V-HOP: Visuo-Haptic 6D Object Pose Tracking

Hongyu Li[1], Mingxi Jia[1], Tuluhan Akbulut[1], Yu Xiang[2], George Konidaris[1], and Srinath Sridhar[1]

[1]Brown University [2]The University of Texas at Dallas
Email: hli230@cs.brown.edu, {mingxi_jia, mete_akbulut, gdk, srinath}@brown.edu, yu.xiang@utdallas.edu

Abstract—**Humans naturally integrate vision and haptics for robust object perception during manipulation. The loss of either modality significantly degrades performance. Inspired by this multisensory integration, prior object pose estimation research has attempted to combine visual and haptic/tactile feedback. Although these works demonstrate improvements in controlled environments or synthetic datasets, they often underperform vision-only approaches in real-world settings due to poor generalization across diverse grippers, sensor layouts, or sim-to-real environments. Furthermore, they typically estimate the object pose for each frame independently, resulting in less coherent tracking over sequences in real-world deployments. To address these limitations, we introduce a novel unified haptic representation that effectively handles multiple gripper embodiments. Building on this representation, we introduce a new visuo-haptic transformer-based object pose tracker that seamlessly integrates visual and haptic input. We validate our framework in our dataset and the Feelsight dataset, demonstrating significant performance improvement on challenging sequences. Notably, our method achieves superior generalization and robustness across novel embodiments, objects, and sensor types (both taxel-based and vision-based tactile sensors). In real-world experiments, we demonstrate that our approach outperforms state-of-the-art visual trackers by a large margin. We further show that we can achieve precise manipulation tasks by incorporating our real-time object tracking result into motion plans, underscoring the advantages of visuo-haptic perception. Project website: https://ivl.cs.brown.edu/research/v-hop.**

I. Introduction

Accurately tracking object poses is a core capability for robotic manipulation, and would enable contact-rich and dexterous manipulations with efficient imitation or reinforcement learning [68, 31, 23]. Recent state-of-the-art object pose estimation methods, such as FoundationPose [70], have significantly advanced visual tracking by leveraging large-scale datasets. However, relying solely on visual information to perceive objects can be challenging, particularly in contact-rich or in-hand manipulation scenarios involving high occlusion and rapid dynamics.

The cognitive science findings show that humans naturally integrate visual and haptic information for robust object perception during manipulation [46, 12, 28]. For instance, Gordon et al. [15] demonstrated that humans use vision to hypothesize object properties and haptics to refine precision grasps. The human "sense of touch" consists of two distinct senses [42, 6]: the *cutaneous sense*, which detects stimulation on the skin surface, and *kinesthesis*, which provides information on static and dynamic body posture. This integration, known as **haptic perception**, allows humans to effectively perceive and manipulate objects [28]. In robotics, analogous capabilities are achieved through tactile sensors (cutaneous sense) and joint sensors (kinesthesis) [46].

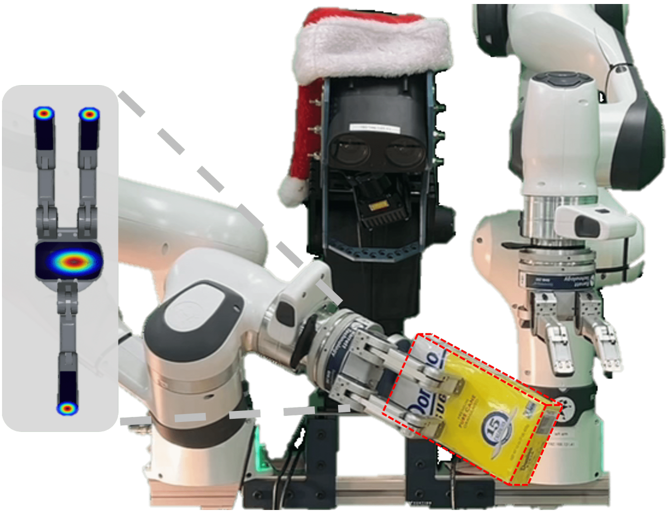


Fig. 1: **Visuo-haptic sensing for 6D object pose tracking**. We fuse *egocentric* visual and haptic sensing to achieve accurate real-time in-hand object tracking.

Drawing inspiration from these human capabilities, researchers have explored the integration of vision and touch in robotics for decades. As early as 1988, Allen [1] proposed an object recognition system that combined these modalities. More recently, data-driven approaches have emerged to tackle object pose estimation using visuo-tactile information [32, 54, 8, 61, 50, 59, 14, 33]. Although promising, these methods face two major barriers that hinder their broader applicability: (i) **Cross-embodiment**: Most approaches overfit specific grippers or tactile sensor layouts, limiting their adaptability. (ii) **Domain generalization**: Compared to visual-only baselines, visuo-tactile approaches struggle to generalize, hindered by insufficient data diversity and model scalability. Moreover, they typically process each frame independently, which can result in less coherent object pose tracking over sequences in real-world deployments. As a result, existing methods are challenging to deploy broadly and often require significant customization to specific robotic platforms.

To address these challenges, we propose **V-HOP** (Fig. 1): a two-fold solution for generalizable visuo-haptic 6D object pose tracking. First, we introduce a novel unified haptic representation that facilitates cross-embodiment learning. We consider the combination of tactile and kinesthesis in the form of a point cloud, addressing a critical yet often overlooked aspect of visuo-haptic learning. Second, we propose a transformer-based object pose tracker to fuse visual and haptic features. We leverage the robust visual prior captured by the visual foundation model while incorporating haptics. V-HOP accommodates diverse gripper embodiments and various objects and generalizes to novel embodiments and objects.

We build a multi-embodied dataset with eight grippers using the NVIDIA Isaac Sim simulator for training and evaluation. Compared to FoundationPose [70], our approach achieves 5% improvement in the accuracy of object pose estimation in terms of ADD-S [72] in our dataset. These results highlight the benefit of fusing visual and haptic sensing. In the FeelSight dataset [54], we benchmark against NeuralFeels [54], an optimization-based visuo-tactile object pose tracker, achieving a 32% improvement in the ADD-S metric and *ten times faster* run-time speed. Finally, we perform the sim-to-real transfer experiments using Barrett Hands. Our method demonstrates remarkable robustness and significantly outperforms FoundationPose, which could lose object tracks entirely (Fig. 5). When integrated into motion plans, our approach achieves 40% higher average task success rates. *To the best of our knowledge, V-HOP is the first data-driven visuo-haptic approach to demonstrate robust generalization across both taxel-based tactile sensors (e.g., Barrett Hand) and vision-based tactile sensors (e.g., DIGIT sensors), as well as on novel embodiments and objects.*

In conclusion, our contributions to this paper are two-fold:

1) **Unified haptic representation**: we introduce a novel haptic representation, enabling cross-embodiment learning and addressing the **cross-embodiment challenge** by improving adaptability across diverse embodiments and objects.
2) **Visuo-haptic transformer**: We present a transformer model that integrates visual and haptic data, improving pose tracking consistency and addressing the **domain generalization challenge**.

II. Background

In this section, we first define the problem formally and then review existing haptic representations and our proposed unified representation.

A. Problem Definition

We tackle the model-based visuo-haptic 6D object pose tracking problem, assuming access to:

- Visual observations: An RGB-D sensor observes the object in the environment.
- Haptic feedback: The object is manipulated by a rigid gripper equipped with tactile sensors.

Our approach takes the following as input:

1) The CAD model $\mathcal{M}_o$ of the object.
2) A sequence of RGB-D images $\mathcal{O} = \{\mathbf{O}_i\}_{i=1}^{t}$, where each observation $\mathbf{O}_i = [\mathbf{I}_i, \mathbf{D}_i]$ includes an RGB image $\mathbf{I}_i$ and a depth map $\mathbf{D}_i$.
3) An initial 6D pose $\mathbf{T}_0 = (\mathbf{R}_0, \mathbf{t}_0) \in \mathrm{SE}(3)$, where $\mathbf{R}_0 \in \mathrm{SO}(3)$ is 3D rotation and $\mathbf{t}_0 \in \mathbb{R}^3$ is 3D translation.

In practice, the ground-truth initial pose $\mathbf{T}_0$ is hard to obtain and can only be estimated through pose estimation [72, 62, 48, 36, 63, 30, 70, 27, 18, 40, 37, 57, 67]. Therefore, we treat $\widehat{\mathbf{T}}_0 = \mathbf{T}_0$ in the following. At each timestep i, our model estimates the object pose $\widehat{\mathbf{T}}_i$ given all the inputs with the initial pose being the estimate $\widehat{\mathbf{T}}_{i-1}$ at the previous timestep.

The above inputs are the standard inputs from the model-based visual pose tracking problem [66, 7], while the inputs below will serve our haptic representation and will be detailed in later sections.

4) Gripper description in Unified Robot Description Format (URDF).
5) Gripper joint positions $\mathbf{j} = \{j_1, j_2, \ldots, j_{DoF}\}$.
6) Tactile sensor data $\mathcal{S}$, including Positions $\mathcal{S}_p$ and readings $\mathcal{S}_r$ of tactile sensors, which will be formally defined in the next section.
7) Transformation between the camera and the robot frames obtained through hand-eye calibration [44].

B. Haptic Representation

The effectiveness of haptic learning hinges on its representation. Prior approaches using raw value [38], image [16], or graph-based [75, 33, 50] representations often struggle to generalize across diverse embodiments. For instance, Wu et al. [71] and Guzey et al. [16] project tactile signals from Xela sensors into a 2D image format. While this allows efficient processing with existing visual models, extending the method to different grippers or sensor layouts proves challenging. Similarly, Li et al. [33] and Rezazadeh et al. [50] employ graph-based mappings, where taxels are represented as vertices. However, variations in sensor layouts result in different graph distributions, creating significant generalization gaps.

In contrast, we adopt a point cloud representation, which naturally encode 3D positions and can flexibly accommodate multi-embodiments. We broadly classify tactile sensors into *taxel-based* and *vision-based*. A more comprehensive review on tactile sensors can be found at [74]. Below, we outline how they are converted into point clouds in prior works [8, 54, 64, 13], paving the way for our unified framework.

Taxel-based Sensors. The tactile data is defined as $\mathcal{S} = \{s_i\}_{i=1}^{n_t}$, which encapsulate n_t taxels. s_i represents individual taxels. The tactile data consists of $\mathcal{S} = (\mathcal{S}_p, \mathcal{S}_r)$:

- Positions ($\mathcal{S}_p$): Defined in the gripper frame and transformed into the camera frame using forward kinematics.
- Readings ($\mathcal{S}_r$): Capturing contact values. Readings are commonly binarized into contact or no-contact states [78, 73, 32, 8, 34] based on a threshold τ.

The set of taxels in contact is:

$$\mathcal{S}_c = \{s_i \in \mathcal{S} \mid \mathcal{S}_r(s_i) > \tau\}, \tag{1}$$

and the corresponding tactile point cloud $\mathcal{S}_{p,c}$ is defined as

$$\mathcal{S}_{p,c} = \{\mathcal{S}_p(s_i) \mid s_i \in \mathcal{S}_c\}. \tag{2}$$

Vision-based sensors. For vision-based tactile sensors [29, 79, 10, 56], the tactile data includes $\mathcal{S} = (\mathcal{S}_p, \mathcal{S}_I)$:

- Positions ($\mathcal{S}_p$): Sensor positions in the camera frame, similar to taxel-based.
- Images ($\mathcal{S}_I$): Capturing contact states using regular RGB image representation. Using the tactile depth estimation model [3, 54, 26, 53, 52, 2], we can convert $\mathcal{S}_I$ into tactile point cloud $\mathcal{S}_{p,c}$.

Yet we are not the first to employ point cloud representations for tactile learning, prior works [8, 54, 64, 13] focus on a single type of sensor and overlook the gripper posture. Our *key contribution* is a unified representation spanning both taxel-based and vision-based sensors on multi-embodiments, empowered by our multi-embodied dataset. We demonstrate generalizability on the Barrett hand (taxel-based) during our real-world experiments and on the Allegro hand (vision-based DIGIT sensor) using the Feelsight dataset [54]. Our novel haptic representation seamlessly integrates the tactile signals with the *gripper posture*, enabling more effective gripper-object interaction reasoning. In subsequent sections, we describe our approach and provide empirical evidence demonstrating that our representation improves generalization capabilities, bridging the gap between heterogeneous tactile sensor modalities.

III. Methodology

We propose V-HOP, a data-driven approach that fuses visual and haptic modalities to achieve accurate 6D object pose tracking. Our goal is to build a *generalizable* visuo-haptic pose tracker that accommodates diverse embodiments and objects. We first outline the core representations used in our haptic modality: gripper and object representations. Our choice for the representations follows the spirit of the render-and-compare paradigm [35]. Later, we introduce our visuo-haptic model and how it is trained.

A. Gripper Representation

Tactile signals only represent the cutaneous stimulation, while haptic sensing combines tactile and kinesthetic feedback to provide a more comprehensive spatial understanding of contact and manipulation. We propose a novel haptic representation that integrates tactile signals and gripper posture in a unified point cloud representation. This gripper-centric representation enables efficient reasoning about spatial contact and gripper-object interaction.

Using the URDF definition and joint positions $\mathbf{j}$, we generate the gripper mesh $\mathcal{M}_h$ through forward kinematics and calculate the surface normals. The mesh is then downsampled to produce a 9-D gripper point cloud $\mathcal{P}_h = \{\mathbf{p}_i\}_{i=1}^{n_h}$:

$$\mathbf{p}_i = (x_i, y_i, z_i, n_{ix}, n_{iy}, n_{iz}, \mathbf{c}) \in \mathbb{R}^9, \tag{3}$$

where x_i, y_i, z_i represent the 3-D coordinate of the point. n_{ix}, n_{iy}, n_{iz} represent the 3-D normal vectors, and $\mathbf{c} \in \mathbb{R}^3$ is a one-hot encoded point label:

- $[1, 0, 0]$: Gripper point in contact.
- $[0, 1, 0]$: Gripper point not in contact.
- $[0, 0, 1]$: Object point (for later integration with the object point cloud).

To obtain the contact state of each point, we map the tactile point cloud $\mathcal{S}_{p,c}$, representing the contact points detected by the tactile sensors (Sec. II-B), onto the downsampled gripper point cloud $\mathcal{P}_h$. Specifically, for each point in $\mathcal{S}_{p,c}$, we find its neighboring points in $\mathcal{P}_h$ within a radius r. These neighboring points are labeled as "in contact", while all others are labeled as "not in contact". The choice of the radius r is randomized during training and determined by the measured effective radius of each taxel during robot deployment. The resulting haptic point cloud, $\mathcal{P}_h$, serves as a unified representation for both tactile and kinesthetic data (Fig. 2).

B. Object Representation

We denote the object model point cloud as $\mathcal{P}_\Phi = \{\mathbf{q}_i\}_{i=1}^{n_o}$. Similar to the gripper point cloud, $\mathbf{q}_i$ follows the same 9-D definitions (Equation 3),

$$\mathbf{q}_i = (x_i, y_i, z_i, n_{ix}, n_{iy}, n_{iz}, \mathbf{c}) \in \mathbb{R}^9,$$

with $\mathbf{c} = [0, 0, 1]$ for all object points. At each timestep $i > 0$, we transform the model point cloud into a hypothesized point cloud $\mathcal{P}_o = \{\mathbf{q}'_i\}_{i=1}^{n_o}$ according to the pose from the previous timestep $\mathbf{T}_{i-1}$. For each point $\mathbf{q}'_i$ in the hypothesized point cloud $\mathcal{P}_o$

$$\mathbf{q}'_i = (x'_i, y'_i, z'_i, n'_{ix}, n'_{iy}, n'_{iz}, \mathbf{c}), \tag{4}$$

where:

$$\begin{bmatrix} x'_i \\ y'_i \\ z'_i \end{bmatrix} = \mathbf{R}_{i-1} \begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} + \mathbf{t}_{i-1}, \quad \begin{bmatrix} n'_{ix} \\ n'_{iy} \\ n'_{iz} \end{bmatrix} = \mathbf{R}_{i-1} \begin{bmatrix} n_{ix} \\ n_{iy} \\ n_{iz} \end{bmatrix}. \tag{5}$$

To enable reasoning about gripper-object interactions, we fuse the gripper point cloud $\mathcal{P}_h$ and the hypothesized object point cloud $\mathcal{P}_o$ to create a gripper-object point cloud $\mathcal{P}$,

$$\mathcal{P} = \mathcal{P}_h \cup \mathcal{P}_o. \tag{6}$$

This novel unified representation adopts the principles of the render-and-compare paradigm from visual approaches [35, 66, 27, 70, 58], in which the rendered image (based on pose hypothesis) is compared against the visual observation. The hypothesized object point cloud $\mathcal{P}_o$ serves as the "rendered" pose hypothesis (Fig. 2). The gripper point cloud $\mathcal{P}_h$ represents the real observation using haptic feedback, which we used to compare with. By leveraging this representation, the model captures the contact-rich interactions between the gripper and the object by learning feasible object poses informed by haptic feedback.

C. Network Design

Visual modality. Unlike prior works, which train the whole visuo-haptic network from scratch, our approach can effectively leverage the pretrained visual foundation model. Our design extends the formulation of FoundationPose [70], as it demonstrates great generalizability on unseen objects and a narrow sim-to-real gap. To harness the high-quality visual

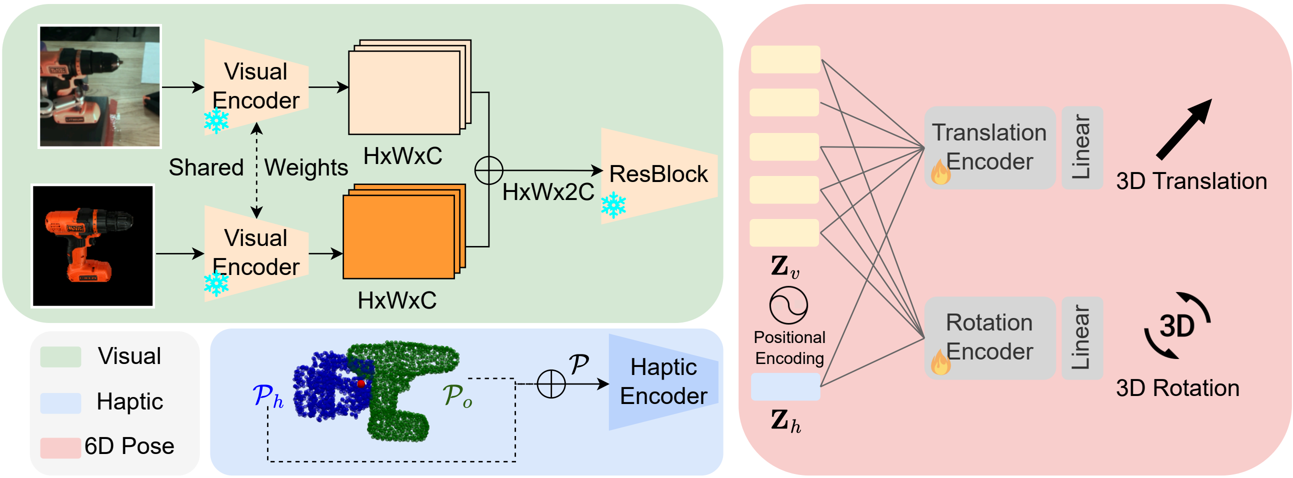


Fig. 2: **Network design of V-HOP.** The visual modality, based on FoundationPose [70], uses a visual encoder to process RGB-D observations (real and rendered) into feature maps, which are concatenated and refined through a ResBlock to produce visual embeddings [11]. The haptic modality encodes a unified gripper-object point cloud, derived from 9D gripper $\mathcal{P}_h$ and object $\mathcal{P}_o$ point clouds, into a haptic embedding that captures gripper-object interactions. The **red dot** in the figure denotes the activated tactile sensor. These visual and haptic embeddings are processed by Transformer encoders to estimate 3D translation and rotation.

prior captured by it, we utilize its visual encoder f_v and freeze it during our training. Using this encoder, We transform the RGB-D observation into visual embeddings $\mathbf{Z}_v = f_v(\mathbf{O})$.

Haptic modality. In parallel, we encode the gripper-object point cloud $\mathcal{P}$ using a haptic encoder f_h, resulting in a haptic embedding $\mathbf{Z}_h = f_h(\mathcal{P})$. By representing all interactions in point cloud space, our novel haptic representation provides the flexibility to utilize any point cloud-based network for encoding. For this purpose, we choose PointNet++ [49] as our haptic encoder f_h. To improve learning efficiency, we canonicalize the point cloud using the centroid of the gripper points, ensuring $\mathcal{P}$ is spatially centered around the gripper during processing.

Visuo-haptic fusion. Integrating visual and haptic modalities, however, poses significant challenges. Existing methods often apply fixed or biased weightings between these modalities [32, 54, 8, 59], which can fail under specific conditions. For example, when contact is absent, the visual modality alone should be leveraged, or when occlusion is severe, haptics should be favored. Inspired by the principle of "optimal integration" in human multisensory perception [12, 19, 28, 55, 20], where the brain dynamically adjusts the weighting of visual and haptic inputs to maximize reliability, we adopt self-attention mechanisms [60] for the adaptive fusion of visual and haptic embeddings. This ensures robustness across varying scenarios, whether the object is in contact or in clear view.

To achieve this fusion, we propose haptic instruction-tuning, inspired by visual instruction-tuning [39]. While keeping the visual encoder f_v frozen, we feed both visual embedding $\mathbf{Z}_v$ and haptic embedding $\mathbf{Z}_h$ into the original visual-only Transformer encoders [60, 70], which are initialized with the pretrained weights from FoundationPose. We then fine-tune the Transformer encoders and the haptic encoder f_h together. Consequently, visual and haptic information is fused adaptively using self-attention blocks, and the model dynamically adjusts the modality weight based on the context (Fig. 9). Following FoundationPose, we disentangle the 6D pose into 3D translation and 3D rotation and estimate them using two output heads (Fig. 2), respectively.

D. Training Paradigm

We train our model by adding noise $(\mathbf{R}_\epsilon, \mathbf{t}_\epsilon)$ to the ground-truth pose $\mathbf{T} = (\mathbf{R}, \mathbf{t})$ to create the hypothesis pose $\widetilde{\mathbf{T}} = (\widetilde{\mathbf{R}}, \widetilde{\mathbf{t}})$:

$$\widetilde{\mathbf{R}} = \mathbf{R}_\epsilon^{-1} \cdot \mathbf{R}, \quad \widetilde{\mathbf{t}} = -\mathbf{t}_\epsilon + \mathbf{t}. \tag{7}$$

The rendered image is generated using $\widetilde{\mathbf{T}}$, while the object point cloud is transformed based on $\widetilde{\mathbf{T}}$; in contrast, the RGB-D image and gripper point cloud represent actual observations. The model estimates the relative pose $\Delta\widehat{\mathbf{T}} = (\Delta\widehat{\mathbf{R}}, \Delta\widehat{\mathbf{t}})$ between the pose hypothesis and observation. The model is optimized using the L_2 loss:

$$\mathcal{L}_{\mathbf{T}} = \|\Delta\widehat{\mathbf{R}} - \mathbf{R}_\epsilon\|_2 + \|\Delta\widehat{\mathbf{t}} - \mathbf{t}_\epsilon\|_2, \tag{8}$$

where we use quaternion representations for rotations. The estimated pose $\widehat{\mathbf{T}} = (\widehat{\mathbf{R}}, \widehat{\mathbf{t}})$ is:

$$\widehat{\mathbf{R}} = \Delta\widehat{\mathbf{R}} \cdot \widetilde{\mathbf{R}}, \quad \widehat{\mathbf{t}} = \Delta\widehat{\mathbf{t}} + \widetilde{\mathbf{t}}. \tag{9}$$

We further incorporate an attractive loss ($\mathcal{L}_a$) and a penetration loss ($\mathcal{L}_p$) to encourage the object to make contact with the tactile point cloud $\mathcal{S}_{p,c}$ and avoid penetrating the gripper point cloud $\mathcal{P}_h$. We first transform the initial hypothesized object pose cloud $\mathcal{P}_o$ using the estimated pose $\widehat{\mathcal{P}}_o = \widehat{\mathbf{T}}\,\widetilde{\mathbf{T}}^{-1}\,\mathcal{P}_o$, where $\mathcal{P}_o$ is in homogenous form.

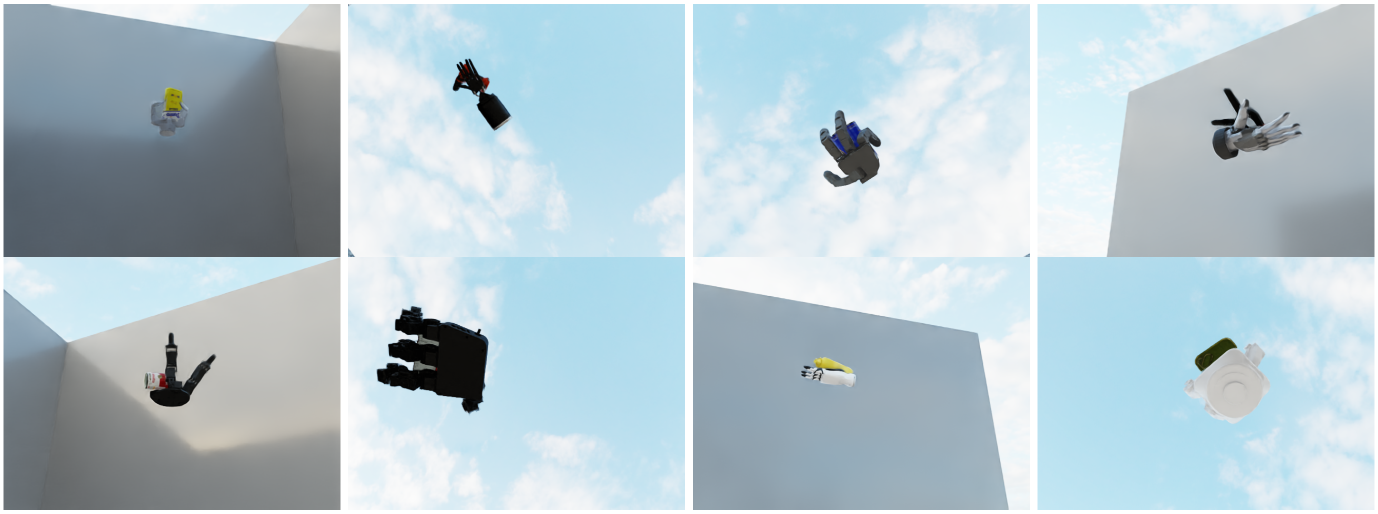


Fig. 3: **Dataset sample visualization**. (Top row) Barrett Hand, Shadow Hand, Allegro Hand, SHUNK SVH. (Bottom row) D'Claw, LEAP Hand, Inspire Hand, Robotiq 3-Finger gripper.

The attractive loss enforces that each activated taxel must make contact with the object:

$$\mathcal{L}_a = \frac{1}{|\mathcal{S}_{p,c}|} \sum_{s_{p,c} \in \mathcal{S}_{p,c}} \min_{p \in \widehat{\mathcal{P}}_o} \|s_{p,c} - p\|^2, \tag{10}$$

which can be interpreted as a single-direction Chamfer distance between the tactile point cloud and the object point cloud.

The penetration loss avoids penetrations between the object and the gripper [76, 77, 4]:

$$\begin{aligned} \mathbf{p}_o &= \arg\min_{\mathbf{q} \in \widehat{\mathcal{P}}_o} \|\mathbf{q} - \mathbf{p}_h\|_2, \\ \mathcal{L}_p &= \sum_{\mathbf{p}_h \in \mathcal{P}_h} e^{\max\{0,\, -\mathbf{n}_o \cdot (\mathbf{p}_h - \mathbf{p}_o)\}} - 1, \end{aligned} \tag{11}$$

where $\mathbf{p}_o$ represents the nearest neighbor of each point $\mathbf{p}_h$ in the gripper point cloud $\mathcal{P}_h$. Our overall loss is:

$$\mathcal{L} = \mathcal{L}_{\mathbf{T}} + \alpha \mathcal{L}_a + \beta \mathcal{L}_p, \tag{12}$$

where we set $\alpha = 1$ and $\beta = 0.001$ empirically. We optimize the model using the AdamW [43] optimizer with an initial learning rate of 0.0004 and train the model for 20 epochs.

IV. Experiments

A. Multi-embodied Dataset

Existing visuo-haptic datasets were not publicly available [8, 33, 61] at the time of completing this work and focused on a single gripper [54], leaving the question of generalization to novel embodiments unanswered. Consequently, we develop a multi-embodied dataset (Fig. 3) using NVIDIA Isaac Sim to enable cross-embodiment learning and thorough evaluation. Our dataset comprises approximately 1,550,000 images collected across eight grippers and thirteen objects. We utilize 85% of the data for training and the rest for validation. The camera trajectories are sampled on the semi-sphere around the gripper, which has a random radius between 0.5 and 2.5 meters. We selected graspable YCB object [5] and grippers used in prior works [9, 45]. Additional details about the dataset can be found in the appendix.

In this paper, we follow the sim-to-real paradigm and utilize only synthetic data for training. Increasing real-world training data could indeed help mitigate the sim-to-real gap. However, as demonstrated in recent work [70], leveraging a large-scale synthetic dataset enriched with domain randomization can yield superior real-world performance compared to small-scale real-world datasets. Our synthesized dataset exemplifies this principle and supports our robust real-world performance. Collecting real-world data with comparable scale and diversity would be both challenging and resource-intensive. Moreover, our unified haptic representation leverages point cloud representation to maintain invariance across various tactile sensors. Consequently, our sim-to-real experiments (Sec. V) demonstrate robust performance and eliminate the need for costly real-world data collection.

B. Pose Tracking Comparison

In the following experiments, we evaluate performance using the metrics:

- Area under the curve (AUC) of ADD and ADD-S [21, 72], and
- ADD(-S)-0.1d [18]: ADD/ADD-S that is less than 10% of the object diameter.

We compare V-HOP against the current state-of-the-art approaches in visual pose tracking (FoundationPose [70], or FP in short) and visuo-tactile pose estimation (ViTa [8]). To ensure a fair comparison, we finetune FoundationPose and train ViTa on our multi-embodied dataset. To verify the generalizability of the novel object and novel gripper, we exclude one object (pudding_box) and one gripper (D'Claw) during training.

Due to the absence of a visuo-haptic pose tracking approach, we compare V-HOP with ViTa, an instance-level visuo-tactile

Object Name	AUC Metric	ViTa	FP	V-HOP
master_chef_can	ADD	5.61	**64.95**	62.88
	ADD-S	80.51	84.60	**86.38**
sugar_box	ADD	11.09	73.21	**74.75**
	ADD-S	74.34	85.27	**89.35**
tomato_soup_can	ADD	32.08	57.02	**59.13**
	ADD-S	84.19	78.45	**83.30**
mustard_bottle	ADD	7.23	72.65	**74.07**
	ADD-S	73.49	86.05	**88.57**
pudding_box (Unseen)	ADD	N/A	69.87	**70.75**
	ADD-S	N/A	84.63	**88.20**
gelatin_box	ADD	43.20	63.89	**69.75**
	ADD-S	86.66	80.16	**86.87**
potted_meat_can	ADD	34.13	65.62	**68.29**
	ADD-S	86.77	82.67	**87.21**
banana	ADD	23.93	63.87	**69.72**
	ADD-S	71.67	79.99	**85.79**
mug	ADD	35.05	**59.60**	58.42
	ADD-S	86.58	82.16	**84.10**
power_drill	ADD	2.58	67.21	**68.56**
	ADD-S	61.02	80.77	**85.77**
scissors	ADD	23.34	66.23	**70.67**
	ADD-S	65.56	81.27	**85.08**
large_marker	ADD	42.43	61.74	**71.10**
	ADD-S	73.69	75.45	**85.00**
large_clamp	ADD	30.56	71.64	**75.63**
	ADD-S	79.20	86.07	**89.09**
All	ADD ↑	23.93	66.29	**68.90**
	ADD-S ↑	76.87	82.37	**86.62**

TABLE I: **Per-object comparison of AUC metrics for ADD and ADD-S**. The row of novel object is grayed out. Both metrics are the higher, the better. The best results are **bolded**.

Gripper Name	AUC Metric	ViTa	FP	V-HOP
Allegro Hand	ADD	24.48	74.45	**76.20**
	ADD-S	77.60	88.74	**90.48**
Barrett Hand	ADD	24.63	77.67	**79.06**
	ADD-S	77.74	88.72	**91.73**
D'Claw (Unseen)	ADD	21.99	48.16	**57.49**
	ADD-S	76.00	77.06	**85.48**
Inspire Hand	ADD	24.56	**70.22**	70.15
	ADD-S	77.65	84.22	**87.28**
LEAP Hand	ADD	23.88	64.17	**69.96**
	ADD-S	77.55	83.06	**88.05**
Robotiq 3-Finger	ADD	23.40	**79.48**	79.14
	ADD-S	76.87	89.39	**90.61**
SCHUNK SVH	ADD	24.40	61.01	**62.75**
	ADD-S	76.74	78.58	**82.96**
Shadow Hand	ADD	23.81	58.77	**60.27**
	ADD-S	75.55	73.24	**79.35**
All	ADD ↑	23.93	66.29	**68.90**
	ADD-S ↑	76.87	82.37	**86.62**

TABLE II: **Per-gripper comparison of AUC metrics for ADD and ADD-S**. Our dataset contains eight grippers. We train the model on seven grippers, leaving one gripper (D'Claw) unseen.

Method	AUC ADD	ADD-0.1d	AUC ADD-S	ADD-S-0.1d
Without Tactile	60.10	43.69	77.33	63.17
Without Visual	32.19	3.72	58.85	31.44
V-HOP (Ours)	**68.90**	**48.55**	**86.62**	**77.83**

TABLE III: **Ablations of input modalities**. Our results confirm the effectiveness of combining visual and haptic modalities.

Fusion Type	AUC ADD	ADD-0.1d	AUC ADD-S	ADD-S-0.1d
Late Fusion	47.56	17.57	70.43	51.66
Early Fusion (Ours)	**68.90**	**48.55**	**86.62**	**77.83**

TABLE IV: **Ablations of fusion strategies**. We evaluate the performance of early fusion and late fusion strategies.

pose estimation approach that operates under different settings. For ViTa, we provide ground-truth segmentation and train a separate model for each object, as it is an instance-level method. In contrast, both FoundationPose and V-HOP handle novel-object estimation and require training only once. For fair evaluation, we run both methods for two iterations per tracking step. For V-HOP, we run one visuo-haptic iteration and one visual iteration.

In Tab. I, we show the performance for each object. V-HOP consistently outperforms ViTa and FoundationPose (FP) on most objects with respect to ADD and across all objects in terms of ADD-S. On average, our approach delivers an improvement of 4% in ADD and 5% in ADD-S compared to FoundationPose. Notably, V-HOP demonstrates strong performance on unseen objects, highlighting the potential of our model to generalize effectively to novel objects.

Similarly, Tab. II illustrates the performance of each gripper. In line with its object performance, V-HOP outperforms its counterparts on most grippers in terms of ADD and across all grippers in ADD-S. Moreover, V-HOP demonstrates robust performance on unseen grippers, further emphasizing the generalizability of our unified haptic representation.

C. Ablation on Modalities

We conduct an ablation study on the input modalities to evaluate the effectiveness of the haptic representation. Specifically, we train two ablated versions of V-HOP: one without tactile feedback and another without visual input, as shown in Tab. III. To exclude tactile input, we remove all "in contact" point labels (Equation 3). Our results indicate that visual input significantly contributes to performance, likely due to the richness of visual information, including texture and spatial details. This finding aligns with previous studies on human perception systems, which suggest that vision plays a dominant role in visuo-haptic integration [24]. Similarly, tactile feedback is crucial; without it, performance degrades notably because reasoning about gripper-object contact during interactions becomes more difficult.

D. Ablation on Fusion Strategies

We perform ablation studies on different modality fusion strategies: early fusion and late fusion. Early fusion refers to fusion at the input or feature level, the one we presented in Fig. 2. Late fusion strategy fuses the visual and tactile modalities at the result level, where each modality has a separate branch to estimate its result [59]. As shown in Tab. IV, the late fusion strategy results in an average ADD score of 47.56 and an ADD-S score of 70.43, which underperforms our early fusion design by 30.97% in ADD and 18.69% in ADD-S. The results confirm the necessity to fuse the visual and haptic modalities at the feature level.

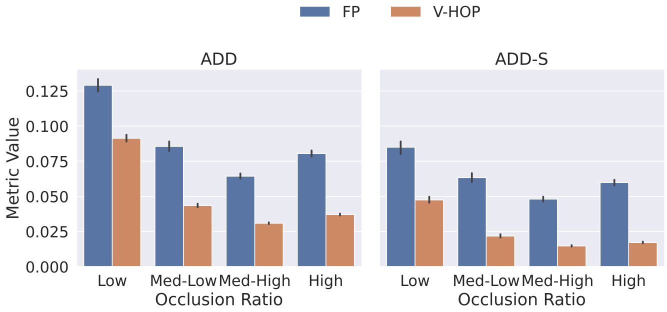


Fig. 4: **Performance under various occlusion ratios**. We use the direct ADD and ADD-S metrics (in meters) in this experiment.

Method	GT Seg	ADD-S ↓	ADD-S-0.1d ↑	FPS ↑
NeuralFeels [54]	✓	2.14	**98.95**	3
V-HOP (Ours)	✗	**1.46**	98.45	**32**

TABLE V: **Performance on the FeelSight Dataset**. For consistency with the metric used in NeuralFeels [54], this experiment reports the direct ADD-S metric [72] (in mm) rather than the AUC of ADD-S used in other experiments.

E. Occlusion's Effect on the Performance

We evaluate the performance of V-HOP and FoundationPose across varying occlusion ratios (Fig. 4). The occlusion ratio is defined as the proportion of pixels in the segmentation mask relative to the total pixels in the rendered object image, generated using the ground-truth pose. Our results show that V-HOP consistently outperforms FoundationPose in both ADD and ADD-S metrics under different levels of occlusion. These results underscore the importance of integrating visual and haptic information to improve performance in challenging occlusion scenarios.

F. Pose Tracking on FeelSight

To evaluate the generalizability of V-HOP, we benchmark it against NeuralFeels [54], a recently introduced optimization-based visuo-tactile pose tracking approach, using their proposed Feelsight dataset. Specifically, we focus on the occlusion subset of the dataset, `FeelSight-Occlusion`, which presents significant challenges due to severe occlusions. This subset requires robust generalization capabilities as it includes a **novel embodiment** (the Allegro hand equipped with DIGIT fingertips), a **novel sensor type** (a vision-based tactile sensor), and a **novel object** (a Rubik's cube). For a fair comparison, we compare against their model-based tracking approach, which uses almost the same inputs as V-HOP but with the ground-truth segmentation mask (GT Seg).

The results are presented in Tab. V. V-HOP achieves a 32% lower ADD-S error compared to NeuralFeels and has a similar ADD-S-0.1d score. It is important to note that NeuralFeels leverages the ground-truth segmentation mask, which helps in more accurate object localization, whereas V-HOP does not have such an input, further underscoring its robustness and adaptability.

In terms of computational efficiency, V-HOP is approximately 10 times faster than NeuralFeels, achieving 32 FPS compared to NeuralFeels' 3 FPS on an NVIDIA RTX 4070 GPU. This substantial improvement in speed highlights the practicality of V-HOP for real-world manipulation applications, as we will demonstrate in the later sections.

V. Sim-to-Real Transfer Experiments

To validate the real-world effectiveness of our approach, we perform sim-to-real experiments using our robot platform (Fig. 1). Our bimanual platform comprises dual Franka Research 3 robotic arms [17] and Barrett Hands BH8-282. Our Barrett Hand has 4 degrees of freedom (DoF) and 96 taxels: 24 taxels on each fingertip and 24 taxels on the palm. Each taxel comprises a capacitive cell capable of detecting forces within a range of 10 N/cm^2 with a resolution of 0.01 N. For egocentric visual input, we use a MultiSense SLB RGB-D camera, which combines a MultiSense S7 stereo camera and a Hokuyo UTM-30LX-EW laser scanner. We utilize FoundationPose to provide the initial frame pose estimate and CNOS [47, 25] to provide the segmentation task.

A. Pose Tracking Experiments

In this experiment (Fig. 5), the gripper stably grasps the object while a human operator guides the robot arm along a random trajectory. This introduces heavy occlusion and high dynamic motion to emulate challenging real-world manipulation scenarios. Under these conditions, FoundationPose often loses tracking due to reliance on visual input alone. In contrast, V-HOP maintains stable object tracking throughout the trajectory, demonstrating the robustness of its visuo-haptic sensing.

B. Bimanual Handover Experiment

In this experiment (Fig. 6), an object is placed on a table within reach of the robot's right arm. The task requires the robot to perform the following sequence of actions:

1) Use the right arm to grasp the object and transport it to the center.
2) Use the left arm to grasp the object from the right gripper and place it into a designated bin.

The robot employs model-based grasping, which depends on real-time object pose estimation. This task presents two key challenges:

1) If the grasp attempt fails, the robot must detect the failure based on the real-time object pose and reattempt the grasp.
2) During transport to the center, the robot must maintain precise tracking of the object's pose to ensure that the left arm can accurately grasp it. Inaccurate tracking results could lead to collision during the handover.

V-HOP enables the motion planner to handle objects in random positions and adapt to dynamic scenarios, such as human perturbations. For instance, a human may move the object during task execution, remove it from the gripper, or reposition it on the table (Fig. 7). Due to the integration of

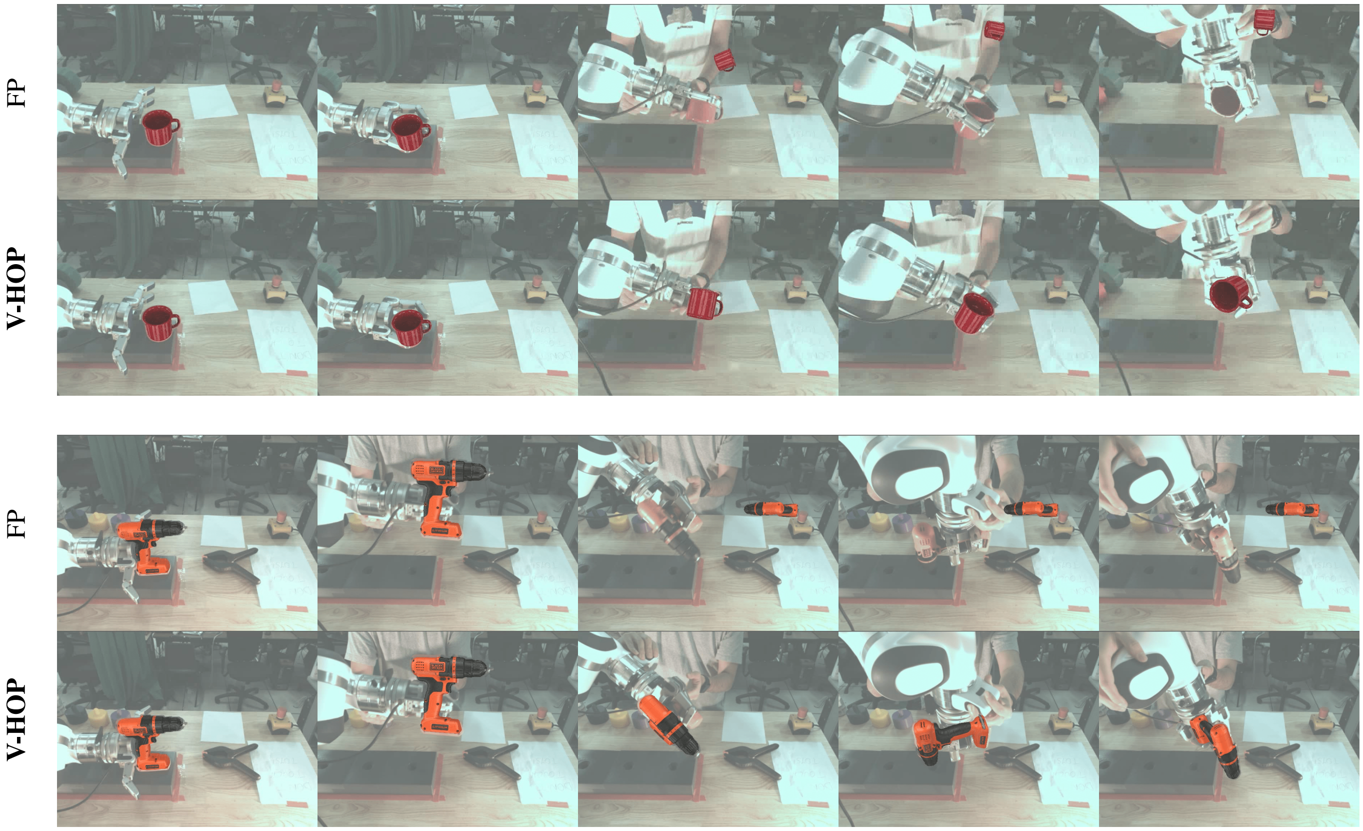


Fig. 5: **Qualitative results of pose tracking sequences**. We verify the performance in the real world using YCB objects. The cup and power drill are highlighted in this figure, while the results of more objects are in the appendix.

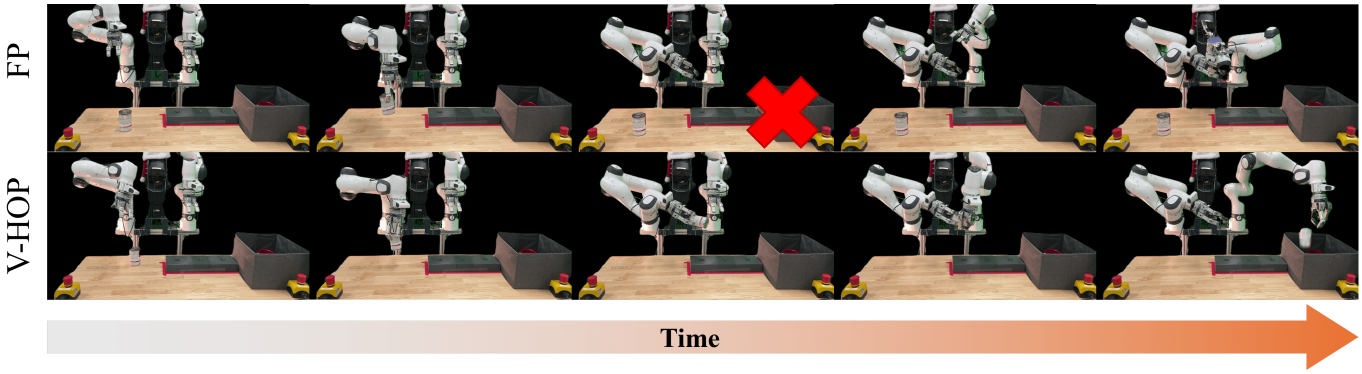


Fig. 6: **Bimanual handover experiment.** In this experiment, the robot performs bimanual manipulation to transport the target object to the box. V-HOP integrates visual and haptic inputs to accurately track the pose of the in-hand object in real-time, resulting in stable handover performance. Results on more objects can be found in the appendix.

Method	Sugar Box	Power Drill	Tomato Can	Average
FP	60	40	20	40
V-HOP	**80**	**80**	**80**	**80**

TABLE VI: Success rate on bimanual handover task.

haptic feedback, V-HOP accurately tracks the object's pose, allowing the robot to promptly detect and respond to changes, such as the object leaving the gripper. On the contrary, FoundationPose loses tracking during handover or grasping failure (Fig. 6) and leads to collisions. In Tab. VI, we show the success rate for each object for five trials. V-HOP has 40% higher success rate on average compared to FoundationPose.

C. Can-in-Mug Experiment

The Can-in-Mug task (Fig. 8) involves grasping a tomato can and inserting it into a mug. The bimanual version requires the robot to also grasp the mug and insert the can in the center. Successful execution hinges on precise pose estimation for both objects, as any noise in their poses can lead to failure.

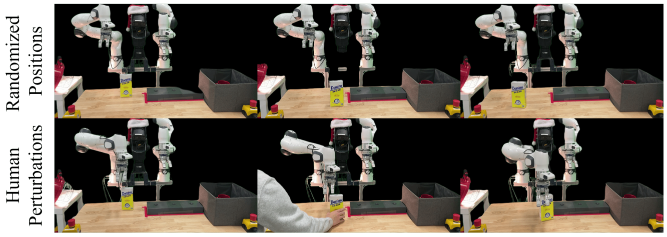
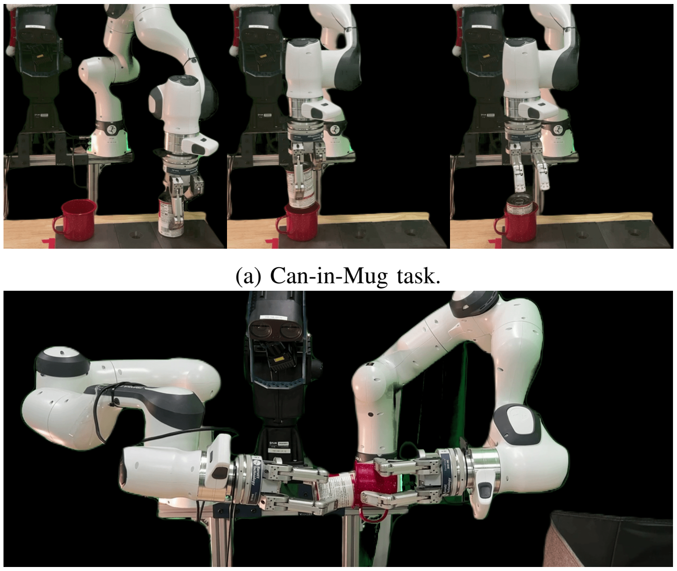


Fig. 7: **Robustness test for the bimanual handover task**. (Left) The object is placed at various randomized positions. (Right) A human perturbs the object by moving it to a different position while the robot attempts to grasp it.

Method	Can-in-Mug	Bimanual Can-in-Mug
FP	20	0
V-HOP	**60**	**20**

TABLE VII: Success rate on Can-in-Mug task.



(a) Can-in-Mug task.

(b) Bimanual Can-in-Mug task.

Fig. 8: **Can-in-Mug tasks.** (top) The robot grasps the can and inserts it into the mug. (bottom) The robot uses bimanual to grasp the can and the mug and insert the can into the mug in the center.

Our results (Tab. VII) demonstrate that V-HOP, by integrating visual and haptic inputs, delivers more stable tracking and a higher overall success rate.

D. Contribution of each modality

In this study, we examine the contribution of visual and haptic inputs to the final prediction. We adapt Grad-CAM [51], utilizing the final normalization layer of the Transformer encoder as the target layer. Figure 9 illustrates the weight distribution across the visual and haptic modalities. Our findings suggest that when the gripper is not in contact with an object, the model predominantly relies on visual inputs. However, as the gripper establishes contact and occlusion becomes more severe, the model increasingly shifts its reliance toward haptic inputs. This finding confirms the choice of self-attention mechanism to emulate human's "optimal integration" principle.

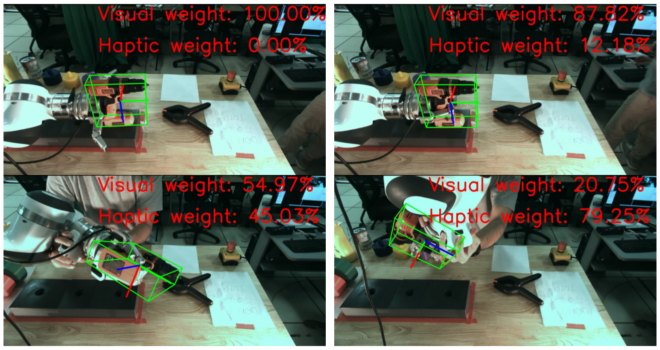


Fig. 9: **Weights of visual and haptic modalities to the final prediction**. We overlay the modality weights calculated using GradCAM [51] in the top-right corner.

VI. Related Works

In this work, we consider the problem of 6D object pose tracking problem, which has been widely studied as a visual problem [66, 35, 70, 7]. In particular, we focus on model-based tracking approaches, which assume access to the object's CAD model. While model-free approaches [65, 69, 54] exist, they fall outside the scope of this work. Visual pose tracking has achieved significant progress on established benchmarks, such as BOP [22]. Despite these successes, deploying such systems in real-world robotic applications remains challenging, especially under scenarios with high occlusion and dynamic interactions, such as in-hand manipulation tasks.

To address these challenges, prior research has explored combining visual and tactile information to improve pose tracking robustness [32, 54, 8, 61, 50, 59, 14, 33]. These approaches leverage learning-based techniques to estimate object poses by fusing visuo-tactile inputs. However, these methods estimate poses on a per-frame basis, which lacks temporal coherence. Additionally, cross-embodiment and domain generalization remain significant hurdles, limiting their scalability and practicality for broad deployment.

More recent works aim to overcome some of these limitations. For example, Liu et al. [41] proposes an optimization-based approach that integrates tactile data with visual pose tracking using an ad-hoc slippage detector and velocity predictor. Suresh et al. [54] extend the model-free tracking frameworks BundleTrack [65] and BundleSDF [69] by combining visual and tactile point clouds within a pose graph optimization framework. However, these approaches are only validated on a single embodiment and suffer from computational inefficiencies [54], which present challenges for real-time deployment in dynamic manipulation tasks.

VII. Limitation

We follow the model-based object pose tracking setting, which assumes that a CAD model is available for the object.

While assuming a CAD model may limit generalization in in-the-wild applications, it is a well-established assumption in industrial settings, such as warehouses or assembly lines [3, 54]. One potential direction to overcome this limitation is to simultaneously reconstruct the object and perform pose tracking, as demonstrated in methods like BundleSDF [69] and NeuralFeels [54], which offer promising and compatible ways to supply a model to our approach.

VIII. Conclusion

We introduced V-HOP, a visuo-haptic 6D object pose tracker that integrates a unified haptic representation and a visuo-haptic transformer. Our experiments demonstrate that V-HOP generalizes effectively to novel sensor types, embodiments, and objects, outperforming state-of-the-art visual and visuo-tactile approaches. Ablation studies highlight the critical role of both visual and haptic modalities in the framework. In the sim-to-real transfer experiments, V-HOP proved robust, delivering stable tracking under high occlusion and dynamic conditions. Furthermore, integrating V-HOP's real-time pose tracking into motion planning enabled accurate manipulation tasks, such as bimanual handover and insertion, showcasing its practical effectiveness.

Acknowledgments

This work is supported by the National Science Foundation (NSF) under CAREER grant #2143576, grant #2346528, and the Office of Naval Research (ONR) grant #N00014-22-1-259. We thank Ying Wang, Tao Lu, Zekun Li, and Xiaoyan Cong for their valuable discussions. We thank the area chair and the reviewers for providing constructive feedback on improving the quality and clarity of our paper. This research was conducted using computational resources and services at the Center for Computation and Visualization, Brown University.

References

[1] Peter K. Allen. Integrating Vision and Touch for Object Recognition Tasks. *The International Journal of Robotics Research*, 7(6):15–33, December 1988.

[2] Rareş Ambrus, Vitor Guizilini, Naveen Kuppuswamy, Andrew Beaulieu, Adrien Gaidon, and Alex Alspach. Monocular Depth Estimation for Soft Visuotactile Sensors. In *2021 IEEE 4th International Conference on Soft Robotics (RoboSoft)*, pages 643–649, April 2021.

[3] Maria Bauza, Oleguer Canal, and Alberto Rodriguez. Tactile Mapping and Localization from High-Resolution Tactile Imprints. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 3811–3817, May 2019.

[4] Samarth Brahmbhatt, Ankur Handa, James Hays, and Dieter Fox. ContactGrasp: Functional Multi-finger Grasp Synthesis from Contact. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2386–2393, November 2019.

[5] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M. Dollar. The YCB object and Model set: Towards common benchmarks for manipulation research. In *2015 International Conference on Advanced Robotics (ICAR)*, pages 510–517, July 2015.

[6] Ravinder S. Dahiya, Giorgio Metta, Maurizio Valle, and Giulio Sandini. Tactile Sensing—From Humans to Humanoids. *IEEE Transactions on Robotics*, 26(1):1–20, February 2010.

[7] Xinke Deng, Arsalan Mousavian, Yu Xiang, Fei Xia, Timothy Bretl, and Dieter Fox. PoseRBPF: A Rao–Blackwellized Particle Filter for 6-D Object Pose Tracking. *IEEE Transactions on Robotics*, 37(5):1328–1342, October 2021.

[8] Snehal Dikhale, Karankumar Patel, Daksh Dhingra, Itoshi Naramura, Akinobu Hayashi, Soshi Iba, and Nawid Jamali. VisuoTactile 6D Pose Estimation of an In-Hand Object Using Vision and Tactile Sensor Data. *IEEE Robotics and Automation Letters*, 7(2):2148–2155, April 2022.

[9] Runyu Ding, Yuzhe Qin, Jiyue Zhu, Chengzhe Jia, Shiqi Yang, Ruihan Yang, Xiaojuan Qi, and Xiaolong Wang. Bunny-VisionPro: Real-Time Bimanual Dexterous Teleoperation for Imitation Learning, July 2024.

[10] Elliott Donlon, Siyuan Dong, Melody Liu, Jianhua Li, Edward Adelson, and Alberto Rodriguez. GelSlim: A High-Resolution, Compact, Robust, and Calibrated Tactile-sensing Finger, May 2018.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021.

[12] Marc O. Ernst and Martin S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, January 2002.

[13] Pietro Falco, Shuang Lu, Andrea Cirillo, Ciro Natale, Salvatore Pirozzi, and Dongheui Lee. Cross-modal visuo-tactile object recognition using robotic active exploration. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5273–5280, May 2017.

[14] Yuan Gao, Shogo Matsuoka, Weiwei Wan, Takuya Kiyokawa, Keisuke Koyama, and Kensuke Harada. In-Hand Pose Estimation Using Hand-Mounted RGB Cameras and Visuotactile Sensors. *IEEE Access*, 11:17218–17232, 2023.

[15] A. M. Gordon, H. Forssberg, R. S. Johansson, and G. Westling. The integration of haptically acquired size information in the programming of precision grip. *Experimental Brain Research*, 83(3):483–488, February 1991.

[16] Irmak Guzey, Ben Evans, Soumith Chintala, and Lerrel Pinto. Dexterity from Touch: Self-Supervised Pre-Training of Tactile Representations with Robotic Play. In *Proceedings of The 7th Conference on Robot Learning*, pages 3142–3166. PMLR, December 2023.

[17] Sami Haddadin, Sven Parusel, Lars Johannsmeier, Saskia Golz, Simon Gabl, Florian Walch, Mohamadreza Sabaghian, Christoph Jähne, Lukas Hausperger, and Simon Haddadin. The Franka Emika Robot: A Reference Platform for Robotics Research and Education. *IEEE Robotics & Automation Magazine*, 29(2):46–64, June 2022.

[18] Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. OnePose++: Keypoint-Free One-Shot Object Pose Estimation without CAD Models. *Advances in Neural Information Processing Systems*, 35: 35103–35115, December 2022.

[19] Hannah B. Helbig and Marc O. Ernst. Optimal integration of shape information from vision and touch. *Experimental Brain Research*, 179(4):595–606, June 2007.

[20] Hannah B. Helbig, Marc O. Ernst, Emiliano Ricciardi, Pietro Pietrini, Axel Thielscher, Katja M. Mayer, Johannes Schultz, and Uta Noppeney. The neural mechanisms of reliability weighted integration of shape information from vision and touch. *NeuroImage*, 60(2): 1063–1072, April 2012.

[21] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes. In Kyoung Mu Lee, Yasuyuki Matsushita, James M. Rehg, and Zhanyi Hu, editors, *Computer Vision – ACCV 2012*, Lecture Notes in Computer Science, pages 548–562, Berlin, Heidelberg, 2013. Springer.

[22] Tomas Hodan, Martin Sundermeyer, Yann Labbe, Van Nguyen Nguyen, Gu Wang, Eric Brachmann, Bertram Drost, Vincent Lepetit, Carsten Rother, and Jiri Matas. BOP Challenge 2023 on Detection Segmentation and Pose Estimation of Seen and Unseen Rigid Objects. pages 5610–5619, 2024.

[23] Cheng-Chun Hsu, Bowen Wen, Jie Xu, Yashraj Narang, Xiaolong Wang, Yuke Zhu, Joydeep Biswas, and Stan Birchfield. SPOT: SE(3) Pose Trajectory Diffusion for Object-Centric Manipulation, November 2024.

[24] Tanja Kassuba, Corinna Klinge, Cordula Hölig, Brigitte Röder, and Hartwig R. Siebner. Vision holds a greater share in visuo-haptic object recognition than touch. *NeuroImage*, 65:59–68, January 2013.

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment Anything. pages 4015–4026, 2023.

[26] Naveen Kuppuswamy, Alejandro Castro, Calder Phillips-Grafflin, Alex Alspach, and Russ Tedrake. Fast Model-Based Contact Patch and Pose Estimation for Highly Deformable Dense-Geometry Tactile Sensors. *IEEE Robotics and Automation Letters*, 5(2):1811–1818, April 2020.

[27] Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. MegaPose: 6D Pose Estimation of Novel Objects via Render & Compare. August 2022.

[28] Simon Lacey and K. Sathian. Chapter 7 - Visuo-haptic object perception. In K. Sathian and V. S. Ramachandran, editors, *Multisensory Perception*, pages 157–178. Academic Press, January 2020.

[29] Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, Dinesh Jayaraman, and Roberto 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*, 5(3):3838–3845, July 2020.

[30] Taeyeop Lee, Jonathan Tremblay, Valts Blukis, Bowen Wen, Byeong-Uk Lee, Inkyu Shin, Stan Birchfield, In So Kweon, and Kuk-Jin Yoon. TTA-COPE: Test-Time Adaptation for Category-Level Object Pose Estimation. pages 21285–21295, 2023.

[31] Albert H. Li, Preston Culbertson, Vince Kurtz, and Aaron D. Ames. DROP: Dexterous Reorientation via Online Planning, October 2024.

[32] Hongyu Li, Snehal Dikhale, Soshi Iba, and Nawid Jamali. ViHOPE: Visuotactile In-Hand Object 6D Pose Estimation With Shape Completion. *IEEE Robotics and Automation Letters*, 8(11):6963–6970, November 2023.

[33] Hongyu Li, Snehal Dikhale, Jinda Cui, Soshi Iba, and Nawid Jamali. HyperTaxel: Hyper-Resolution for Taxel-Based Tactile Signals Through Contrastive Learning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7499–7506, October 2024.

[34] Hongyu Li, James Akl, Srinath Sridhar, Tye Brady, and Taskin Padir. ViTa-Zero: Zero-shot Visuotactile Object 6D Pose Estimation, April 2025.

[35] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. DeepIM: Deep Iterative Matching for 6D Pose Estimation. pages 683–698, 2018.

[36] Yuelong Li, Yafei Mao, Raja Bala, and Sunil Hadap. MRC-Net: 6-DoF Pose Estimation with MultiScale Residual Correlation. pages 10476–10486, 2024.

[37] Jiehong Lin, Lihua Liu, Dekun Lu, and Kui Jia. SAM-6D: Segment Anything Model Meets Zero-Shot 6D Object Pose Estimation. pages 27906–27916, 2024.

[38] Toru Lin, Yu Zhang, Qiyang Li, Haozhi Qi, Brent Yi, Sergey Levine, and Jitendra Malik. Learning Visuotactile Skills with Two Multifingered Hands, April 2024.

[39] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, December 2023.

[40] Yuan Liu, Yilin Wen, Sida Peng, Cheng Lin, Xiaoxiao Long, Taku Komura, and Wenping Wang. Gen6D: Generalizable Model-Free 6-DoF Object Pose Estimation from RGB Images. In Shai Avidan, Gabriel Brostow,

Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 298–315, Cham, 2022. Springer Nature Switzerland.

[41] Yun Liu, Xiaomeng Xu, Weihang Chen, Haocheng Yuan, He Wang, Jing Xu, Rui Chen, and Li Yi. Enhancing Generalizable 6D Pose Tracking of an In-Hand Object with Tactile Sensing. *IEEE Robotics and Automation Letters*, 9(2):1106–1113, February 2024.

[42] Jack M. Loomis and Susan J. Lederman. Tactual perception. In *Handbook of perception and human performance, Vol. 2: Cognitive processes and performance*, pages 1–41. John Wiley & Sons, Oxford, England, 1986.

[43] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. September 2018.

[44] E. Marchand, F. Spindler, and F. Chaumette. ViSP for visual servoing: a generic software platform with a wide class of robot control skills. *IEEE Robotics & Automation Magazine*, 12(4):40–52, December 2005.

[45] Luis Felipe Casas Murrilo, Ninad Khargonkar, Balakrishnan Prabhakaran, and Yu Xiang. MultiGripperGrasp: A Dataset for Robotic Grasping from Parallel Jaw Grippers to Dexterous Hands, March 2024.

[46] Nicolás Navarro-Guerrero, Sibel Toprak, Josip Josifovski, and Lorenzo Jamone. Visuo-haptic object perception for robots: an overview. *Autonomous Robots*, 47 (4):377–403, April 2023.

[47] Van Nguyen Nguyen, Thibault Groueix, Georgy Ponimatkin, Vincent Lepetit, and Tomas Hodan. CNOS: A Strong Baseline for CAD-Based Novel Object Segmentation. pages 2134–2140, 2023.

[48] Kiru Park, Timothy Patten, and Markus Vincze. Pix2Pose: Pixel-Wise Coordinate Regression of Objects for 6D Pose Estimation. pages 7668–7677, 2019.

[49] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[50] Alireza Rezazadeh, Snehal Dikhale, Soshi Iba, and Nawid Jamali. Hierarchical Graph Neural Networks for Proprioceptive 6D Pose Estimation of In-hand Objects. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2884–2890, May 2023.

[51] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2):336–359, February 2020.

[52] Sudharshan Suresh, Zilin Si, Joshua G. Mangelson, Wenzhen Yuan, and Michael Kaess. ShapeMap 3-D: Efficient shape mapping through dense touch and vision. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 7073–7080, May 2022.

[53] Sudharshan Suresh, Zilin Si, Stuart Anderson, Michael Kaess, and Mustafa Mukadam. MidasTouch: Monte-Carlo inference over distributions across sliding touch. In *Proceedings of The 6th Conference on Robot Learning*, pages 319–331. PMLR, March 2023.

[54] Sudharshan Suresh, Haozhi Qi, Tingfan Wu, Taosha Fan, Luis Pineda, Mike Lambeta, Jitendra Malik, Mrinal Kalakrishnan, Roberto Calandra, Michael Kaess, Joseph Ortiz, and Mustafa Mukadam. NeuralFeels with neural fields: Visuotactile perception for in-hand manipulation. *Science Robotics*, November 2024.

[55] Chie Takahashi and Simon Justin Watt. Visual-haptic integration with pliers and tongs: signal "weights" take account of changes in haptic sensitivity caused by different tools. *Frontiers in Psychology*, 5, February 2014.

[56] Ian H. Taylor, Siyuan Dong, and Alberto Rodriguez. GelSlim 3.0: High-Resolution Measurement of Shape, Force and Slip in a Compact Tactile-Sensing Finger. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 10781–10787, May 2022.

[57] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep Object Pose Estimation for Semantic Robotic Grasping of Household Objects. In *Proceedings of The 2nd Conference on Robot Learning*, pages 306–316. PMLR, October 2018.

[58] Jonathan Tremblay, Bowen Wen, Valts Blukis, Balakumar Sundaralingam, Stephen Tyree, and Stan Birchfield. Diff-DOPE: Differentiable Deep Object Pose Estimation, September 2023.

[59] Yuyang Tu, Junnan Jiang, Shuang Li, Norman Hendrich, Miao Li, and Jianwei Zhang. PoseFusion: Robust Object-in-Hand Pose Estimation with SelectLSTM. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6839–6846, October 2023.

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[61] Zhaoliang Wan, Yonggen Ling, Senlin Yi, Lu Qi, Wang Wei Lee, Minglei Lu, Sicheng Yang, Xiao Teng, Peng Lu, Xu Yang, Ming-Hsuan Yang, and Hui Cheng. VinT-6D: A Large-Scale Object-in-hand Dataset from Vision, Touch and Proprioception. In *Proceedings of the 41st International Conference on Machine Learning*, pages 49921–49940. PMLR, July 2024.

[62] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martin-Martin, Cewu Lu, Li Fei-Fei, and Silvio Savarese. DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion. pages 3343–3352, 2019.

[63] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized Object Coordinate Space for Category-Level 6D Object Pose and Size Estimation. pages 2642–2651, 2019.

[64] David Watkins-Valls, Jacob Varley, and Peter Allen. Multi-Modal Geometric Learning for Grasping and Ma-

nipulation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7339–7345, May 2019.

[65] Bowen Wen and Kostas Bekris. BundleTrack: 6D Pose Tracking for Novel Objects without Instance or Category-Level 3D Models, August 2021.

[66] Bowen Wen, Chaitanya Mitash, Baozhang Ren, and Kostas E. Bekris. se(3)-TrackNet: Data-driven 6D Pose Tracking by Calibrating Image Residuals in Synthetic Domains. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10367–10373, October 2020.

[67] Bowen Wen, Chaitanya Mitash, Sruthi Soorian, Andrew Kimmel, Avishai Sintov, and Kostas E. Bekris. Robust, Occlusion-aware Pose Estimation for Objects Grasped by Adaptive Hands. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6210–6217, May 2020.

[68] Bowen Wen, Wenzhao Lian, Kostas Bekris, and Stefan Schaal. You Only Demonstrate Once: Category-Level Manipulation from Single Visual Demonstration, May 2022.

[69] Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Muller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. BundleSDF: Neural 6-DoF Tracking and 3D Reconstruction of Unknown Objects, March 2023.

[70] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects. In *CVPR*, 2024.

[71] Bing Wu, Qian Liu, and Qiang Zhang. Tactile Pattern Super Resolution with Taxel-based Sensors. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3644–3650, October 2022.

[72] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. In *RSS*, volume 14, June 2018.

[73] Zhengrong Xue, Han Zhang, Jingwen Cheng, Zhengmao He, Yuanchen Ju, Changyi Lin, Gu Zhang, and Huazhe Xu. ArrayBot: Reinforcement Learning for Generalizable Distributed Manipulation through Touch, June 2023.

[74] Akihiko Yamaguchi and Christopher G. Atkeson. Recent progress in tactile sensing and sensors for robotic manipulation: can we turn tactile sensing into vision? *Advanced Robotics*, 33(14):661–673, July 2019.

[75] Linhan Yang, Bidan Huang, Qingbiao Li, Ya-Yen Tsai, Wang Wei Lee, Chaoyang Song, and Jia Pan. TacGNN: Learning Tactile-Based In-Hand Manipulation With a Blind Robot Using Hierarchical Graph Neural Network. *IEEE Robotics and Automation Letters*, 8(6):3605–3612, June 2023.

[76] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. CPF: Learning a Contact Potential Field to Model the Hand-Object Interaction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11077–11086, October 2021.

[77] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Junming Zhang, Jiefeng Li, and Cewu Lu. Learning a Contact Potential Field for Modeling the Hand-Object Interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5645–5662, August 2024.

[78] Zhao-Heng Yin, Binghao Huang, Yuzhe Qin, Qifeng Chen, and Xiaolong Wang. Rotating without Seeing: Towards In-hand Dexterity through Touch. volume 19, July 2023.

[79] Wenzhen Yuan, Siyuan Dong, and Edward H. Adelson. GelSight: High-Resolution Robot Tactile Sensors for Estimating Geometry and Force. *Sensors*, 17(12):2762, December 2017.

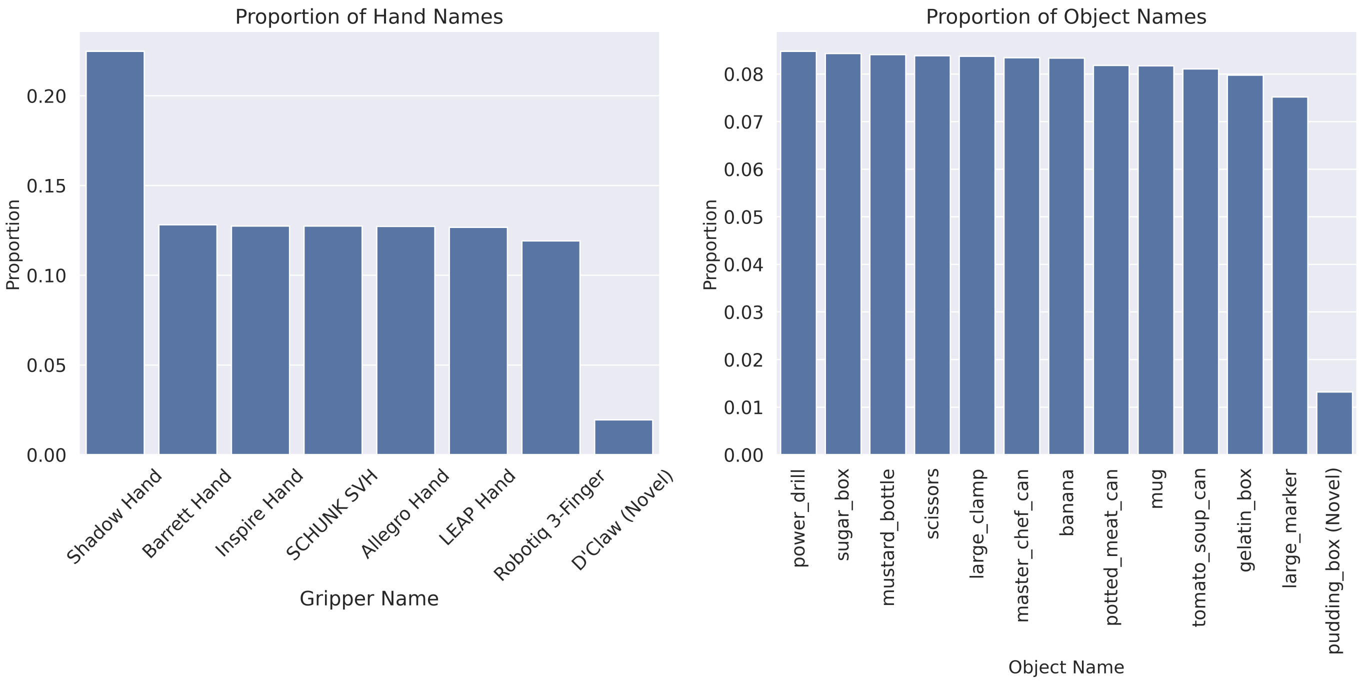


Fig. 10: **Distributions of embodiments and objects.** The novel gripper and object have fewer samples as they are only used for evaluation and not during training.

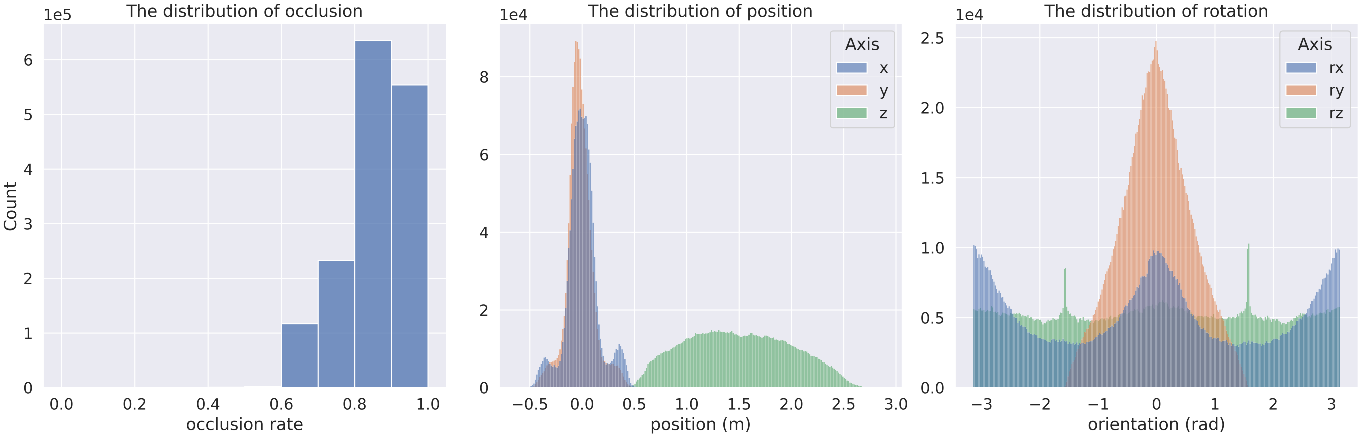


Fig. 11: **Dataset distributions**. We view the occlusion rate, position, and rotation distribution of our data samples.

Appendix A
Dataset

In Fig. 11, we provide a visualization of the dataset's distribution. The dataset focuses on in-hand object pose tracking and addresses challenges such as heavy occlusion during in-hand manipulation. Our visualization reveals that the positions and rotations of the data samples are well-distributed, following either normal or uniform distributions, ensuring a comprehensive evaluation of pose tracking performance.

Appendix B
Experiments

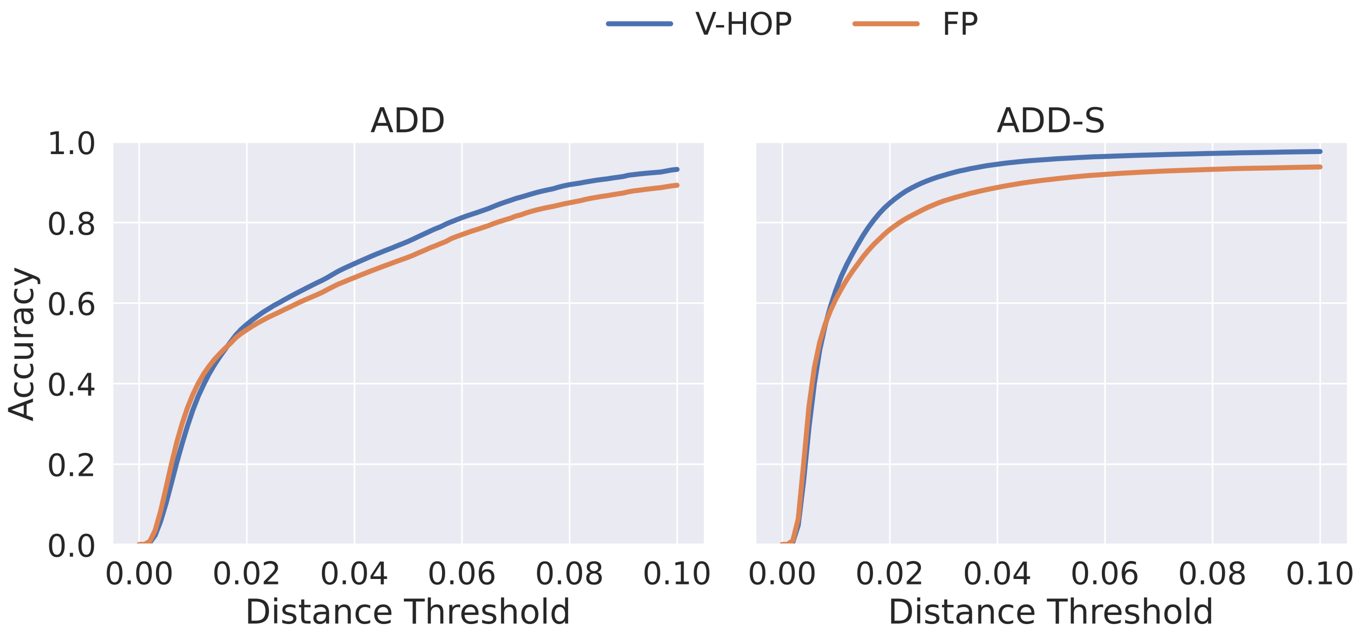


Fig. 12: **Accuracy-threshold curve [72] on our dataset.** V-HOP consistently demonstrates stronger or similar performance as FoundationPose (FP) under various thresholds.

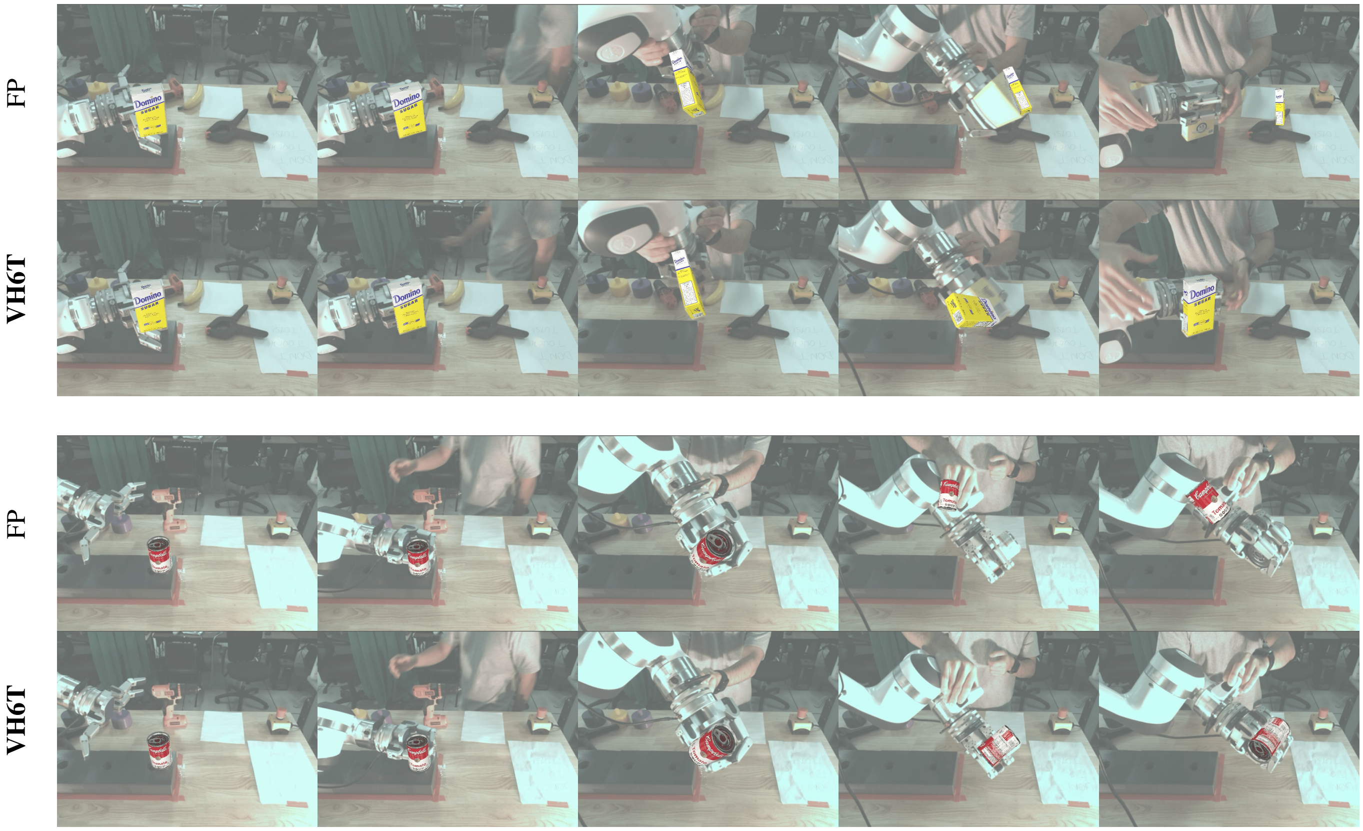


Fig. 13: **Qualitative results of pose tracking sequences**. We perform qualitative comparisons on more objects. Our results demonstrate that V-HOP consistently outperforms FP by a large margin.

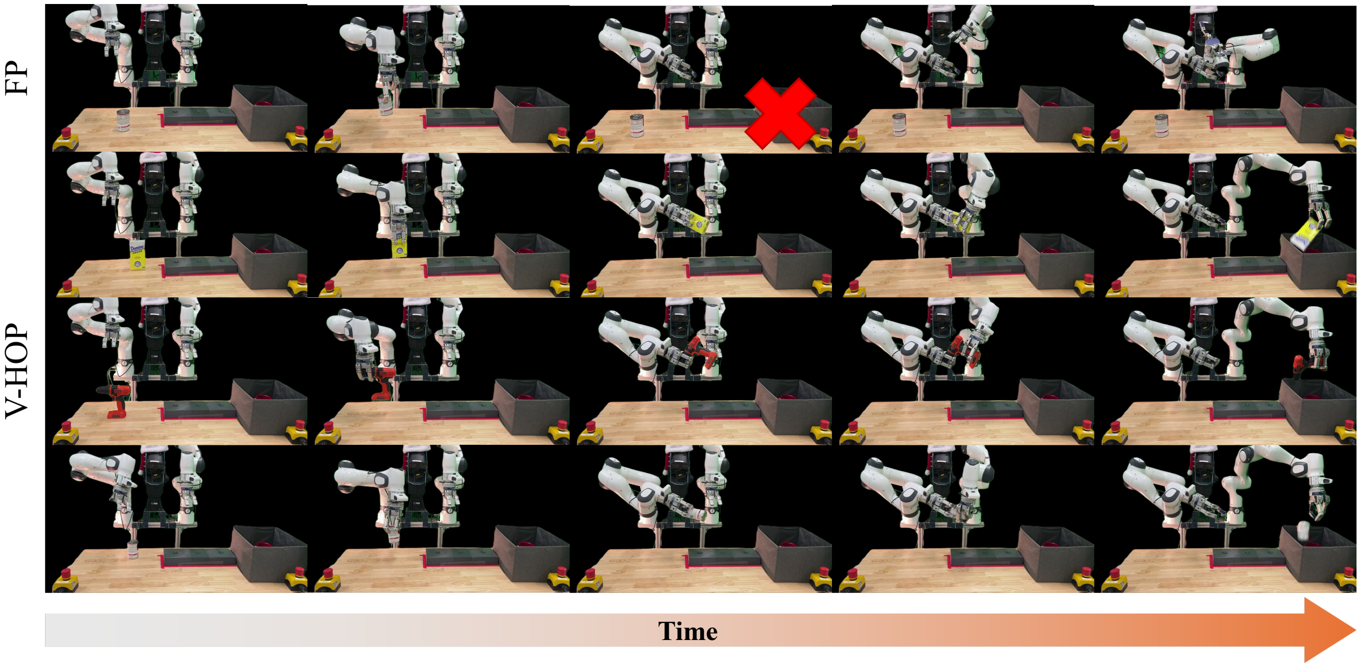


Fig. 14: **Bimanual handover experiments.** We perform bimanual handover experiments on more objects.