}, c_{1:k}) = \eta p(\mathcal{V}, \mathcal{T} | \mathbf{x}_k) p(c | \mathbf{x}_k) \tag{5}$$

This yields particle weights of the form:

$$w_k^{(n)} = p(\mathcal{V}_k, \mathcal{T}_k | \mathbf{x}_k^{(n)}) p(c_k | \mathbf{x}_k^{(n)}) \tag{6}$$

Each weight comprises two independent terms: a visuo-tactile likelihood and a contact likelihood. The visuo-tactile likelihood is taken directly from the matching score between observed and rendered masks for pose

$$p(\mathcal{V}_k, \mathcal{T}_k | \mathbf{x}_k^{(n)}) = s_k^{(n)} \tag{7}$$

When we observe contact between connector and the receptacle, we update the pose particles using a soft likelihood

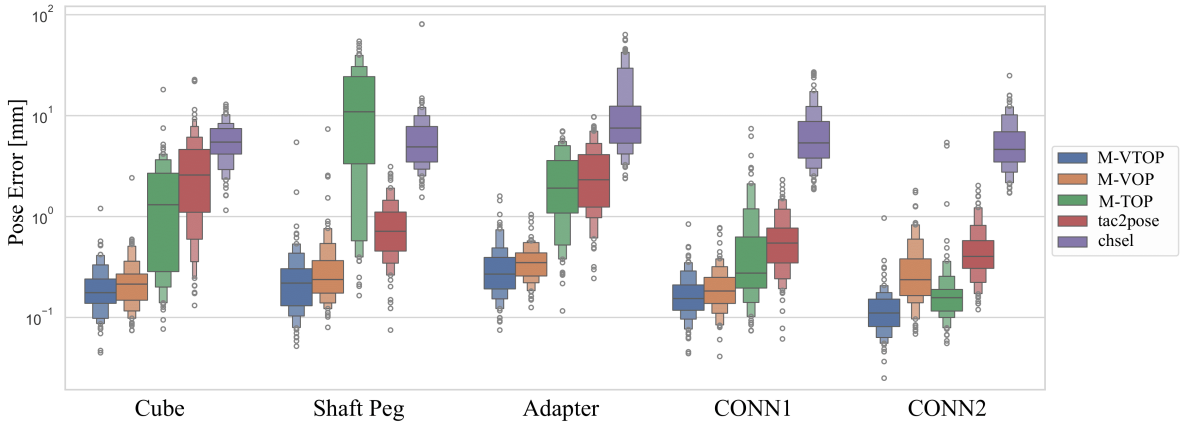

Fig. 4: **Simulation Method Benchmark:** We compare M-VTOP (ours, blue) with ablations M-VOP (vision-only, orange) and M-TOP (tactile-only, green), plus baselines Tac2Pose (tactile-only, red) and Chsel (vision + touch, purple). Y-axis: log-scale pose error (mean bounding-box distance to ground truth). X-axis: object type. Each object is tested on 100 poses.

based on signed distance between the pose particles and the receptacle in the simulator:

$$\pi(\boldsymbol{x}) = \sigma\Big(\tfrac{\tau-\phi(\boldsymbol{x})}{\sigma_d}\Big), \tag{8}$$

$$p(c_k \mid \boldsymbol{x}_k^{(n)}) = z_k^{(n)} = \begin{cases} \pi(\boldsymbol{x}_k^{(n)}), & c_k = 1, \\ 1, & c_k = 0\ . \end{cases} \tag{9}$$

Here, $\phi(\boldsymbol{x})$ is the signed distance between the object at pose $\boldsymbol{x}$ and the environment ($\phi > 0$ means contact-free), $\sigma(\cdot)$ is the logistic, and σ_d, τ are tunable softness and margin parameters. Signed distances are computed with FCL [34].

A hard gate (e.g., $|\phi| \leq \varepsilon$) collapses particle weights near a zero-measure contact manifold and is brittle to noise, compliance, and minor model errors. The soft mapping $\phi \mapsto \pi$ preserves probability mass around contact, reduces weight variance, and prevents particle depletion while retaining the same downstream update $w_k^{(n)} \propto s_k^{(n)}\, z_k^{(n)}$.

Finally, given particles representing the pose belief $\mathcal{X}_k$, we compute the pose estimate $\hat{\boldsymbol{x}}_k$ as a weighted average:

$$\hat{\mathbf{x}}_k = \sum_{n=1}^{N} w_k^{(n)} \mathbf{x}_k^{(n)}$$

This estimate supports downstream tasks such as insertion or in-hand pose correction.

Algorithm 4: Motion Model

Data: $\mathcal{X}_{k-1}, u_k, c_k$
Result: $\mathcal{X}_k$

$\mathcal{X}_k \leftarrow \emptyset$;
for $\mathbf{x}_{k-1}^{(n)} \in \mathcal{X}_{k-1}$ **do**
  if $c_k = 1$ **then**
    $\sigma_n \leftarrow \sigma^H$
  else
    $\sigma_n \leftarrow \sigma^L$
  $\mathbf{x}_k^{(n)} \leftarrow \mathrm{H}(u_k)\mathbf{x}_{k-1} + \mathcal{N}(0, \sigma_n)$;
  $\mathcal{X}_k \leftarrow \mathcal{X}_k + \langle \mathbf{x}_k^{(n)} \rangle$;

D. Robot Motion Model

Our framework updates pose uncertainty during robot motions u_k, accounting for potential relative movements due to contact or robot interactions. This is especially relevant in constrained environments, such as lowering an object to complete an insertion. The process for applying robot motion and updating pose particles is outlined in Algorithm 4.

Given a motion u_k, each particle is transformed as if rigidly attached to the end-effector via the homogeneous transform $\boldsymbol{H}(u_k)$. To account for possible relative displacements, we add Gaussian noise with different variances depending on whether the robot experiences external collision ($c_k = 1$) or remains collision-free ($c_k = 0$). This approach assumes that the object remains relatively stable in the grasp when no external contact is detected but is more likely to shift upon environmental interaction. To capture this variability, we introduce two parameters, σ^H and σ^L, which regulate the uncertainty introduced in the motion update. These parameters can be adjusted based on external constraints, such as whether the object is attached to a cable, making it more prone to slipping or shifting within the grasp.

V. Experiments and Results

We evaluate our framework on autonomous assembly with diverse industrial insertion tasks using novel objects, in both simulation and real-world trials.

A. Sensor Modality Ablation

We compare our proposed method against several ablations and baseline approaches. Specifically, we evaluate:

- **Ablations**: *Vision-only* and *Tactile-only* configurations.
- **Baselines**: *Tac2Pose* (tactile), and *Chsel* (visuo-tactile).

We assess performance on five different objects with varying sizes and geometries (Figure 5). For each object, we render 100 different poses. The results are shown in Figure 4. Our method consistently outperforms all baselines and ablations, achieving sub-millimeter accuracy and demonstrating the effectiveness of combining vision and touch for robust shape and pose estimation. Notably, for objects like the

TABLE I: Success rate (%) of insertion across modalities with the proposed framework.

	Masking Backbone	Cube	Shaft Peg	PLUG Adapter	CONN1	CONN2	Avg.
M-VOP	SAM2	100	33	100	67	100	80
	MaskRCNN	100	100	100	100	67	**93**
M-TOP	SAM2	100	0	33	33	33	40
	MaskRCNN	67	100	0	67	33	**53**
M-VTOP	SAM2	100	100	100	100	67	93
	MaskRCNN	100	100	100	100	100	**100**

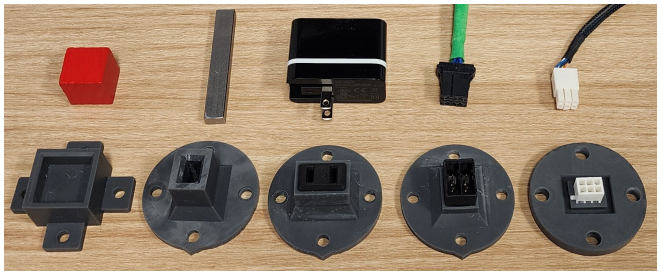

Fig. 5: Objects (top) and matching receptacles (bottom) used in the insertion experiments; each column is a pair. Left to right: **Wooden Cube** (25 mm) with receptacle; **Shaft Peg** (8 × 12 mm) with slot receptacle; **Power Plug Adapter** (NEMA 1–15P) with plug receptacle (NEMA 1–15R); 6-pin rectangular connector housing **CONN1** (3.81 mm pitch) with vertical header (3.81 mm); 6-pin mini connector housing **CONN2** (4.14 mm pitch) with vertical header (4.14 mm).

Shaft Peg, the tactile-only ablation suffers significantly due to partial observations and tactile geometric aliasing. Tac2Pose observations also suffer from this effect, but it is trained to minimize the predicted error which results into better predictions on average. Integrating vision resolves these issues by providing complementary information. Additionally, we observe that *Chsel*, which operates on point clouds, is affected by noisy depth estimations. To simulate real-world conditions, we introduce small depth noise during evaluation.

B. Real-World Experiments

In this section, we describe the real-world experiments conducted to validate our methodology, in which we focus to assess the generalization, modality ablation, spatial invariance, and robustness of our approach. The experimental setup is illustrated in Fig. 1. A MELFA Assista manipulator equipped with a Schunk WSG-50 parallel-jaw gripper serves as the primary platform. Two Intel RealSense cameras are mounted to capture front and side views of the workspace. To enable tactile sensing, a pair of GelSight Mini sensors are embedded in the gripper with their sensing surfaces aligned to the inner contact faces, providing per-finger tactile feedback from grasped objects.

Tactile depth is reconstructed from the RGB images of each GelSight sensor. Object contact deforms the elastomer, altering light transport through refraction and reflection, and thereby encoding surface geometry in color variations. A dense, high-resolution depth map is recovered by relating these appearance changes to surface normals and heights using stereophotometry combined with a deep neural network trained on large-scale contact geometry data.

Insertion objects are held in the robot's grasp, while their corresponding receptacles featuring mating geometries are securely mounted on an electrical board to emulate industrial fixturing. Five distinct component types are used (Fig. 5), spanning a range of sizes, functions, pin counts, and visual appearances, thereby offering a diverse evaluation set. The overall task requires perceiving, transporting, and inserting objects into their designated receptacles. The insertion process demands sub-millimeter precision.

The proposed approach performs object pose estimation and continuously refines the estimates in closed-loop control, to guide, align, and insert the object into its receptacle with high precision. To evaluate modularity, the method was tested under three sensing modalities: vision-only observations (M-VOP), tactile-only observations (M-TOP), and combined visuo-tactile observations (M-VTOP). For each component and modality, experiments were repeated across three different grasp poses that varied the in-hand translation and orientation of the object (Fig. 6, Left). In addition, two separate masking backends were evaluated for each modality, further testing the robustness and generality of the approach.

Table I presents the average successful insertion rates for the five test objects shown in Fig. 5, comparing the combined and independent sensing modalities.

We evaluate the effectiveness of the pose estimation method for object insertion by analyzing translational and rotational errors relative to ground truth. In addition to insertion success rate, we measure these errors at the final timestep, with the ground-truth pose defined as the object's nominal in-receptacle configuration. Figs. 7 present per-object box plots comparing the zero-shot segmentation backend with the supervised baseline. Across objects, our method achieves sub-millimeter translational and sub-degree rotational accuracy under both segmentation settings while running at a frequency of 0.4 Hz for the whole system, satisfying the demands of high-precision insertion. Experiments further demonstrate a strong correlation between lower pose error and higher success rates, highlighting the importance of accurate object pose estimation for reliable robotic insertion.

VI. DISCUSSION

Beyond the challenges described above, robotic manipulation in unstructured, geometrically complex environments exposes the limitations of zero-shot perception models trained on broad, out-of-domain datasets. In our system, the zero-shot segmentation backend occasionally produces erroneous

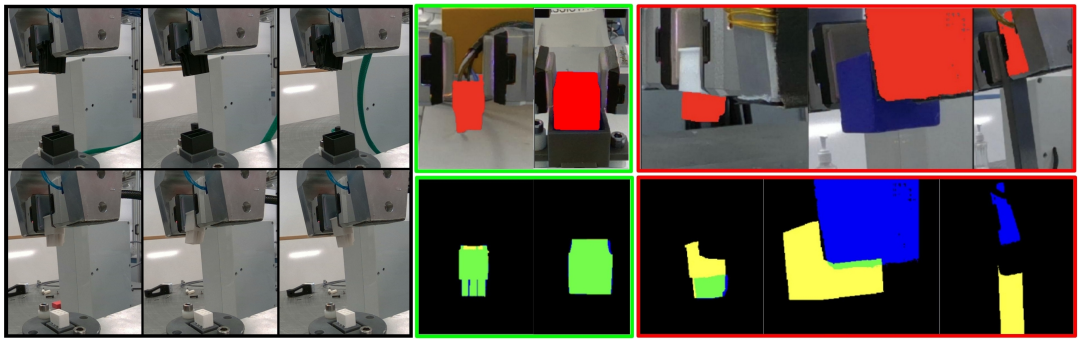

Fig. 6: *Left:* Figure shows the grasp variations used in the experimental setup for in-hand pose estimation and insertion, demonstrated for CONN1 and CONN2. *Right:* Figure shows a sample of good and bad cases for masking (top-row) and matching (bottom-row) with the zero-shot backend, for the objects CONN2, Cube, and Shaft Peg, where predicted masks in blue and rendered mask are shown in yellow. Green border: successful matching; Red border: partial or incorrect matching.

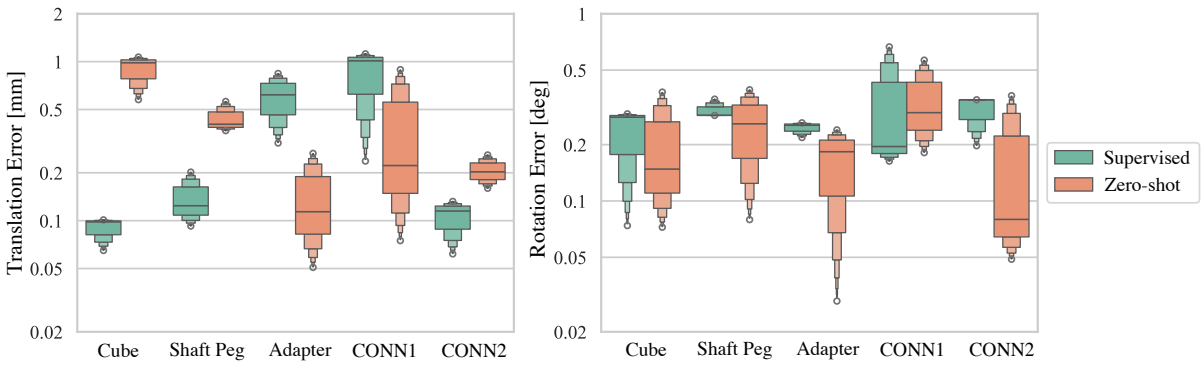

Fig. 7: Log-scale translation (left) and rotation (right) errors per object for M-VTOP with different masking backbones.

masks, for example, under-segmentation of the target caused by variations in geometry, texture, or illumination. These errors can lead to poor alignment with the rendered mask (Fig. 6, Right). Nevertheless, the pipeline remains robust: its modular design and particle-filter estimator prevent inconsistent measurements from heavily influencing the final estimate, so failures from a single sensor rarely cause catastrophic performance degradation. As a result, even with a zero-shot backend, our method achieves insertion success rates and translation/rotation errors that are competitive with a supervised baseline (Mask R-CNN trained on in-domain, human-labeled data), as shown in Table I and Fig. 7.

In our experiments, we use a high-resolution tactile sensor (GelSight Mini). In principle, however, our framework could also operate with lower-resolution tactile sensors. The primary limitation would be reduced pose precision, since finer geometric details and small misalignments may not be reliably captured by the measurement model, particularly in tasks that require tight translational and angular tolerances. Nevertheless, for applications with less stringent accuracy requirements or objects with larger geometric features, lower-resolution sensors may offer an attractive trade-off among cost, robustness, and computational efficiency.

Beyond sensing resolution, the intrinsic difficulty of the manipulation task also varies significantly across objects. Electrical connectors present the greatest challenge: successful insertion requires tight translational and angular tolerances, and even small yaw or roll errors can induce pin-socket misalignment. Attached cables further complicate the task by applying torque that biases the grasp pose and by introducing occlusions. The peg object color and surface finish closely matched those of the gripper, resulting in low visual contrast and frequent segmentation failures. Finally, although we incorporate a contact-based cue, its contribution to the final accuracy is limited for sub-millimeter requirements. By the time contact occurs, the particle set has typically already converged to a narrow hypothesis, so soft-threshold scores derived from a binary contact signal provide little additional information to significantly shift the posterior.

In addition to task-specific challenges, our framework relies on several modeling assumptions that may affect its ability to generalize. We assume rigid-body objects, and the motion model implicitly assumes that the object does not deviate significantly from the grasp point as a result of robot motion. When these assumptions are violated, larger-than-expected relative displacements may occur. In principle, the framework can accommodate such scenarios by increasing the motion deviation parameter σ_n, allowing the particle filter to account for greater motion uncertainty. However, this adjustment may lead to slower convergence and potentially reduced estimation accuracy.

Finally, the computational complexity of our pipeline is primarily driven by the particle filter and the mask-scoring procedure. For each particle, we render the predicted object pose and compute its similarity with tactile and vision mask proposals. This results in a complexity that scales linearly

with the number of particles and quadratically with the number of mask proposals per modality. In practice, however, the number of mask proposals is relatively small compared to the number of particles, so the overall runtime is effectively linear in the number of particles.

VII. CONCLUSION

We introduced M-VTOP, a modular framework for 6D object pose estimation that fuses vision, tactile, and contact sensing to support high-precision robotic manipulation. Using a pose particle filter to integrate multimodal observations, our method achieves zero-shot adaptation to unseen object geometries without requiring task-specific training. The framework is flexible and supports different sensor combinations, including externally mounted cameras and tactile sensors, enabling deployment across a wide range of robotic setups. Our evaluations on fine manipulation task, such as insertion, demonstrates both the robustness and generality of the approach. Future work may extend the approach to unstructured environments, incorporate additional sensing modalities, and support more diverse manipulation tasks.

REFERENCES

[1] Y. Labbé, L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, J. Carpentier, M. Aubry, D. Fox, and J. Sivic, "Megapose: 6d pose estimation of novel objects via render & compare," *arXiv preprint arXiv:2212.06870*, 2022.

[2] V. N. Nguyen, T. Groueix, M. Salzmann, and V. Lepetit, "Gigapose: Fast and robust novel object pose estimation via one correspondence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[3] G. Wang, F. Manhardt, F. Tombari, and X. Ji, "Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[4] H. Chang, A. Boularias, and S. Jain, "Insert-one: One-shot robust visual-force servoing for novel object insertion with 6-dof tracking," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.

[5] 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.

[6] G. M. Caddeo, N. A. Piga, F. Bottarel, and L. Natale, "Collision-aware in-hand 6d object pose estimation using multiple vision-based tactile sensors," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[7] N. Kuppuswamy, A. Castro, C. Phillips-Grafflin, A. Alspach, and R. Tedrake, "Fast model-based contact patch and pose estimation for highly deformable dense-geometry tactile sensors," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1811–1818, 2019.

[8] J. Bimbo, S. Rodriguez-Jimenez, H. Liu, X. Song, N. Burrus, L. D. Senerivatne, M. Abderrahim, and K. Althoefer, "Object pose estimation and tracking by fusing visual and tactile information," in *2012 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, 2012.

[9] S. Dikhale, K. Patel, D. Dhingra, I. Naramura, A. Hayashi, S. Iba, and N. Jamali, "Visuotactile 6d pose estimation of an in-hand object using vision and tactile sensor data," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2148–2155, 2022.

[10] H. Li, S. Dikhale, S. Iba, and N. Jamali, "Vihope: Visuotactile in-hand object 6d pose estimation with shape completion," *IEEE Robotics and Automation Letters*, 2023.

[11] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the IEEE international conference on computer vision*, 2015.

[12] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *arXiv preprint arXiv:1711.00199*, 2017.

[13] S. Zakharov, I. Shugurov, and S. Ilic, "Dpod: 6d pose object detector and refiner," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.

[14] A. Caraffa, D. Boscaini, A. Hamza, and F. Poiesi, "Freeze: Training-free zero-shot 6d pose estimation with geometric and vision foundation models," in *European Conference on Computer Vision*, 2025.

[15] Y. Su, M. Saleh, T. Fetzer, J. Rambach, N. Navab, B. Busam, D. Stricker, and F. Tombari, "Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[16] Y. Labbé, J. Carpentier, M. Aubry, and J. Sivic, "Cosypose: Consistent multi-view multi-object 6d pose estimation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*. Springer, 2020, pp. 574–591.

[17] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox, "Deepim: Deep iterative matching for 6d pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[18] 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.

[19] M. Oller, M. P. i Lisbona, D. Berenson, and N. Fazeli, "Manipulation via membranes: High-resolution and highly deformable tactile sensing and control," in *Conference on Robot Learning*. PMLR, 2022, pp. 1850–1859.

[20] S. Rodriguez, Y. Dou, M. Oller, A. Owens, and N. Fazeli, "Touch2touch: Cross-modal tactile generation for object manipulation," *arXiv preprint arXiv:2409.08269*, 2024.

[21] J. A. Eyzaguirre, M. Oller, and N. Fazeli, "Tactile neural de-rendering," *arXiv preprint arXiv:2409.13923*, 2024.

[22] M. Oller, D. Berenson, and N. Fazeli, "Tactilevad: Geometric aliasing-aware dynamics for high-resolution tactile control," in *Conference on Robot Learning*. PMLR, 2023, pp. 3083–3099.

[23] J. Zhao, M. Bauza, and E. H. Adelson, "Fingerslam: Closed-loop unknown object localization and reconstruction from visuo-tactile feedback," *arXiv preprint arXiv:2303.07997*, 2023.

[24] S. Suresh, Z. Si, S. Anderson, M. Kaess, and M. Mukadam, "Midastouch: Monte-carlo inference over distributions across sliding touch," in *Conference on Robot Learning*. PMLR, 2023, pp. 319–331.

[25] Y. Tu, J. Jiang, S. Li, N. Hendrich, M. Li, and J. Zhang, "Posefusion: Robust object-in-hand pose estimation with selectlstm," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023.

[26] J. Lee and N. Fazeli, "Vitascope: Visuo-tactile implicit representation for in-hand pose and extrinsic contact estimation," *arXiv preprint arXiv:2506.12239*, 2025.

[27] Y. Gao, S. Matsuoka, W. Wan, T. Kiyokawa, K. Koyama, and K. Harada, "In-hand pose estimation using hand-mounted rgb cameras and visuotactile sensors," *IEEE Access*, vol. 11, pp. 17 218–17 232, 2023.

[28] L. Manuelli and R. Tedrake, "Localizing external contact using proprioceptive sensors: The contact particle filter," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016.

[29] A. Sipos and N. Fazeli, "Multiscope: Disambiguating in-hand object poses with proprioception and tactile feedback," *arXiv preprint arXiv:2305.14204*, 2023.

[30] F. Von Drigalski, S. Taniguchi, R. Lee, T. Matsubara, M. Hamaya, K. Tanaka, and Y. Ijiri, "Contact-based in-hand pose estimation using bayesian state estimation and particle filtering," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[31] F. von Drigalski, K. Hayashi, Y. Huang, R. Yonetani, M. Hamaya, K. Tanaka, and Y. Ijiri, "Precise multi-modal in-hand pose estimation using low-precision sensors for robotic assembly," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[32] V. N. Nguyen, T. Groueix, G. Ponimatkin, V. Lepetit, and T. Hodan, "Cnos: A strong baseline for cad-based novel object segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[33] M. Matl *et al.*, "Pyrender," *GitHub Repository*, 2019.

[34] J. Pan, S. Chitta, and D. Manocha, "Fcl: A general purpose library for collision and proximity queries," in *2012 IEEE International Conference on Robotics and Automation*, 2012.