

# The MOTIF Hand: A Robotic Hand for Multimodal Observations with Thermal, Inertial, and Force Sensors

Hanyang Zhou\*, Haozhe Lou\*, Wenhao Liu\*, Enyu Zhao,  
Yue Wang, and Daniel Seita (\* equal contribution)

Department of Computer Science, Viterbi School of Engineering  
University of Southern California, Los Angeles CA 90089, USA,  
`{zhouhany, haozhelo, liuwenha, enyuzhao, yue.w, seita}@usc.edu`

**Abstract.** Advancing dexterous manipulation with multi-fingered robotic hands requires rich sensory capabilities, while existing designs lack on-board thermal and torque sensing. In this work, we propose the MOTIF hand, a novel multimodal and versatile robotic hand that extends the LEAP hand by integrating: (i) dense tactile information across the fingers and palm, (ii) a depth sensor, (iii) a thermal camera, (iv), IMU sensors, and (v) a visual sensor. The MOTIF hand is designed to be relatively low-cost (under 4000 USD) and easily reproducible. We validate our hand design through experiments that leverage its multimodal sensing for two representative tasks. First, we integrate thermal sensing into 3D reconstruction to guide temperature-aware, safe grasping. Second, we show how our hand can distinguish objects with identical appearance but different masses—a capability beyond methods that use vision only. For more details, please check our website: <https://slurm-lab-usc.github.io/motif-hand>

**Keywords:** dexterous hands, multimodal sensing, robot manipulation

## 1 Introduction and Motivation

Dexterous manipulation using multi-fingered hands has advanced rapidly, driven by innovations in both hardware design [19, 23] and algorithms for control [11, 29]. Research in this area predominantly uses third-person or wrist cameras for perception, along with tactile sensors such as XELA uSkin sensors [4] or lower-cost alternatives like force-sensing resistors [27, 29]. However, these sensing methods are insufficient in more complex real-world tasks that require reasoning about multiple sensing modalities. Onboard depth and force sensing could assist in tasks such as reaching into a bag or box, while thermal sensing is essential for cooking or safe human-robot interaction, yet current robotic hands lack these capabilities. While prior work has used thermal cameras to perceive liquids in robot manipulation [20, 21], no existing robotic hand integrates a thermal camera into a multi-fingered dexterous platform.

To address these limitations, we introduce the MOTIF hand (**M**ultimodal **O**bervation with **T**hermal, **I**nertial, and **F**orce sensors). This hand is built on

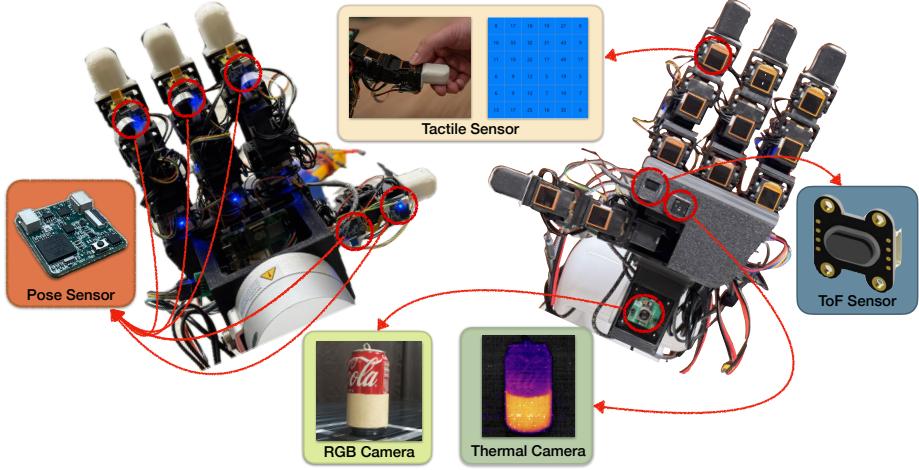


Fig. 1: The proposed MOTIF hand, where we show the back (left) and front (right). We label the key components of the MOTIF hand and show an example of data from the RGB and thermal cameras (using the can). See Section 2 for more details.

the LEAP hand [23] and enhances its design with dense force and tactile sensing, along with onboard depth and thermal sensors (see Fig. 1). By integrating these diverse sensing modalities, we believe the MOTIF hand is **one of the most comprehensive multimodal hands available**. We will open-source our hand design and continually improve it to benefit the robotics community.

### 1.1 Related Work

Prior work on dexterous manipulation with multi-fingered robotic hands falls into two broad categories: software-driven approaches and hardware innovations. Software-based approaches include learning-based algorithms for tasks such as in-hand reorientation using tactile-only feedback [13, 27] or visuo-tactile fusion [14, 29]. Other works study tactile-driven manipulation using high-quality sensors on the hand [4]. Hardware innovations focus on improved tactile-based sensors. For example, modular magnetic sensors such as ReSkin [1] and AnySkin [2] offer low-cost and replaceable sensing. Optical sensors such as the GelSight [28] and Digit 360 [8] provide high-resolution touch feedback while pressure-based sensors convert mechanical force into electrical signals [6]. In contrast to studies on individual sensors, we study full-hand designs, of which the LEAP hand [23] has become a widely-adopted platform in the robot manipulation community. Other recent hands include a five-finger variation of the LEAP hand, tested on dexterous cable manipulation [30] and the EyeSight hand [19], a 7-DoF humanoid hand that incorporates dense vision-based tactile sensors. However, no prior work has combined vision, depth, tactile, torque, and thermal sensing into a single hand, as we do with the MOTIF hand.

## 2 Technical Approach: The MOTIF Hand

We design our novel multisensory hand on top of the LEAP hand [23]. The MOTIF hand includes thin film tactile sensors covering the fingers, a depth sensor, a thermal camera, and a high-resolution RGB camera on the wrist, as shown in Fig. 1. We position accelerometers, gyroscopes, and magnetoresistive sensors at each finger joint and on the back of the hand to provide acceleration and orientation information.

For the finger pads, we use mature commercial tactile sensors for tactile perception input. The basic principle involves sensing force magnitude by measuring analog voltage signals in corresponding sampling areas of special materials with pressure-variable resistance values through array scanning. Our design provides sampling precision with six rows and six columns, each sub-unit having a resolution of 2.5mm. The trigger value for each sampling area is 20 g. While we include finger sensors, we do not use *fingertip* sensors, as there are relatively mature and modular sensors such as the Digit 360 [8] which we can use in a future version of the MOTIF hand.

The MOTIF hand uses three processing tiers. On the **first-tier** processing unit at each finger joint, we install a motion tracking unit that combines a 3-axis gyroscope and a 3-axis accelerometer. Finger joint movement features short distances, high instantaneous acceleration, and high degrees of freedom. Therefore, using traditional six-axis pose sensors for integration produces significant drift. We adopt a nine-axis sensor solution widely used in drone applications, employing the BMM350 3-axis magnetic sensor. We also use high-shielding conductive adhesive tape to isolate electromagnetic interference from the servomotors to the magnetic sensors.

Each first-tier processing unit on the finger joints connects via an RS485 serial bus and Modbus Protocol to the **second-tier** module on the palm responsible for data integration, encoding, and transmission. Similarly, we install nine-axis sensing units for pose calculation on this circuit board. To collect signals over a larger hand area, the MOTIF hand uses four centrally symmetric Inertial Measurement Units (IMUs) for multi-point sampling fusion. Tactile data, gyroscope data, and acceleration data from the palm and from 11 finger joints encode and transmit through UART via USB-C port to the host computer.

For our **third-tier** data processing unit, we directly use a Raspberry Pi 5 which provides increased computational power for integrating multimodal sensory data. FLIR Lepton 3.5 collects infrared temperature data and a small time of flight (ToF) sensor provides real-time distance from the palm to contact surfaces. Infrared temperature data initially collects at  $160 \times 120$  resolution and outputs at  $1080 \times 960$  resolution after interpolation. For temperature sensing, we can either directly use the resulting temperature value matrices or assign colors based on relative temperatures to render a thermal image. We use the Raspberry Pi Camera Module 2 on the wrist to provide RGB image data at 30fps.

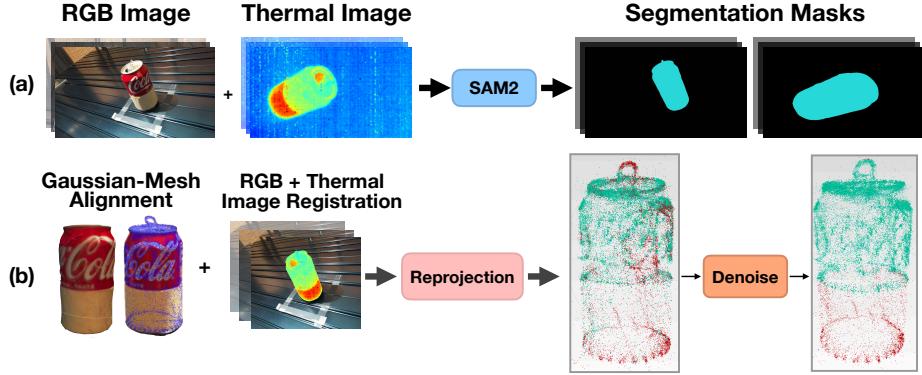


Fig. 2: Data processing pipeline for thermal-based grasping. (a) First, we collect images from diverse viewpoints of the object and use SAM2 [18] to extract the object mask. (b) We then reconstruct the 3D mesh and point cloud, perform thermal-RGB data alignment, and do reprojection. After denoising, we obtain our thermal affordance map which we pass to simulation for grasp pose optimization. See Section 3.1 for details.

### 3 Experiments

We perform two sets of experiments to demonstrate the capabilities and versatility of the MOTIF hand. First, in Section 3.1, we evaluate how its thermal sensors enable safe grasping of objects with high-temperature regions. Then, in Section 3.2, we investigate how its force-based sensors can distinguish objects with different mass but the same shape through a simple “flick” action.

#### 3.1 Temperature-Aware Object Modeling for Safe Grasping

In this experiment, we present a representative example of how the MOTIF hand’s multimodal sensing can inform grasping. We implement a “Real2Sim” pipeline that uses sensor data to build a geometric and thermal representation of an object for use in a physics simulator. It consists of reconstruction, alignment, denoising, and training a grasp policy. See Figures 2 and 3 for an overview.

**Reconstruction** This phase consists of three primary steps: camera pose estimation, dense reconstruction, and mesh extraction. We begin by moving the robot’s end-effector to capture approximately 90 images of the target object from various viewpoints. Using Structure-from-Motion (SfM) [22], we obtain a sparse 3D point cloud along with estimated camera poses. Subsequently, we apply Gaussian Splatting [5, 7] to produce a dense reconstruction of the scene that captures fine geometric details and color variations. Utilizing the resulting dense Gaussian splats (3DGS) combined with normal constraints [9, 26], we extract a refined mesh that accurately represents the object’s surface geometry. This mesh can be exported to a physics simulation engine such as MuJoCo [25].

**Alignment** We next align the reconstructed simulation scene with the physical environment [9], with a focus on synchronizing thermal and RGB images

to ensure scale and spatial consistency. First, we employ the Scale-Invariant Feature Transform (SIFT) algorithm [10] to establish correspondences between thermal and RGB camera images. After identifying these correspondences, we apply a reprojection procedure to map values from the thermal images onto the reconstructed 3D points using the dense 3D Gaussian Splat representation. Let  ${}^w T_c \in SE(3)$  denote the camera pose in the world coordinate system, and let  $\mathbf{K} \in \mathbb{R}^{3 \times 3}$  represent the camera intrinsic matrix. The transformation from pixel-level thermal coordinates  $\mathbf{x}_{2D}$  to 3D coordinates  $\mathbf{x}_{3D}$  is:

$$\mathbf{x}_{3D} = {}^w T_c^{-1} \left( d \cdot \mathbf{K}^{-1} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \right) \quad (1)$$

where:  $\mathbf{x}_{2D} = [x, y]^T$  is the pixel coordinate in the thermal image and  $d$  is its associated depth. Once we establish the correspondence between the point cloud and pixel coordinates, thermal values for each pixel can be assigned to the respective points in the 3D Gaussian Splats [31]. The result is a static 3D thermal affordance map, which highlights thermal distribution across the object’s surface.

**Denoising** Noise from low-resolution thermal images is inevitable. As shown in Figure 2, the Coke can’s top hole introduces spurious hot signals during reprojection. To mitigate this, we assume a planar boundary separates the liquid versus air regions inside the object. We identify this boundary by analyzing thermal gradients through cross-sectional slicing. We compute “thermal values” using a weighted sum of RGB values, with a higher coefficient for red values since high-temperature pixels have stronger red responses.

We then search for maximum negative transitions from positive thermal values (undersurface, warm) to negative values (above-surface, cool). The priority goes to transitions where the upper layer exceeds a threshold, the lower layer falls below another threshold, and the jump exceeds 30 units. Otherwise, we select the largest negative transition among adjacent layers. This approach relies on expecting significant temperature gradients on water surfaces. Using the identified boundary, we analyze point color values in both directions to determine dominant color patterns. Anomalous points, such as hot spots above water and cold spots below, are treated as noise and replaced with the dominant colors for each layer. This removes artifacts while preserving legitimate thermal features.

**Grasping Policy Conditioned on Thermal Observations** Given the 3D thermal affordance map and the Real2Sim alignment, we develop a grasping policy explicitly conditioned on thermal observations. The policy aims to avoid grasping points that could lead to interactions with thermally sensitive or potentially hazardous areas. We utilize imitation learning based on human demonstrations [16] and employ MuJoCo simulation [25] to infer contact regions between the robotic grasping poses and the object. Specifically, we extract 5000 demonstration frames from 12 episodes, retarget them to our MOTIF hand, and

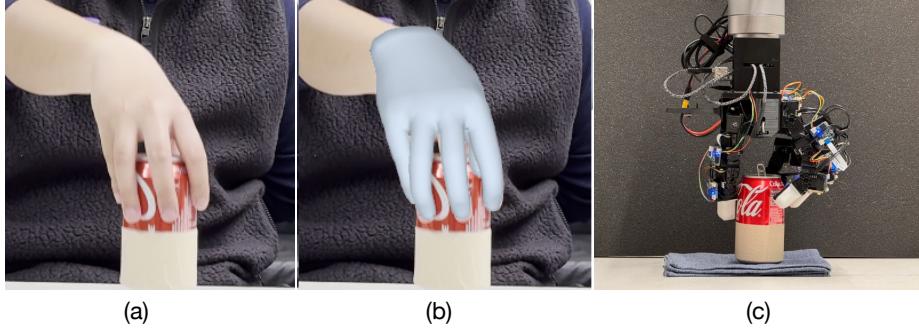


Fig. 3: Hand pose and grasping demonstration pipeline. (a) A human hand performing one episode of demonstration which shows a natural grasping motion on a cylindrical object, avoiding the high-temperature region. (b) Pose extraction from the human demonstration. (c) Robotic hand trained to replicate the demonstrated grasping behavior, successfully learning to avoid the hazardous tape area like humans.

subsequently use these frames to train the policy [12, 17]. If the robotic hand encounters a high-temperature region on the object, the policy adjusts by shifting the robot’s end-effector grasping pose away from the hazardous region.

Grasp poses are filtered based on proximity to high-temperature zones on the object’s thermal map. Only grasping configurations that avoid these heated regions are selected for execution. Finally, validated grasps are performed in the real-world robotic manipulation setup, ensuring safe and efficient interactions with the object based on its thermal characteristics.

### 3.2 Multimodal Sensor Object Mass Classification via Flicks

Next, we conduct an experiment to test how the MOTIF hand’s multimodal accelerometer, gyroscope, and magnetometer sensors can help distinguish objects of different masses via a simple fingertip “flick” action.

**Experiment** For humans, a simple way to distinguish the weights of different objects is to perform a short-distance movement, such as gently flicking them with the fingertips. Thus, we design an experiment where the hand uses its index fingertip to strike three U-shaped objects of the same material but of different weights, applying the same torque. As shown in Fig. 4(a), the lightest red object (82 g) has the highest displacement, the medium-weight blue object (125 g) has a slightly lower displacement than the first one, and the heaviest purple object (219 g) has the shortest displacement.

**Data Collection** The act of flicking objects with varying physical properties results in measurably different hand motion signatures, as recorded by sensors (accelerometer, gyroscope, and magnetometer) across their three-axis measurement frameworks. Here, we focus on the measurements from the first board near the fingertip which is at the point of contact. To enable real-time data acquisition, the system transmits data at an update interval of 2 ms, with a Baud

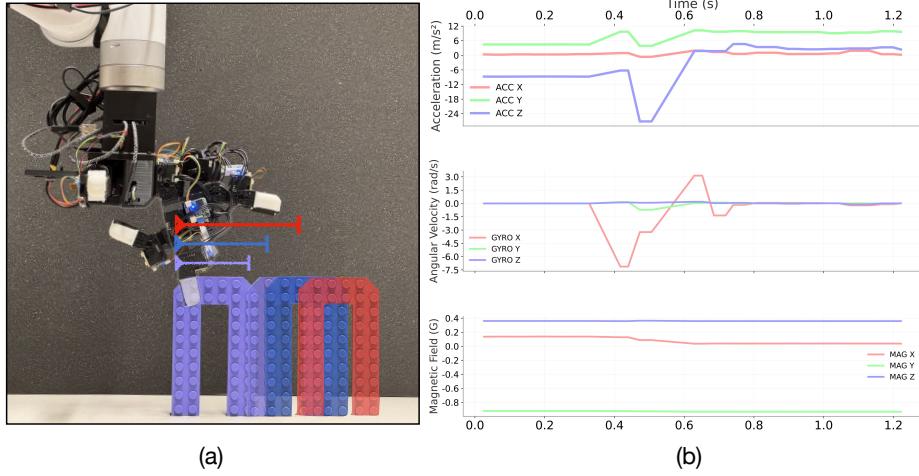


Fig. 4: Fingertip flicking experiment setup and sensor data acquisition. (a) A visualization of how the MOTIF hand’s flick action affects the configuration of three U-shaped objects of different masses: purple (219 g), blue (125 g), and red (82 g). (b) Representative sensor data from the hand’s accelerometer, gyroscope, and magnetometer during a single flicking trial, captured from 0.5 seconds before to 1 seconds after the flick action.

Rate of 115200 Bits/s. Data recording starts 0.25 seconds before performing the flick, and continues for another 1.0 seconds (see Fig. 4(b)). We conducted 50 experiments on each object.

**Classification Results and Analysis** To validate the discriminative capability of the collected IMU sensor data, we performed Linear Discriminant Analysis (LDA) [3] on the sensor features from the flicking experiments. From the raw time-series data collected during each flicking motion, we extracted 42 statistical features to capture the essential characteristics. These features were derived from the sensors as follows: for the axis (X, Y, Z) of the accelerometer, gyroscope, and magnetometer, we computed four statistical measures: minimum, maximum, mean, and standard deviation. Additionally, we calculated the range (max-min) and magnitude mean (average of the three-axis combined magnitude) for each sensor type. This provides 14 features per sensor (3 axes × 4 statistics + 1 range + 1 magnitude), totaling 42 features across all three sensors ( $14 \times 3 = 42$ ).

The results, shown in Fig. 5, demonstrate strong separability among the three classes. The first discriminant direction (LD1) explains 77.5% of the inter-class variance, while the second discriminant direction (LD2) accounts for the remaining 22.5%, together capturing 100% of the discriminative information. As shown in Fig. 5(a), the three object categories form distinct clusters in the LDA feature space, with minimal overlap between their 95% confidence ellipses.

The feature contribution analysis in Fig. 5(b) reveals physically meaningful patterns that align with the biomechanics of the fingertip flicking motion. The primary discriminant direction is mainly influenced by acceleration-related features, particularly ACC\_Z\_Min (49.48) and ACC\_Range (47.98). This is bio-

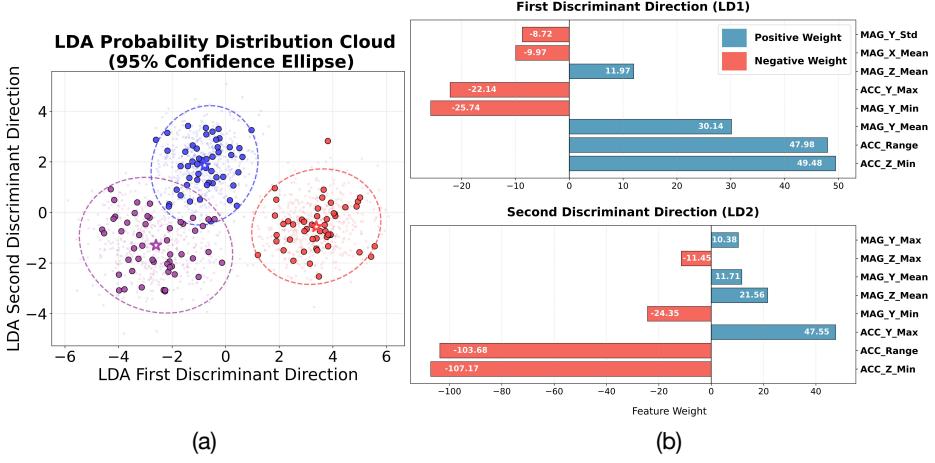


Fig. 5: LDA analysis of multimodal sensor data from fingertip flicking experiments. (a) Two-dimensional LDA feature space showing separation of three object classes (red: 82 g, blue: 125 g, purple: 219 g) within 95% confidence ellipses. (b) Feature contribution weights for the first and second discriminant directions, highlighting the dominant role of acceleration-based features in object mass classification.

mechanically reasonable because during the flick motion, the finger initially accelerates downward (negative Z-direction) to contact the object, then rapidly decelerates upon impact with an object of different weight. Heavier objects require greater contact force and create stronger resistance, resulting in more pronounced negative Z-acceleration values and larger acceleration ranges. The significant contribution of `MAG_Y_Mean` (30.14) reflects the lateral finger movement during the flicking gesture, as different object masses affect the finger's trajectory and stabilization requirements.

The second discriminant direction shows `ACC_Z_Min` (-107.17) and `ACC_Range` (-103.68) with strong negative contributions, while `ACC_Y_Max` (47.55) provides positive discrimination. This pattern captures the finger's recovery motion after contact—lighter objects allow for quicker finger retraction with higher upward acceleration (Y-direction), while heavier objects result in more prolonged contact and different rebound dynamics.

## 4 Experimental Insights

Multiple sensory signals, such as tactile, force/torque, and temperature, play critical roles in human skin sensing. Tactile sensing helps reveal an object's geometry and contact properties. Force sensing indicates the strength of interactions, enabling better control. Temperature sensing expands perception beyond touch, allowing us to infer thermal affordances and avoid damage from extreme temperatures. IMU sensors can support dynamic or closed-loop contact-rich tasks, such as in-hand manipulation and reorientation [15]. We are investigating the essential combination of sensory modalities necessary for robotic manipulation tasks

to ensure generalization across common indoor environments, such as kitchens. Therefore, we integrate visual, tactile, and proprioceptive sensors into our MOTIF hand. Compared to relying solely on vision, our multimodal system enhances a robot’s ability to detect, grasp, and manipulate objects, enabling more effective and safer interactions within complex real-world settings.

**Real-World Applications.** Manipulation in cooking environments is challenging due to the diverse and delicate foods, tools, and surroundings. The presence of hot fire, metal or wood cooking devices, and smooth-surfaced objects make it difficult to deploy traditional vision-based manipulation policies. However, leveraging the high-frequency in-hand sensors of the MOTIF hand offer potential to capture underlying physical properties that are hard to determine from visual data alone. Beyond cooking, our hand can be valuable for industrial tasks such as welding and screw-tightening tasks in factories. A robotic welding system requires knowledge of when its welding gun tip is hot to complete the task. Additionally, screw-tightening demands high-precision positioning and real-time feedback of finger force and torque to achieve stable and accurate operations. By using the MOTIF hand, we aim to make progress towards these tasks.

**Future Developments.** More broadly, we envision that our MOTIF hand has potential to be the first of many future iterations for improved multimodal robotic hands. This mirrors prior progress in robotic hands. After the original LEAP hand [23] was developed, researchers extended its design, such as by adding a fifth finger [30]. Furthermore, the original research team recently proposed the second generation of the LEAP hand [24], which adds a soft exterior skin to create an anthropomorphic hand for robot learning. We plan to continually improve and expand our MOTIF hand with more features. For example, the fingertips do not have sensors, so we will integrate the Digit 360 sensors [8] on them, pending availability. We hope that this inspires future work in developing methods for multimodal sensory robotic dexterous hands for complex tasks.

## References

- [1] R. Bhirangi *et al.*, “ReSkin: versatile, replaceable, lasting tactile skins,” in *Conference on Robot Learning (CoRL)*, 2021.
- [2] R. Bhirangi *et al.*, “AnySkin: Plug-and-play Skin Sensing for Robotic Touch,” *arXiv preprint arXiv:2409.08276*, 2024.
- [3] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
- [4] I. Guzey *et al.*, “Dexterity from Touch: Self-Supervised Pre-Training of Tactile Representations with Robotic Play,” in *Conference on Robot Learning*, 2023.
- [5] B. Huang *et al.*, “2d gaussian splatting for geometrically accurate radiance fields,” in *ACM SIGGRAPH 2024 conference papers*, 2024, pp. 1–11.
- [6] B. Huang *et al.*, “3D ViTac: Learning Fine-Grained Manipulation with Visuo-Tactile Sensing,” in *Conference on Robot Learning (CoRL)*, 2024.
- [7] B. Kerbl *et al.*, “3d gaussian splatting for real-time radiance field rendering,” *ACM Trans. Graph.*, 2023.
- [8] M. Lambeta *et al.*, “Digitizing Touch with an Artificial Multimodal Fingertip,” *arXiv preprint arXiv:2411.02479*, 2024.

- [9] H. Lou *et al.*, “Robo-gs: A physics consistent spatial-temporal model for robotic arm with hybrid representation,” *arXiv preprint arXiv:2408.14873*, 2024.
- [10] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” *International Journal of Computer Vision (IJCV)*, 2004.
- [11] OpenAI *et al.*, “Solving Rubik’s Cube with a Robot Hand,” *arXiv preprint arXiv:1910.07113*, 2019.
- [12] G. Pavlakos *et al.*, “Reconstructing hands in 3D with transformers,” in *CVPR*, 2024.
- [13] H. Qi *et al.*, “In-Hand Object Rotation via Rapid Motor Adaptation,” in *Conference on Robot Learning (CoRL)*, 2022.
- [14] H. Qi *et al.*, “General In-Hand Object Rotation with Vision and Touch,” in *Conference on Robot Learning (CoRL)*, 2023.
- [15] H. Qi *et al.*, *From simple to complex skills: The case of in-hand object reorientation*, 2025. arXiv: 2501.05439 [cs.RO].
- [16] Y. Qin *et al.*, “Dexmv: Imitation learning for dexterous manipulation from human videos,” in *European Conference on Computer Vision (ECCV)*, 2022.
- [17] Y. Qin *et al.*, “Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system,” in *Robotics: Science and Systems*, 2023.
- [18] N. Ravi *et al.*, *Sam 2: Segment anything in images and videos*, 2024. arXiv: 2408.00714 [cs.CV].
- [19] B. Romero *et al.*, “Eyesight hand: Design of a fully-actuated dexterous robot hand with integrated vision-based tactile sensors and compliant actuation,” in *International Conference on Intelligent Robots and Systems (IROS)*, 2024.
- [20] C. Schenck and D. Fox, “Visual Closed-loop Control for Pouring Liquids,” in *International Conference on Robotics and Automation (ICRA)*, 2017.
- [21] C. Schenck and D. Fox, “Perceiving and Reasoning About Liquids Using Fully Convolutional Networks,” in *International Journal of Robotics Research*, 2017.
- [22] J. L. Schonberger and J.-M. Frahm, “Structure-from-Motion Revisited,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [23] K. Shaw, A. Agarwal, and D. Pathak, “LEAP Hand: Low-Cost, Efficient, and Anthropomorphic Hand for Robot Learning,” in *Robotics: Science and Systems (RSS)*, 2023.
- [24] K. Shaw and D. Pathak, “LEAP hand v2: Dexterous, low-cost anthropomorphic hybrid rigid soft hand for robot learning,” in *RSS Workshop*, 2024.
- [25] E. Todorov, T. Erez, and Y. Tassa, “MuJoCo: A Physics Engine for Model-Based Control,” in *International Conference on Intelligent Robots and Systems (IROS)*, 2012.
- [26] C. Ye *et al.*, “Stablenormal: Reducing diffusion variance for stable and sharp normal,” *ACM Transactions on Graphics (TOG)*, vol. 43, no. 6, pp. 1–18, 2024.
- [27] Z.-H. Yin *et al.*, “Rotating without Seeing: Towards In-hand Dexterity through Touch,” in *Robotics: Science and Systems (RSS)*, 2023.
- [28] W. Yuan, S. Dong, and E. H. Adelson, “GelSight: High-Resolution Robot Tactile Sensors for Estimating Geometry and Force,” *Sensors*, vol. 17, no. 12, 2017.
- [29] Y. Yuan *et al.*, “Robot Synesthesia: In-Hand Manipulation with Visuotactile Sensing,” in *International Conference on Robotics and Automation*, 2024.
- [30] S. Zhao *et al.*, “Dexterous Cable Manipulation: Taxonomy, Multi-Fingered Hand Design, and Long-Horizon Manipulation,” *arXiv preprint arXiv:2502.00396*, 2025.
- [31] S. Zhou *et al.*, *Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields*, 2024. arXiv: 2312.03203 [cs.CV].