Tracking Dexterous Hands: System Trade-offs for Scaling Robot Learning

Shumo Chu and GI Team
General Intelligence Labs
shumo@gilabs.xyz
gilabs.xyz

Abstract

Motion capture is essential for data-driven robotics, providing observations required for imitation learning, reinforcement learning from demonstrations, and sim-to-real transfer. Yet each sensing technology embodies trade-offs between accuracy, drift, environmental robustness, and infrastructure requirements. This article systematically examines motion capture technologies for robotics, with emphasis on dexterous manipulation where these trade-offs are especially consequential.

We organize tracking systems into three categories. *External global tracking systems* such as optical motion capture, Lighthouse-style tracking, and electromagnetic systems provide drift-free absolute pose but require infrastructure and suffer from occlusion or environmental sensitivity. *Onboard sensing* such as inertial measurement units and visual SLAM offers autonomy and scalability but accumulates drift. *Internal joint-state sensing* via magnetic encoders delivers precise articulation measurements without spatial awareness.

Because no single technology meets the demands of dexterous manipulation, we examine how tracking systems can be combined, discussing physical constraints on sensor compatibility, fusion architectures, and case studies of recent systems. Our goal is to equip practitioners with the knowledge to design motion capture systems matched to their robotic learning and deployment requirements.

1 Introduction

Data-driven robotics increasingly learns from demonstration data: imitation learning, demonstration-guided reinforcement learning, and sim-to-real transfer all depend on captured motion data whose quality directly determines which behaviors a robot can learn (Argall et al., 2009; Nair et al., 2018). For dexterous manipulation in particular, where success depends on precise coordination of end-effector pose, finger articulation, and contact dynamics, accurate motion capture is indispensable. Yet selecting an appropriate tracking system is far from straightforward. Optical systems achieve sub-millimeter accuracy but fail under occlusion; inertial sensors operate autonomously but accumulate drift; electromagnetic trackers ignore line-of-sight constraints but distort near metal; joint encoders provide precise articulation measurements but reveal nothing about global pose. These trade-offs become especially acute in dexterous manipulation, where hands routinely self-occlude, interact with diverse objects, and operate in uncontrolled environments.

This article provides a systematic examination of motion capture technologies for robotics. We organize sensing modalities into three categories:

- **External global tracking systems** (Section 3) estimate absolute 6-DoF pose in a fixed world coordinate frame using infrastructure-mounted sensors. These include passive and active optical motion capture and Lighthouse-style tracking.
- **Onboard and relative motion sensing** (Section 4) recovers motion using sensors carried by the robot itself, producing estimates relative to an initial state or incrementally built map. This category encompasses inertial measurement units, electromagnetic tracking, and SLAM-based approaches.
- **Internal joint-state sensing** (Section 5) measures the configuration of articulated mechanisms directly, providing joint angles rather than spatial pose. Magnetic encoders exemplify this tracking

Table 1: Comparison of motion capture sensing modalities for robotics. Accuracy values represent typical performance under favorable conditions; actual performance varies with hardware, calibration, and environmental factors. **LoS**: line-of-sight required. **Infra.**: infrastructure requirements. **Env. Sensitivity**: primary environmental factors affecting accuracy.

Tracking system	**Accuracy**	**Drift**	**LoS**	**Infra.**	**Env. Sensitivity**
Passive Optical	<1 mm	None	Yes	High	Lighting, occlusion
Active Optical	<1 mm	None	Yes	High	Lighting, occlusion
Lighthouse	1–10 mm	None	Yes	Med.	Reflections, geometry
IMU	0.1–2° (ori.)	Accumulates	No	None	Magnetic disturbances
Electromagnetic	∼0.5 mm	None	No	None	Metals (critical)
VI-SLAM[†]	cm-scale	Bounded	No	None	Texture, dynamics
Joint Encoders	0.1–0.4°	None	No	None	Minimal

[†]Consumer XR headsets (Meta Quest 3, Apple Vision Pro) employ VI-SLAM for headset localization. Quest 3 additionally uses active optical controller tracking (low-millimeter accuracy relative to the headset); Vision Pro relies on markerless hand tracking (sub-centimeter). See Section 6.

system.

Table 1 summarizes the key trade-offs. Modalities offering high absolute accuracy and drift-free measurements typically require external infrastructure, while autonomous, portable approaches sacrifice global consistency.

Beyond individual modalities, we examine how multiple sensing approaches can be combined to overcome the limitations of any single technology (Section 6).

2 What Does a Motion Capture System Measure?

Motion capture systems record how objects move and orient in three-dimensional space. The most fundamental measurement is the *pose* of a physical entity—its position and orientation—commonly represented as a rigid-body transformation with six degrees of freedom (6 DoF): three translational and three rotational components (Craig, 2018; Siciliano et al., 2009).

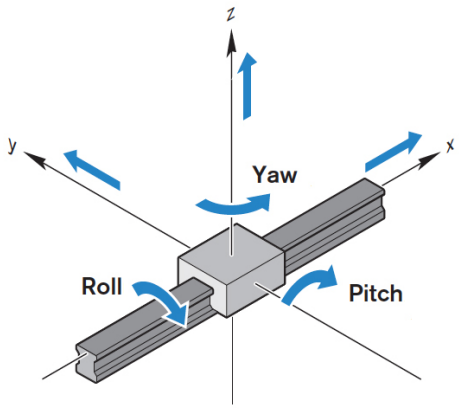


Figure 1: Illustration of a 6 DoF pose consisting of three translational components (x, y, z) and three rotational components (roll, pitch, yaw).

Position: Where is the Object. The position component specifies the location of a reference point on the object relative to a global reference frame, represented by three coordinates (x, y, z) along orthogonal spatial axes. These three translational degrees of freedom answer the question: *where is the object located in space?* (Craig, 2018)

(a) PhaseSpace motion capture suit equipped with approximately 50 active-pulse LED markers.

(b) Example of a passive infrared marker motion capture system (Noitom).

Figure 2: Examples of active and passive optical motion capture systems used in robotics.

Orientation: How an Object is Rotated. The orientation component describes how an object is rotated with respect to the global reference frame. Orientation is characterized by three rotational degrees of freedom and can be represented using Euler angles, rotation matrices, or unit quaternions (Siciliano et al., 2009; Murray et al., 1994). In robotics, orientation is critical because task success often depends not only on end-effector position but also on its alignment during interaction with objects (Craig, 2018).

6 DoF Pose vs. Joint Degrees of Freedom. A 6 DoF pose provides a complete spatial description of a rigid body at a given instant. Motion capture systems typically output time-indexed sequences of such poses, enabling trajectory reconstruction (Siciliano et al., 2009).

This notion of 6 DoF pose should be distinguished from degrees of freedom in robotic manipulators, which refer to independently actuated joints (Craig, 2018). The 6 DoF pose measured by a motion capture system describes rigid-body translation and rotation of a specific component—such as an end-effector or the palm of a dexterous hand—within a global coordinate frame, rather than enumerating joints or actuation.

3 External Global Tracking Systems

External global tracking systems estimate motion by observing objects directly within a fixed world coordinate frame. Unlike onboard or relative sensing approaches, these systems rely on environment-mounted optical infrastructure to provide absolute 6 DoF pose measurements that do not accumulate drift over time, making them a common reference for benchmarking, dataset generation, and evaluation in robotics.

3.1 Passive Infrared Marker Systems

Passive optical systems use high-speed infrared cameras with synchronized illumination to track retroreflective markers via multi-view triangulation, recovering drift-free rigid-body 6 DoF poses in a calibrated global frame (Vicon Motion Systems, 2026). Commercial platforms such as OptiTrack provide low-latency pose streams widely used as ground truth for state estimation, control, and sim-to-real validation (OptiTrack, 2026). High-end systems like the Noitom Perception Neuron Hybrid report sub-millimeter accuracy (0.08–0.10 mm) at frame rates exceeding 100 Hz (Noitom Ltd., 2026).

However, passive optical systems require line-of-sight visibility, making them sensitive to occlusion from self-occluding hands or dense object interaction. They also require a multi-camera cage setup with dedicated calibration and controlled lighting, making deployment labor-intensive and impractical for in-the-wild data collection. As a result, they are largely confined to instrumented laboratory environments for benchmarking and evaluation.

3.2 Active Marker Systems

Active optical motion capture systems replace passive reflective markers with powered light-emitting markers, often LEDs, that emit uniquely identifiable signals. Unlike passive systems that rely on reflective intensity, active markers broadcast a coded signal that allows the tracking software to immediately recognize and label each marker uniquely (PhaseSpace, 2026b). This inherent identification eliminates marker-swapping and reduces the need for post-processing or template-based labeling, which is a common source of error in passive optical systems.

PhaseSpace systems exemplify this approach at a high technical level. Their Impulse series uses high-resolution linear detector cameras capable of capturing active LED positions at up to 960 Hz with a native resolution of 3600×3600 pixels and extended sub-pixel resolution up to $36{,}000 \times 36{,}000$ (PhaseSpace, 2026a). By combining high spatial resolution with high frame rate, these systems support sub-millimeter precision tracking of multiple active LED markers and rigid bodies in real time. The active LED markers themselves each carry a unique identifier, which allows the system to maintain correspondence even when markers re-enter the field of view after occlusion.

The PhaseSpace architecture is highly configurable: Impulse systems can be configured with anywhere from a small number of cameras for compact capture volumes to dozens of cameras covering large spaces, enabling simultaneous tracking of many markers or subjects. In robotics contexts, this scalability supports dense multi-object tracking scenarios such as robot arms with tools, multiple mobile robots in a shared workspace, or human–robot interactions involving dexterous hand motions.

While active systems retain the benefits of global pose estimation and low drift, they introduce trade-offs related to marker power requirements, added mass or wiring for some marker designs, and a continued dependence on line-of-sight visibility to the tracking cameras. Additionally, the infrastructure cost and calibration effort remain non-trivial, although active marker designs can simplify data cleanup and labeling compared to passive systems.

Inside-out active optical tracking. The same active optical principle—cameras observing known LED constellations and solving pose via Perspective-n-Point (PnP)—has been adopted *inside-out* in consumer XR controller tracking. Meta Quest 3's Touch Plus controllers embed IR LEDs in known geometric patterns on the controller faceplate; the headset's four onboard cameras detect these LEDs and recover 6-DoF controller pose, fused with controller IMU data and a concurrent markerless hand model (Meta, 2023). This inverts the PhaseSpace architecture: cameras ride on the user's head rather than being mounted in the environment, eliminating fixed infrastructure at the cost of reduced observation geometry (fewer cameras, limited baseline). Controller tracking accuracy *relative to the headset* is estimated in the low-millimeter range; global accuracy remains bounded by the headset's VI-SLAM precision. Notably, Apple Vision Pro takes a fundamentally different approach—it ships without controllers entirely, relying on markerless hand tracking and eye gaze as primary inputs. The practical implications of these architectural differences for robotics data collection are discussed in Section 6.

3.3 Lighthouse-Style Optical Tracking Systems

Lighthouse-style optical tracking systems estimate pose using a timing-based optical sensing principle rather than camera-based observation. In these systems, fixed base stations mounted in the environment emit precisely timed infrared (IR) laser sweeps, while tracked objects passively sense the incoming signals using arrays of photodiodes distributed across their surface. The known spatial configuration of these photodiodes, combined with the timing of horizontal and vertical laser sweeps, enables reconstruction of the object's global 6 DoF pose without requiring cameras in the environment (University of Massachusetts Amherst Robotics Lab, 2026).

Popularized by the SteamVR Lighthouse system, this approach offers a lower-cost and lower-infrastructure alternative to multi-camera optical motion capture installations. Because tracked objects only receive signals and do not emit light, Lighthouse systems scale well to multiple devices and provide low-latency pose updates suitable for real-time interaction and teleoperation.

Quantitative evaluations show that Lighthouse tracking occupies an intermediate accuracy regime between high-end optical motion capture and onboard sensing. A recent study evaluating SteamVR Track-

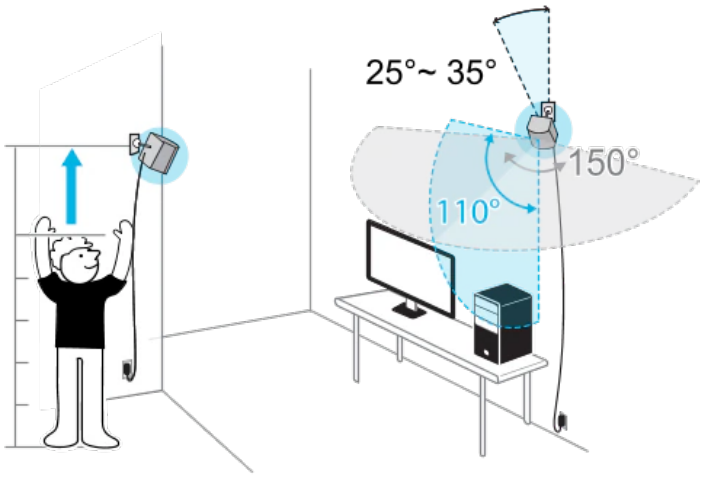


Figure 3: Recommended SteamVR Base Station 2.0 setup. Each base station provides approximately a 150° horizontal and 110° vertical field of view; mounting stations above head height (ideally ≥2 m) and tilting them downward (roughly 25–35°) helps maximize coverage of the tracking area. Adapted from HTC VIVE setup guidance (HTC VIVE, 2026).

ing 2.0 with multiple base stations observed that both static and slow dynamic translational tracking errors remained in the *sub-millimeter range* (below 1 mm RMSE) under controlled conditions when using four base stations (de Canaviri et al., 2023). This indicates that, with sufficient base station coverage and careful calibration, Lighthouse systems can approach the precision required for precise robotic evaluation tasks.

However, accuracy can degrade under more dynamic or less controlled conditions. Independent motion capture comparisons involving the HTC Vive Tracker 3.0 found *spatial differences of approximately 10.4 mm ± 4.5 mm at the knee* and *11.3 mm ± 5.1 mm at the ankle* compared to a Vicon system during athletic movement tracking, illustrating how real-world motions and base station configuration influence measurement error (Merker et al., 2023).

Other work on earlier versions of the SteamVR tracking system also reports position errors on the order of *millimeters to centimeters*, with median errors under a centimeter for static tests but increased variability under occlusions or when base station geometry is suboptimal (Sansone et al., 2022). These results underscore that while Lighthouse systems can yield precise tracking in optimal setups, their absolute accuracy may fall short of dedicated optical motion capture for highly precise robotics tasks.

As a result, Lighthouse-style systems are commonly adopted in robotics research for indoor 6 DoF tracking in teleoperation, prototyping, and temporary experimental setups (University of Massachusetts Amherst Robotics Lab, 2026). While they provide absolute pose estimates without long-term drift, their sensitivity to line-of-sight constraints, reflective surfaces, and workspace geometry generally prevents them from matching the consistency and metrological reliability of dedicated camera-based optical motion capture systems.

3.4 Summary of Trade-offs

Table 2 summarizes the key trade-offs across external global tracking modalities. All provide absolute, drift-free 6 DoF pose in a shared coordinate frame via environment-mounted optical infrastructure, but differ in accuracy, infrastructure burden, and environmental sensitivity. Reliance on external infrastructure limits scalability for deployment; nevertheless, their globally consistent, drift-free measurements make them indispensable for data collection, evaluation, and sim-to-real transfer.

4 Onboard and Relative Motion Sensing

Onboard sensing estimates motion using sensors carried by the robot itself, producing pose estimates relative to an initial state or incrementally built map. These methods trade absolute accuracy for auton-

Table 2: Comparison of external global tracking systems.

System	Accuracy	LoS	Infra.	Cost	Env. Sensitivity	Best Suited For
Passive Optical	$<$1 mm	Yes	High (camera cage)	\$30K–300K+	Lighting, occlusion	Ground-truth benchmarking, evaluation
Active Optical	$<$1 mm	Yes	High (camera cage)	\$30K–300K+	Lighting, occlusion	Ground-truth benchmarking, evaluation
Lighthouse	1–10 mm	Yes	Med. (base stations)	\$300–1K	Reflections, geometry	Teleoperation, robotic data collection

LoS: Line-of-sight required. **Infra.**: Infrastructure requirements. The active optical principle also appears inside-out in Meta Quest controller tracking (see text and Section 6).

Figure 4: VRTRIX PRO IMU-based data glove. The glove integrates multiple 9-axis MEMS IMUs distributed across the hand to estimate finger and wrist articulation, providing 6 DoF hand pose and joint orientation measurements for motion capture and teleoperation (VRTRIX, 2023).

omy and scalability, enabling operation in uninstrumented environments at the cost of drift accumulation (Barfoot, 2017; Thrun et al., 2005).

4.1 Inertial Measurement Units (IMUs)

An inertial measurement unit (IMU) measures angular velocity via gyroscopes and specific force via accelerometers, typically in the body frame. In strapdown inertial navigation, these measurements are integrated to propagate orientation, velocity, and position over time (Barfoot, 2017; Groves, 2013). IMUs enable fully self-contained motion sensing with low latency and no external infrastructure.

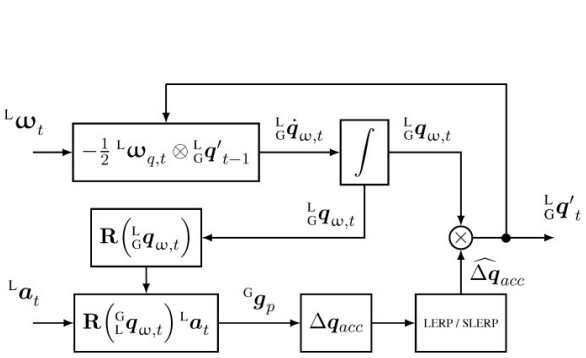
Measurement model. A standard IMU model expresses measured angular velocity and specific force as

$$\tilde{\boldsymbol{\omega}}(t) = \boldsymbol{\omega}(t) + \mathbf{b}_g(t) + \mathbf{n}_g(t), \tag{1}$$

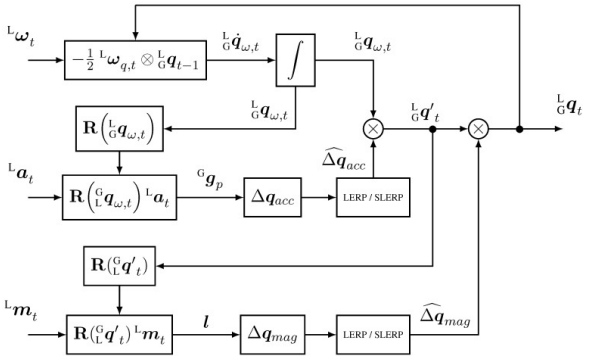
$$\tilde{\mathbf{f}}(t) = \mathbf{R}^\top(t)\big(\mathbf{a}(t) - \mathbf{g}\big) + \mathbf{b}_a(t) + \mathbf{n}_a(t), \tag{2}$$

where $\mathbf{R}(t) \in SO(3)$ is the body-to-world rotation, $\mathbf{a}(t)$ is translational acceleration in the world frame, $\mathbf{g}$ is gravity, $\mathbf{b}_g, \mathbf{b}_a$ are slowly varying sensor biases, and $\mathbf{n}_g, \mathbf{n}_a$ are measurement noise terms (Barfoot, 2017; Groves, 2013). Biases are often modeled as random walks, reflecting temperature dependence and time-varying sensor characteristics.

From 6-axis to 9-axis IMUs. A 6-axis IMU combines a 3-axis accelerometer with a 3-axis gyroscope. The accelerometer measures specific force, which under quasi-static conditions resolves the gravity vector and thereby estimates *tilt* (pitch and roll) but cannot distinguish heading. The gyroscope measures angular velocity; integrating it yields relative orientation changes but accumulates drift. A 9-axis IMU adds a 3-axis *magnetometer* that measures the local magnetic field vector. Because the Earth's magnetic



(a) 6-axis IMU (no magnetometer): gyroscope prediction corrected by accelerometer delta quaternion $\Delta\hat{\mathbf{q}}_{acc}$ for roll and pitch only. Yaw drifts over time.



(b) 9-axis MARG (with magnetometer): a second correction stage computes $\Delta\hat{\mathbf{q}}_{mag}$ from the tilt-compensated magnetic field, providing absolute yaw.

Figure 5: Complementary orientation filters in quaternion form. (a) The 6-axis filter fuses gyroscope and accelerometer, yielding relative heading. (b) The 9-axis filter adds magnetometer-based heading correction for absolute yaw. Adapted from Valenti et al. (Valenti et al., 2015) (CC-BY).

field has a known horizontal component pointing toward magnetic north, the magnetometer provides an *absolute heading (yaw) reference* that neither accelerometers nor gyroscopes can supply (Groves, 2013).

Extracting heading from a magnetometer requires *tilt compensation*: the raw magnetic reading $\mathbf{m} = (m_x, m_y, m_z)$ must be projected onto the horizontal plane using accelerometer-derived pitch θ and roll ϕ before computing the heading angle ψ (Groves, 2013). This two-stage process—tilt from accelerometer, heading from tilt-compensated magnetometer—is the foundation of all attitude and heading reference systems (AHRS).

Orientation estimation via sensor fusion. Raw gyroscope integration drifts; accelerometer and magnetometer readings are noisy and sensitive to dynamic motion and magnetic disturbances. Sensor fusion algorithms combine these complementary signals to produce stable orientation estimates. Three families dominate practice (Valenti et al., 2015; Caruso et al., 2021):

- **Complementary filters** blend high-frequency gyroscope data with low-frequency accelerometer and magnetometer corrections in the frequency domain, requiring minimal tuning (a single cutoff parameter) and the least computation.
- **Gradient-descent filters** (e.g., Madgwick (Madgwick et al., 2011)) minimize a quaternion-space error between measured and predicted gravity and magnetic field vectors via gradient descent. A single gain parameter β controls the correction strength, yielding efficient and robust orientation estimation well suited for embedded systems.
- **Extended Kalman filters (EKF)** maintain a probabilistic state estimate with explicit process and measurement noise covariance matrices $\mathbf{Q}$ and $\mathbf{R}$, offering theoretically optimal fusion under Gaussian assumptions at the cost of higher computational complexity and more involved tuning.

In practice, accuracy differences among these approaches are small—typically under 1° in well-tuned setups—so the choice depends primarily on computational budget and real-time constraints (Caruso et al., 2021). Figure 5 illustrates the signal flow for both 6-axis and 9-axis complementary filter architectures.

The 6-axis vs. 9-axis trade-off. Magnetometers provide absolute heading but are sensitive to magnetic disturbances. The Earth's magnetic field is weak (approximately 25–65 μT), making magnetometers vulnerable to *hard-iron* interference (constant offsets from nearby permanent magnets or magnetized components) and *soft-iron* interference (field distortion from ferromagnetic materials that stretches the measurement sphere into an ellipsoid) (Kuncar et al., 2016). Indoor environments with steel structures, electronic equipment, and motors commonly exhibit magnetic distortions that degrade heading accuracy.

A 6-axis (magnetometer-free) configuration avoids these issues entirely: yaw becomes relative rather

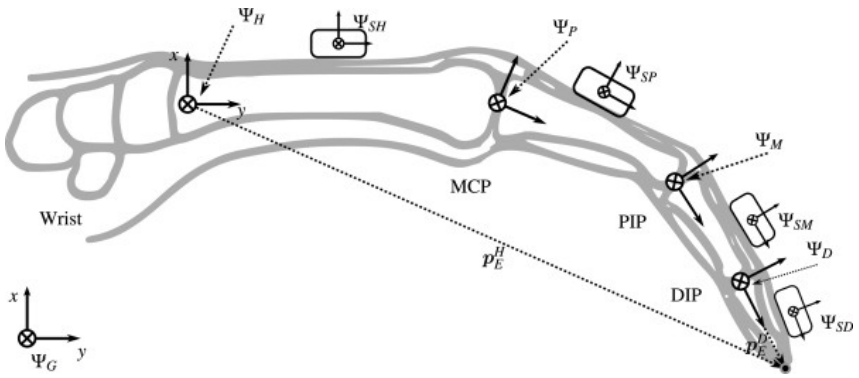


Figure 6: Sagittal view of the left index finger modeled as a kinematic chain of four rigid bodies. Coordinate frames are attached to the hand (Ψ_H), proximal (Ψ_P), medial (Ψ_M), and distal (Ψ_D) phalanges. IMU sensor frames (Ψ_{SH}, Ψ_{SP}, Ψ_{SM}, Ψ_{SD}) are mounted on each segment. Joints are labeled MCP, PIP, and DIP. Fingertip position $\mathbf{p}_E$ is computed via forward kinematics through the chain. Adapted from Kortier et al. (Kortier et al., 2014) (CC-BY).

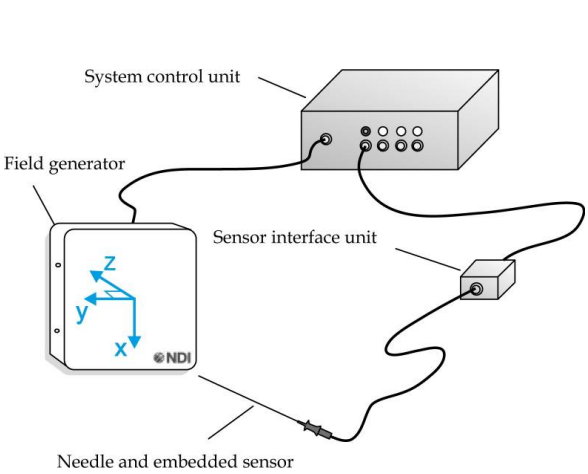
than absolute, drifting over time, but the estimate is immune to magnetic interference. This trade-off is especially important for IMU gloves used in conjunction with electromagnetic tracking systems (Section 6), where the EM field generator—orders of magnitude stronger than Earth's field at close range—would overwhelm the magnetometer (Marín et al., 2020). Adaptive approaches offer a middle ground: the magnetometer is trusted only when the ambient magnetic field is stable (detected via quasi-static field monitoring) and ignored during disturbances, seamlessly falling back to 6-axis operation (Ettlinger et al., 2024).

Drift accumulation. Because gyroscope and accelerometer biases are integrated (once for orientation/velocity, twice for position), even small biases produce errors that grow over time. Gyroscope bias causes orientation drift, which misaligns gravity direction and injects spurious acceleration into the translational channel, leading to rapidly increasing position error (Barfoot, 2017; Groves, 2013). Without external references, IMU-only navigation cannot maintain bounded absolute pose error.

Observability and excitation. Bias estimation requires sufficient motion excitation—rotations and accelerations that provide informative signal changes. Near-constant velocity or low rotational excitation degrades estimation quality (Barfoot, 2017; Mourikis and Roumeliotis, 2007), which is why IMUs are typically fused with exteroceptive sensing or task constraints (e.g., zero-velocity updates).

IMU placement and finger kinematic reconstruction. IMU-based data gloves distribute multiple MEMS IMUs across the hand to reconstruct finger articulation. A full instrumentation places one IMU per phalanx segment—proximal, middle, and distal—plus the palm, requiring 15–18 IMUs for a complete hand (Kortier et al., 2014; Lin et al., 2018). Budget designs use fewer sensors (6–7 per hand) and infer intermediate joint angles via biomechanical coupling constraints, such as the commonly assumed DIP/PIP flexion coupling ratio (Lin et al., 2018).

Each finger is modeled as a serial kinematic chain of rigid links (Figure 6). The metacarpophalangeal (MCP) joint provides 2 DoF (flexion/extension and abduction/adduction), while the proximal interphalangeal (PIP) and distal interphalangeal (DIP) joints each provide 1 DoF (flexion/extension), yielding 4 DoF per finger and approximately 20 DoF for the full hand (Kortier et al., 2014). Joint angles are extracted from the *relative orientation* between adjacent IMUs: for two consecutive segments i and j, the joint rotation is $\mathbf{R}_{ij} = \mathbf{R}_i^\top \mathbf{R}_j$, where each $\mathbf{R}$ is the orientation estimate from the corresponding IMU's sensor fusion filter. This differential measurement cancels common-mode errors—if both adjacent sensors experience the same magnetic distortion, the relative angle remains accurate—which is why joint angle estimates are typically more precise than absolute orientation (Seel et al., 2014). Given joint angles and known phalanx lengths, fingertip positions are recovered via forward kinematics through the serial chain.



(a) Simplified representation of the Aurora electromagnetic tracking system, illustrating the fixed external planar field generator and the tracked sensor used to determine 6 DoF pose via induced magnetic field measurements (Boutaleb et al., 2015).

(b) Manus Quantum Metagloves — wearable EMF glove system providing millimeter-level tracking of finger pose relative to the wrist without reliance on optical line-of-sight.

Figure 7: Examples of electromagnetic tracking systems used for hand and finger motion capture.

Autonomy vs. absolute accuracy. IMUs operate in darkness, without texture, and without line-of-sight, providing high-rate signals valuable for control and dynamics learning. However, IMU-only estimates exhibit unbounded drift and are unsuitable as ground truth. The VRTRIX data glove illustrates achievable wearable IMU performance: dynamic orientation accuracy of $\pm 0.5^\circ$ for pitch/roll and $\pm 1^\circ$–2° for yaw, at up to 120 Hz per hand (VRTRIX, 2023). Hybrid architectures such as the Rokoko Smartgloves combine IMU sensors with a local electromagnetic field coil, using the EMF signal for drift-free absolute position when within range and falling back to IMU-only tracking otherwise (Rokoko, 2026). Orientation is estimated reliably across these systems, while absolute position must be inferred through kinematic constraints or external references, accumulating error over time.

4.2 Electromagnetic Tracking Systems

Electromagnetic (EM) tracking systems estimate 6 DoF pose by exploiting the physics of alternating magnetic fields and induced currents. A field generator emits a quasi-static, low-frequency electromagnetic field, while small receiver coils (sensors) mounted on a tracked object measure the induced electromagnetic response. By modeling the spatial variation of the magnetic field (e.g., using magnetic dipole or multi-coil field models), these induced signals can be inverted to recover the *position and orientation* of the sensor relative to the generator's coordinate frame (Northern Digital Inc. (NDI), 2026b; He et al., 2025). In wrist-mounted configurations—the most common form factor for dexterous hand tracking—the field generator is attached to the wrist and moves with the user, so finger poses are recovered *relative to the wrist frame* rather than a fixed global frame. This makes EM-based finger tracking an onboard, relative sensing modality analogous to IMU-based gloves, rather than a traditional external tracking system.

In practice, an EM tracking system typically consists of:

- a *field generator* that creates a known electromagnetic field,
- one or more *sensor coils* that detect field variations, and
- a control unit that processes induced voltages into real-time pose estimates.

The absence of reliance on optical line-of-sight enables EM tracking to operate in occluded or visually cluttered environments, with *drift-free pose estimates* as long as the generated field is well characterized.

Representative EM Tracking Systems. *Polhemus* offers a family of precision EM tracking systems in which a transmitter produces a controlled magnetic field and sensors deliver 6 DoF pose at real-time

Figure 8: Example SLAM sensing rig used for manipulation experiments, consisting of wrist-mounted cameras and fiducial markers for visual tracking and mapping (UMI Project, 2024; Chi et al., 2024).

update rates (e.g., on the order of 100–120 Hz). These systems are capable of sub-millimeter tracking accuracy in controlled environments and are widely used for instrument and hand motion capture where optical visibility is limited (Social, Life, and Engineering Sciences Imaging Center, Penn State, 2026; He et al., 2025).

NDI Aurora is a medical-grade EM tracking solution designed for real-time tracking of small sensors embedded in instruments such as probes, catheters, or guidewires during image-guided procedures. Aurora systems typically report positional accuracies on the order of approximately 0.5 mm RMS and orientation accuracies in the sub-degree range within the specified operating volume, and support multiple simultaneous sensors without line-of-sight constraints (Northern Digital Inc. (NDI), 2026a).

Manus EMF-based gloves integrate electromagnetic field (EMF) sensors into wearable gloves for detailed hand and finger motion capture. These EMF sensors continuously sample the local magnetic field produced by a small wrist-mounted transmitter and convert field variations into stable 3D pose estimates for each finger segment relative to the wrist. EMF-based gloves offer millimeter-level hand tracking that remains robust in cluttered or visually occluded environments (MANUS, 2026).

Trade-offs and Limitations. EM tracking introduces *environmental sensitivity*: nearby ferromagnetic or conductive materials can distort the field and degrade accuracy (Northern Digital Inc. (NDI), 2026a). In practice, most tabletop and laboratory settings have sufficiently little metal for reliable operation, but proximity to large ferromagnetic objects (e.g., industrial fixtures) remains problematic.

Because the wrist-mounted field generator moves with the hand, workspace size is not a constraint for finger tracking, and finger poses remain drift-free within the wrist frame. However, recovering absolute global hand pose still requires an external reference or additional sensing. EM tracking is best suited for occlusion-heavy scenarios with minimal magnetic interference, such as dexterous hand tracking in desktop or non-metallic environments.

4.3 SLAM-Based Sensing

Simultaneous Localization and Mapping (SLAM) estimates motion by jointly inferring the robot's pose and a map of the environment from onboard sensors—cameras, LiDAR, or depth sensors. By leveraging persistent environmental structure, SLAM constrains motion estimates beyond what IMU-only integration allows (Thrun et al., 2005; Cadena et al., 2016).

How SLAM works. SLAM formulates localization as joint inference over the robot trajectory and map. Modern systems decompose this into a front-end that extracts features and establishes correspondences across time, and a back-end that optimizes these constraints—typically as a factor graph—to obtain locally consistent trajectory and map estimates. Loop closures introduce additional constraints that reduce accumulated drift (Grisetti et al., 2010).

Vision algorithms in SLAM. Classical approaches use sparse feature detection (ORB, FAST/BRIEF) with epipolar geometry and PnP formulations, representing the map as 3D landmarks refined through bundle adjustment (Mur-Artal et al., 2015; Cadena et al., 2016). Direct methods minimize photometric error over image intensities, improving performance in low-texture scenarios at the cost of illumination sensitivity (Engel et al., 2018). LiDAR-based systems use geometric registration (ICP, scan matching) on dense point clouds (Grisetti et al., 2010). Learning-based methods have explored neural replacements for classical feature extraction, improving robustness to illumination and texture-poor scenes but with reduced geometric guarantees and increased domain sensitivity (Wang et al., 2017; Teed and Deng, 2021).

Drift and environment dependence. SLAM estimates pose relative to an incrementally built map, so error accumulates with distance and time. Loop closure reduces but cannot eliminate drift in large-scale or non-revisiting trajectories (Cadena et al., 2016). Performance depends strongly on environmental structure: visual SLAM degrades in low-texture, repetitive, or dynamic scenes.

Autonomy vs. absolute accuracy. SLAM scales naturally to large environments without external infrastructure, making it indispensable for mobile robots and field deployment. However, it does not provide globally consistent, drift-free pose without external references, making SLAM estimates unsuitable as ground truth when precise cross-trial alignment is required. SLAM typically serves as the primary localization mechanism during deployment, complemented by external tracking during data collection and evaluation. Consumer XR headsets have become the most widely deployed VI-SLAM devices in recent robotics teleoperation work. Meta Quest 3 (four tracking cameras and one IMU) achieves approximately 0.77 cm relative pose error (RPE), while Apple Vision Pro (six tracking cameras, four IMUs, a LiDAR scanner, and a dedicated R1 co-processor) achieves approximately 0.52 cm RPE (Hanisch et al., 2025). Their practical role as tracking backbones for dexterous manipulation systems is discussed in Section 6.

5 Internal Joint-State Sensing

Internal joint-state sensing measures the configuration of an articulated mechanism rather than its spatial pose. In contrast to external tracking and onboard estimation, joint-state sensing directly observes internal degrees of freedom such as joint angles—fundamental to kinematic control and repeatable execution, but insufficient for global spatial information (Craig, 2018; Siciliano et al., 2009).

5.1 Magnetic Encoders

Magnetic encoders are widely used for joint-state sensing in robotic arms and dexterous hands. A small permanent magnet is attached to a rotating joint, and magnetic field sensors—commonly Hall-effect or magnetoresistive—measure the local field to infer joint angle. Compared to optical encoders, magnetic encoders are compact, contactless, and robust to dust, vibration, and wear, making them well suited for densely packed joints and wearable devices (Siciliano et al., 2009).

Joint accuracy and repeatability. Magnetic encoders provide direct, absolute joint angle measurements with high repeatability and no drift. The AS5600 Hall-effect encoder, widely used in robotic applications, provides 12-bit resolution (4096 positions per revolution, $\approx 0.087^\circ$ per step) with I^2C, PWM, and analog outputs (ams OSRAM, 2026). Practical implementations report angular precision of 0.1°–0.4° depending on magnet alignment and calibration (Adafruit Industries, 2025).

Magnetic encoders do not directly observe spatial pose. End-effector position and orientation can only be inferred through forward kinematics, which requires accurate kinematic models and known base pose. Uncertainty in link lengths, joint offsets, or base motion propagates into task-space error (Craig, 2018). Joint encoders alone cannot recover absolute pose and are insufficient for tasks requiring global localization.

Joint-state sensing is especially important in dexterous hands, which may contain dozens of closely spaced joints with complex kinematic coupling. Joint encoders often provide the only reliable means of observing fine-grained finger articulation and contact configurations. While external or onboard sensing can estimate hand motion as a whole, precise manipulation and force control depend critically on

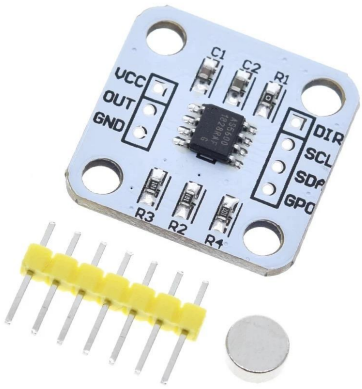


Figure 9: AS5600 magnetic rotary position sensor module. The device measures absolute joint angle using contactless Hall-effect sensing of a diametrically magnetized on-axis permanent magnet, providing 12-bit angular resolution for internal joint-state sensing (ams OSRAM, 2026).

joint-level measurements. High-quality magnetic encoders are therefore a cornerstone of dexterous manipulation systems, even though they must be combined with spatial sensing for full task awareness.

6 Combining Multiple Sensing Modalities

Dexterous manipulation demands simultaneous knowledge of global pose, limb kinematics, finger articulation, and often contact state: no single sensor class addresses all these requirements. This reality has driven researchers toward multimodal systems that combine complementary sensors. However, physical constraints impose hard limits on which modalities can coexist: electromagnetic fields interfere with magnetometers, metallic objects distort EM tracking, and optical systems require controlled lighting or fail under occlusion. Successful system design must account for both complementary strengths and mutual interference patterns.

This section examines how sensing modalities are combined in practice: the physical constraints governing sensor compatibility (6.1), architectural patterns for sensor fusion (6.2), and recent systems deployed for dexterous manipulation data collection (6.3).

6.1 Physical Constraints and Interference

Electromagnetic and Inertial Sensor Incompatibility. EM tracking systems determine pose by measuring magnetic fields from a transmitter, creating a direct conflict with 9-axis IMUs that use magnetometers to correct yaw drift. The EM field—orders of magnitude stronger than Earth's field at close range—overwhelms the magnetometer, rendering yaw estimates unreliable. More broadly, IMU magnetometers are sensitive to indoor magnetic disturbances from structural steel and electronic equipment (Marín et al., 2020). Systems combining EM finger tracking with IMU-based arm orientation must either spatially isolate the subsystems or operate magnetometer-free, accepting increased heading drift (Marín et al., 2020).

Electromagnetic Tracking and Metallic Environments. Conductive metals induce eddy currents that distort the EM tracker's field; ferromagnetic materials cause static, nonlinear distortions (MacLachlan et al., 2016). Compatibility with ferromagnetic metals remains unsolved, though some trackers tolerate eddy currents from nonferromagnetic metals at lower frequencies (MacLachlan et al., 2016). Clinical studies confirm that distortion is setup-specific, with sources including patient couches, C-arms, and imaging equipment (Yaniv et al., 2009). This limits EM tracking in industrial settings with metal fixtures and workpieces.

Optical and Visual-Wearable Sensitivities. Infrared optical systems are sensitive to direct sunlight (saturating sensors or creating false detections), reflective surfaces (generating spurious returns), and

ambient vibrations (Skurowski and Pawlyta, 2019). These sensitivities explain why optical mocap typically requires dedicated environments with matte surfaces, darkened windows, and rigidly mounted cameras. Beyond environmental factors, self-occlusion during contact-rich manipulation is the dominant failure mode for vision-based finger tracking: fingers disappear behind grasped objects precisely when accurate pose estimation matters most. Additionally, wearing gloves or exoskeletons degrades markerless vision-based hand tracking, as these devices alter the hand's visual appearance. DexUMI (Xu et al., 2025) addresses this by inpainting glove-wearing hands in RGB observations, recovering natural hand appearance for downstream policy learning.

Implications for System Design. These constraints impose hard boundaries on sensor combinations. Systems requiring both EM finger tracking and IMU arm orientation must spatially isolate subsystems or drop magnetometer fusion. Portable systems cannot rely on EM tracking where metallic objects may be present. Optical systems moving outdoors must adopt active LED markers with electronic shutters rather than passive retroreflective spheres. Vision-based hand tracking systems must anticipate both self-occlusion during manipulation and visual interference from wearable sensors, motivating fusion with non-optical finger sensing or image-space corrections.

6.2 Architectural Patterns for Sensor Fusion

Visual-inertial sensing has emerged as the dominant portable tracking backbone for dexterous manipulation systems. It is worth distinguishing two levels: visual-inertial odometry (VIO) fuses camera and IMU data to estimate incremental pose but accumulates drift, while visual-inertial SLAM (VI-SLAM) adds loop closure and map management to bound drift over longer horizons. In practice, dedicated devices such as the Intel RealSense T265 run VI-SLAM pipelines internally, and consumer XR headsets run proprietary VI-SLAM as their primary headset localization mechanism. Systems using GoPro footage with ORB-SLAM3 also operate at the VI-SLAM level. We refer readers to Section 4.3 for technical details on SLAM algorithms underlying these systems.

However, the term "consumer XR tracking" masks significant architectural differences. Meta Quest 3 is a *hybrid* system combining three distinct subsystems at different accuracy tiers: (1) VI-SLAM headset localization via four tracking cameras and one IMU, achieving approximately 0.77 cm relative pose error (RPE); (2) active IR LED constellation tracking for controllers, where the headset cameras observe known LED patterns on the Touch Plus controllers and solve pose via PnP, fused with controller IMU and a concurrent markerless hand model—the same active optical principle as PhaseSpace (Section 3) but inside-out—achieving low-millimeter accuracy *relative to the headset* (Meta, 2023); and (3) markerless hand tracking via computer vision, with approximately 1.73 cm average positional error (Hanisch et al., 2025; Bateman et al., 2025). Critically, controller global accuracy remains bounded by the headset's VI-SLAM precision—the low-millimeter figure describes controller-to-headset relative pose, not world-frame accuracy. Apple Vision Pro takes a fundamentally different approach: it provides superior VI-SLAM headset tracking (0.52 cm RPE via six cameras, four IMUs, a LiDAR scanner, and a dedicated R1 co-processor (Hanisch et al., 2025)) but ships without controllers entirely, relying on markerless hand tracking and eye gaze as primary inputs. For teleoperation, the controller-to-headset relative precision is what determines control quality during an episode, giving Quest controllers (e.g., ARCap) a meaningful advantage over Vision Pro's markerless hand tracking (e.g., Open-TeleVision, Bunny-VisionPro) for fine manipulation—though both share centimeter-scale global accuracy bounded by their respective VI-SLAM pipelines. For comparison, SteamVR Lighthouse (Section 3) achieves sub-millimeter static accuracy in a fixed world frame via base stations—true global precision, but requiring dedicated infrastructure.

Given this visual-inertial tracking backbone, the challenge shifts to fusing finger-level, tactile, or body-level sensing. We identify five architectural patterns.

Hierarchical Fusion. In hierarchical fusion, different sensors are assigned to disjoint degrees of freedom with minimal overlap. DexCap (Wang et al., 2024) exemplifies this: SLAM cameras on each hand estimate wrist pose, while EMF gloves track fingertip positions relative to the palm. Because the two

subsystems measure disjoint quantities—global wrist pose versus local finger configuration—they operate independently, and the final hand pose is reconstructed by composing the wrist transform with finger joint angles through forward kinematics. OSMO (Yin et al., 2024) similarly separates vision-based hand tracking from tactile contact sensing, assigning each to a distinct information channel. Hierarchical fusion is well suited to articulated systems where different kinematic chains have different observability properties.

Redundant Fusion with Filtering. When multiple sensors measure overlapping quantities, filtering algorithms—typically Extended Kalman Filter (EKF) variants—combine their outputs to exceed the accuracy of either alone. Cuadrado et al. (Cuadrado et al., 2021) demonstrated an EKF that handles missing optical markers by zeroing out their measurement entries and Jacobian rows, achieving lower errors than traditional interpolation. Pan et al. (Pan et al., 2023) fuse monocular RGB with sparse IMU signals, using a hidden state feedback mechanism that adaptively weights visual versus inertial inputs based on signal quality. VIST (Lee et al., 2021) combines glove-mounted IMUs with a stereo camera and visual markers in an EKF framework, achieving robust hand tracking through occlusion, electromagnetic interference, and physical contact—conditions that defeat any single modality. HandCept (Huang et al., 2025) fuses wrist-mounted RGB-D with finger IMUs via an EKF for proprioceptive hand pose estimation. The key insight across these systems: sensor weaknesses are often uncorrelated—optical systems fail during occlusion while IMUs drift—so combining both inherits strengths while mitigating individual failures.

Reference Anchoring. A third pattern uses one sensor to provide an absolute reference frame anchoring drift-prone sensors. ARCap (Chen et al., 2024) combines a VR headset (Meta Quest 3) with motion capture gloves, using the headset's inside-out tracking as the global coordinate frame. Test-time alignment replaces offline calibration: users visually match a virtual robot to the physical robot in AR, establishing the tracking-to-robot transform in seconds. EgoAllo (Yi et al., 2025) takes a different approach: given VI-SLAM-derived head pose from an egocentric device, a conditional diffusion model estimates full body and hand pose, effectively using the head as a single anchor point. HaWoR (Zhang et al., 2025b) reconstructs world-space hand trajectories from egocentric video using an adaptive SLAM pipeline that handles the rapid viewpoint changes inherent to head-mounted cameras. Reference anchoring trades absolute accuracy for portability—the anchor sensor need only provide a stable, drift-free reference frame, not laboratory-grade precision.

Learning-Based Fusion. Rather than hand-designing filter models, learning-based approaches train neural networks to weight and combine multimodal inputs. Cross-modal transformers can learn attention-based weighting between visual and tactile streams, adapting to context-dependent reliability without explicit sensor models. This contrasts with model-based EKF approaches that require specifying noise covariances and measurement models a priori. While learning-based fusion can handle complex sensor interactions that are difficult to model analytically, it requires substantial training data and may generalize poorly to novel sensor configurations.

Embodiment-Isomorphic Design. A growing class of systems sidesteps fusion and retargeting entirely by mechanically matching the human operator's hand to the robot end-effector. DEXOP (Fang et al., 2025) and MILE (Du et al., 2025) use exoskeletons whose joint structure mirrors the target robot hand, so human finger motion maps directly to robot commands without kinematic retargeting. DexUMI (Xu et al., 2025) attaches robot fingers directly to a human hand via a lightweight interface, using the physical coupling as an implicit "sensor" for finger pose. These embodiment-isomorphic designs eliminate the fusion problem by construction, though they constrain the operator to the robot's kinematic workspace.

6.3 Systems in Practice

Table 3 summarizes representative multimodal systems for dexterous manipulation data collection, spanning VIO-anchored, vision-only, exoskeleton-based, and tactile-augmented architectures.

Table 3: Comparison of multimodal motion capture systems for dexterous manipulation data collection.

System	Global Tracking	Hand/Finger	Tactile/Haptic	Fusion	Cost
DexCap (Wang et al., 2024)	SLAM (T265)	EMF gloves	—	Hierarchical	$3–5k
ARCap (Chen et al., 2024)	Quest 3	Mocap gloves	Haptic warnings	Ref. anchor	$1–2k
Open Teach (Iyer et al., 2024)	SteamVR	SteamVR	—	Ref. anchor	$1–2k
Open-TeleVision (Cheng et al., 2024)	Vision Pro	Vision Pro hands	—	Ref. anchor	$4–5k
Bunny-VisionPro (Ding et al., 2024)	Vision Pro	Vision Pro hands	Haptic finger cots	Ref. anchor	$4–5k
AnyTeleop (Qin et al., 2023)	RGB-D	Vision (hand det.)	—	Vision-only	$1–2k
DEXOP (Fang et al., 2025)	Ext. tracking	Exoskeleton	—	Isomorphic	—
MILE (Du et al., 2025)	—	Exo. + tactile	Fingertip tactile	Isomorphic	—
HATO (Lin et al., 2025)	—	Allegro teleop	Fingertip tactile	Hierarchical	—
DOGlove (Zhang et al., 2025a)	—	Glove (hall effect)	Force feedback	Hierarchical	$50
OSMO (Yin et al., 2024)	Vision-based	—	12×3-axis tactile	Hierarchical	—

= open-source code available.

(a) DexCap front view (b) DexCap back view (c) Details of the camera setup

Figure 10: DexCap portable motion capture system. (a,b) The wearable rig consists of EMF gloves for finger tracking and wrist-mounted SLAM cameras, powered by a backpack-housed Intel NUC and battery. (c) Quick-release camera mounts enable transfer between human and robot for calibration and data collection (Wang et al., 2024).

VI-SLAM and Lighthouse-Anchored Systems. DexCap (Wang et al., 2024) pairs Intel RealSense T265 cameras on each hand (VI-SLAM at 60 Hz) with EMF gloves for finger joint angles, using a quick-release buckle for camera transfer between human and robot. ARCap (Chen et al., 2024) uses a Meta Quest 3 as both tracking anchor and AR feedback channel, combined with motion capture gloves—the Quest 3's internal VI-SLAM provides the global reference frame, and its test-time AR alignment replaces offline calibration. Open-TeleVision (Cheng et al., 2024) and Bunny-VisionPro (Ding et al., 2024) leverage Apple Vision Pro's VI-SLAM for 6-DoF wrist pose and finger articulation from a single consumer device, with Bunny-VisionPro adding haptic finger cots for force feedback. Open Teach (Iyer et al., 2024) takes a different approach, using SteamVR lighthouse tracking (external IR base stations) rather than visual-inertial sensing, providing a modular, open-source framework with a unified teleoperation interface for bimanual dexterous tasks.

Vision-Only and Exoskeleton Systems. AnyTeleop (Qin et al., 2023) uses only RGB-D cameras for both arm and hand tracking, detecting hand pose from monocular RGB and using depth for global localization—eliminating wearable hardware entirely at the cost of robustness to occlusion. At the other extreme, DEXOP (Fang et al., 2025) and MILE (Du et al., 2025) use exoskeletons whose kinematics mirror the target robot hand, mapping human finger motion to robot commands without retargeting. MILE additionally integrates fingertip visuotactile sensors (GelSight Mini) on the robot side, closing the tactile perception loop. DexUMI (Xu et al., 2025) takes a hybrid approach, attaching robot fingers directly to the human hand so that the physical coupling serves as the control interface, while inpainting the operator's gloved hand in RGB observations to preserve visual policy learning.

Tactile-Augmented Systems. OSMO (Yin et al., 2024) augments vision-based hand tracking with 12 three-axis magnetic tactile sensors (fingertips and palm), measuring normal and shear forces up to 80 N; the same glove worn by demonstrators mounts on the robot, eliminating the tactile embodiment gap. On a contact-intensive wiping task, policies trained with tactile data achieved 72% success versus

approximately 30% for vision-only baselines. HATO (Lin et al., 2025) enables bimanual visuotactile teleoperation with two Allegro hands augmented with fingertip tactile sensors, learning contact-rich skills like in-hand reorientation. DOGlove (Zhang et al., 2025a) provides an extremely low-cost ($50) open-source haptic glove using hall-effect sensors for finger pose and electromagnetic actuators for force feedback, demonstrating that tactile teleoperation need not require expensive custom hardware.

Common Challenges. *Spatial calibration* remains a universal requirement: each system must establish transforms between sensor frames and between human anatomy and robot kinematics. DexCap's quick-release mount and ARCap's AR alignment represent different tradeoffs between precision and usability. *Temporal synchronization* across sensors at different rates (60 Hz SLAM, 90 Hz Vision Pro, 1 kHz tactile) requires careful timestamp alignment; hardware synchronization trades complexity for reliability. *Embodiment retargeting* is dominated by fingertip inverse kinematics (Handa et al., 2020)—given human fingertip positions, solving for robot joints at corresponding locations—though recent work shows that wrist adjustment improves accuracy (Li et al., 2024) and that preserving inter-finger distances captures grasp intent more faithfully than position alone.

7 Conclusion

No single motion capture technology meets all the demands of dexterous manipulation. As summarized in Table 1, every modality trades off accuracy, drift, portability, and environmental robustness, which is why practical systems combine multiple sensors.

Several trends emerge from our analysis of multimodal systems (Table 3). First, consumer XR devices have become a practical tracking backbone without dedicated infrastructure, though their accuracy profiles differ substantially. Both Meta Quest 3 and Apple Vision Pro run VI-SLAM for headset localization at sub-centimeter relative pose error, but Quest 3 adds active IR LED constellation tracking for its controllers (low-millimeter accuracy *relative to the headset*—the same principle as PhaseSpace, but inside-out), while Vision Pro ships without controllers and relies on markerless hand tracking as its primary input. For teleoperation, the controller-to-headset relative precision is what determines control quality during an episode, giving Quest controllers a meaningful advantage over Vision Pro's markerless hand tracking ($\sim$1.73 cm) for fine manipulation. Regardless of subsystem, these devices have shifted the design challenge from global tracking toward finger-level and contact sensing. Second, tactile augmentation is proving consequential: OSMO's 72% versus 30% success rate gap between tactile-equipped and vision-only policies on contact-rich tasks illustrates that contact state is not merely supplementary but essential for manipulation. Third, embodiment-isomorphic designs (DEXOP, MILE, DexUMI) represent a growing trend that sidesteps fusion and retargeting entirely through mechanical matching, trading operator ergonomics for engineering simplicity. Fourth, the open-source ecosystem has matured rapidly—the majority of systems in Table 3 release code and hardware designs, and costs range from $50 (DOGlove) to a few thousand dollars, a dramatic reduction from the $30K–300K+ price of optical mocap infrastructure.

These trends point toward several open challenges. Accuracy characterization remains inconsistent: few multimodal systems report end-to-end pose error against ground truth, making principled system comparison difficult. Temporal synchronization across heterogeneous sensors (60 Hz SLAM, 90 Hz XR tracking, 1 kHz tactile) lacks standardized protocols. Embodiment retargeting—mapping human hand motion to kinematically dissimilar robot hands—remains largely heuristic, with recent work suggesting that preserving inter-finger distances and adjusting wrist pose can improve grasp fidelity, but no consensus on best practices. Finally, as robotic learning methods scale to larger and more diverse demonstration datasets, the motion capture systems that produce this data will become an increasingly important determinant of what behaviors robots can acquire. We hope this survey provides practitioners with the technical grounding to navigate these trade-offs and design capture systems matched to their learning and deployment requirements.

Acknowledgements

We thank Yu Xiang from UT Dallas for insightful discussions on IMU sensing and sensor fusion, and Jessica Yin from NVIDIA for discussions on electromagnetic tracking.

References

Adafruit Industries. 2025. As5600 magnetic angle sensor: Practical performance. https://learn.adafruit.com/adafruit-as5600-magnetic-angle-sensor/overview. Reports practical precision and implementation considerations.

ams OSRAM. 2026. As5600 magnetic rotary position sensor. https://ams-osram.com/products/sensor-solutions/position-sensors/ams-as5600-position-sensor. 12-bit contactless magnetic rotary encoder with I^2C, PWM, and analog outputs.

Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. 2009. A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 57(5):469–483.

Timothy D. Barfoot. 2017. *State Estimation for Robotics*. Cambridge University Press.

Scott Bateman and 1 others. 2025. Robotic characterization of markerless hand-tracking on Meta Quest Pro and Quest 3 virtual reality headsets. *IEEE Transactions on Visualization and Computer Graphics*. Quest 3 markerless hand tracking: 1.73 cm avg positional error, 1.11 cm jitter; Quest Pro: 1.22 cm avg error.

Samir Boutaleb, Emmanuel Racine, Olivier Fillion, Antonio Bonillas, Gilion Hautvast, Dirk Binnekamp, and Luc Beaulieu. 2015. Performance and suitability assessment of a real-time 3d electromagnetic needle tracking system for interstitial brachytherapy. *Journal of Contemporary Brachytherapy*, 7(4):280–289.

Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, Jose Neira, Ian Reid, and John J. Leonard. 2016. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 32(6):1309–1332.

Marco Caruso, Angelo Maria Sabatini, Daniel Laidig, Thomas Seel, Alice Mantoan, Andrea Cereatti, Marta Susi, and 1 others. 2021. Analysis of the accuracy of ten algorithms for orientation estimation using inertial and magnetic sensing under optimal conditions: One size does not fit all. *Sensors*, 21(7):2580.

Sirui Chen, Chen Wang, Kaden Nguyen, Li Fei-Fei, and C. Karen Liu. 2024. ARCap: Collecting high-quality human demonstrations for robot learning with augmented reality feedback. In *arXiv preprint*.

Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. 2024. Open-TeleVision: Teleoperation with immersive active visual feedback. In *Conference on Robot Learning (CoRL)*.

Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. 2024. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In *Proceedings of Robotics: Science and Systems (RSS)*.

John J. Craig. 2018. *Introduction to Robotics: Mechanics and Control*, 4 edition. Pearson.

Javier Cuadrado, Florian Michaud, Urbano Lugrís, and Manuel Pérez Soto. 2021. Using accelerometer data to tune the parameters of an extended Kalman filter for optical motion capture: Preliminary application to gait analysis. *Sensors*, 21(2):427.

L. Kuhlmann de Canaviri and 1 others. 2023. Static and dynamic accuracy and occlusion robustness of steamvr tracking 2.0 in multi-base station setups. *Sensors*. Reports sub-millimeter translational accuracy with four base stations.

Runyu Ding, Yuzhe Qin, Jiyue Zhu, Chengzhe Jia, Shiqi Yang, Ruihan Yang, Xiaojuan Qi, and Xiaolong Wang. 2024. Bunny-VisionPro: Real-time bimanual dexterous teleoperation for imitation learning. In *arXiv preprint*.

Jinda Du, Jieji Ren, Qiaojun Yu, Ningbin Zhang, Yu Deng, Xingyu Wei, Yufei Liu, Guoying Gu, and Xiangyang Zhu. 2025. MILE: A mechanically isomorphic exoskeleton data collection system with fingertip visuotactile sensing for dexterous manipulation. *arXiv preprint arXiv:2512.00324*.

Jakob Engel, Vladlen Koltun, and Daniel Cremers. 2018. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(3):611–625.

Andreas Ettlinger, Andreas Wieser, and Hans Neuner. 2024. Robust determination of smartphone heading by mitigation of magnetic anomalies. *NAVIGATION: Journal of the Institute of Navigation*, 71(1).

Hao-Shu Fang, Branden Romero, Yichen Xie, Arthur Hu, Bo-Ruei Huang, Juan Alvarez, Matthew Kim, Gabriel Margolis, Kavya Anbarasu, Masayoshi Tomizuka, Edward Adelson, and Pulkit Agrawal. 2025. DEXOP: A device for robotic transfer of dexterous human manipulation. *arXiv preprint arXiv:2509.04441*.

Giorgio Grisetti, Rainer Kümmerle, Cyrill Stachniss, and Wolfram Burgard. 2010. A tutorial on graph-based slam. *IEEE Intelligent Transportation Systems Magazine*, 2(4):31–43.

Paul D. Groves. 2013. *Principles of GNSS, Inertial, and Multisensor Integrated Navigation Systems*, 2 edition. Artech House.

Ankur Handa, Karl Van Wyk, Wei Yang, Jacky Liang, Yu-Wei Chao, Qian Wan, Stan Birchfield, Nathan Ratliff, and Dieter Fox. 2020. DexPilot: Vision-based teleoperation of dexterous robotic hand-arm system. In *IEEE International Conference on Robotics and Automation (ICRA)*.

Lennart Hanisch, Bálint Dvid, Alexander Eichler, and Christian Holz. 2025. XR reality check: What commercial devices deliver for spatial tracking. *arXiv preprint arXiv:2508.08642*. Compares headset tracking accuracy of Quest 3 (0.77 cm RPE), Vision Pro (0.52 cm RPE), HoloLens 2, and others against OptiTrack/Qualisys ground truth.

Mingshan He, Aoji Zhu, and Lidong Yang. 2025. Electromagnetic tracking system for medical micro devices: Working principles and system overview. *Micromachines*, 16(10):1175. Provides a review of EM tracking principles based on electromagnetic induction and pose estimation mechanisms.

HTC VIVE. 2026. Tips for setting up the base stations (steamvr base station 2.0). `https://www.vive.com/us/support/vive-pro-eye/category_howto/tips-for-setting-up-the-base-stations.html`. Accessed 2026.

Junda Huang, Jianshu Zhou, Honghao Guo, and Yunhui Liu. 2025. HandCept: A visual-inertial fusion framework for accurate proprioception in dexterous hands. *arXiv preprint arXiv:2505.08213*.

Aadhithya Iyer, Zhuoran Peng, Yinlong Dai, Irmak Guzey, Siddhant Haldar, Soumith Chintala, and Lerrel Pinto. 2024. OPEN TEACH: A versatile teleoperation system for robotic manipulation. In *Conference on Robot Learning (CoRL)*.

Henk G. Kortier, Victor I. Sluiter, Daniel Roetenberg, and Peter H. Veltink. 2014. Assessment of hand kinematics using inertial and magnetic sensors. *Journal of NeuroEngineering and Rehabilitation*, 11:70. Open access, CC-BY.

Ales Kuncar, Martin Sysel, and Tomas Urbanek. 2016. Calibration of low-cost triaxial magnetometer. *MATEC Web of Conferences*, 76:05008.

Yongseok Lee, Wonkyung Do, Hanbyeol Yoon, Jinuk Heo, WonHa Lee, and Dongjun Lee. 2021. Visual-inertial hand motion tracking with robustness against occlusion, interference, and contact. *Science Robotics*, 6(58):eabe1315.

Haoran Li, Haoyu Wang, Yuxiang Li, and He Wang. 2024. Analyzing key objectives in human-to-robot retargeting for dexterous manipulation. *arXiv preprint*.

Bor-Shing Lin, I-Jung Lee, Shu-Yu Yang, Yi-Chiao Lo, Junghsi Lee, and Jean-Lon Chen. 2018. A modular data glove system for finger and hand motion capture based on inertial sensors. *Journal of Medical and Biological Engineering*, 38(4):532–540.

Toru Lin, Yu Zhang, Qiyang Li, Haozhi Qi, Brent Yi, Sergey Levine, and Jitendra Malik. 2025. Learning visuotactile skills with two multifingered hands. In *IEEE International Conference on Robotics and Automation (ICRA)*.

Robert A MacLachlan, Nicholas Parody, Shohin Mukherjee, Ralph Hollis, and Cameron N Riviere. 2016. Electromagnetic tracker for active handheld robotic systems. In *Proceedings of IEEE Sensors*. Published 2017.

Sebastian O. H. Madgwick, Andrew J. L. Harrison, and Ravi Vaidyanathan. 2011. Estimation of IMU and MARG orientation using a gradient descent algorithm. In *IEEE International Conference on Rehabilitation Robotics (ICORR)*, pages 1–7.

MANUS. 2026. Metagloves pro — high-precision, low-latency data gloves. `https://www.manus-meta.com/products/metagloves-pro`. Accessed 2026.

Javier Marín, Teresa Blanco, Juan de la Torre, and José J Marín. 2020. Gait analysis in a box: A system based on magnetometer-free IMUs or clusters of optical markers with automatic event detection. *Sensors*, 20(12):3338.

S. Merker, S. Pastel, D. Bürger, A. Schwadtke, and K. Witte. 2023. Measurement accuracy of the htc vive tracker 3.0 compared to vicon system for generating valid positional feedback in virtual reality. *Sensors*, 23(17):7371. Reports 10.4–11.3 mm differences vs Vicon under dynamic motion.

Meta. 2023. Tracking technology explained: LED matching. https://developers.meta.com/horizon/blog/tracking-technology-explained-led-matching/. Describes how Quest 3 Touch Plus controllers use IR LED constellations tracked by headset cameras via PnP pose estimation.

Anastasios I. Mourikis and Stergios I. Roumeliotis. 2007. A multi-state constraint kalman filter for vision-aided inertial navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 3565–3572.

Raul Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. 2015. Orb-slam: A versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163.

Richard M. Murray, Zexiang Li, and S. Shankar Sastry. 1994. *A Mathematical Introduction to Robotic Manipulation*. CRC Press.

Ashvin Nair, Bob McGrew, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. 2018. Overcoming exploration in reinforcement learning with demonstrations. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 6292–6299. IEEE.

Noitom Ltd. 2026. Perception neuron hybrid: Optical–inertial motion capture system. https://www.noitom.com.cn/perception-neuron-hybrid.html. Reported tracking accuracy of 0.08–0.10 mm under controlled conditions.

Northern Digital Inc. (NDI). 2026a. Aurora field generators – electromagnetic tracking solutions. https://www.ndigital.com/electromagnetic-tracking-technology/aurora/aurora-field-generators/. Describes different Aurora electromagnetic field generators and their measurement volumes.

Northern Digital Inc. (NDI). 2026b. Electromagnetic tracking — technology overview. https://www.ndigital.com/electromagnetic-tracking-technology/. Accessed 2026.

OptiTrack. 2026. Motion capture for robotics. https://optitrack.com/applications/robotics. Accessed 2026.

Shaohua Pan, Qi Ma, Xinyu Yi, Weifeng Hu, Xiong Wang, Xingkang Zhou, Jijunnan Li, and Feng Xu. 2023. Fusing monocular images and sparse IMU signals for real-time human motion capture. In *SIGGRAPH Asia 2023 Conference Papers*.

PhaseSpace. 2026a. Impulse x2e motion capture system specifications. https://www.phasespace.com/x2e-motion-capture/. Scalable camera counts and high-speed tracking for multi-subject setups.

PhaseSpace. 2026b. Phasespace motion capture products. https://www.phasespace.com/productsMain.html. Accessed 2026.

Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. 2023. AnyTeleop: A general vision-based dexterous robot arm-hand teleoperation system. In *Robotics: Science and Systems (RSS)*.

Rokoko. 2026. Smartgloves — affordable quality finger and hand motion capture. https://www.rokoko.com/products/smartgloves. Hybrid IMU and EMF sensor fusion via Volta Tracking Technology.

L. G. Sansone and 1 others. 2022. Robustness and static-positional accuracy of the steamvr tracking system. *Virtual Reality*. Reports low centimeter errors in room-scale static accuracy tests.

Thomas Seel, Jörg Raisch, and Thomas Schauer. 2014. IMU-based joint angle measurement for gait analysis. *Sensors*, 14(4):6891–6909.

Bruno Siciliano, Lorenzo Sciavicco, Luigi Villani, and Giuseppe Oriolo. 2009. *Robotics: Modelling, Planning and Control*. Springer.

Przemysław Skurowski and Magdalena Pawlyta. 2019. On the noise complexity in an optical motion capture facility. *Sensors*, 19(20):4435.

Social, Life, and Engineering Sciences Imaging Center, Penn State. 2026. 3d motion tracking system by polhemus (fastrak) — specifications. https://www.imaging.psu.edu/facilities/human-electrophysiology/3d-motion-tracking-system-polhemus. Accessed 2026.

Zachary Teed and Jia Deng. 2021. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in Neural Information Processing Systems (NeurIPS)*, 34.

Sebastian Thrun, Wolfram Burgard, and Dieter Fox. 2005. *Probabilistic Robotics*. The MIT Press.

UMI Project. 2024. Umi gripper hardware guide. https://umi-gripper.github.io/. Hardware documentation for wrist-mounted sensing used in manipulation and SLAM.

University of Massachusetts Amherst Robotics Lab. 2026. Motion capture for robotics. https://www.umass.edu/robotics/mrrl/research/motion-capture. Accessed 2026.

Roberto G. Valenti, Ivan Dryanovski, and Jizhong Xiao. 2015. Keeping a good attitude: A quaternion-based orientation filter for IMUs and MARGs. *Sensors*, 15(8):19302–19330. Open access, CC-BY.

Vicon Motion Systems. 2026. What is motion capture? how it works, and what it's used for. https://www.vicon.com/about-us/what-is-motion-capture/. Accessed 2026.

VRTRIX. 2023. Vrtrix datasheet. Product datasheet. Specifications for VRTRIX IMU-based data gloves, including orientation accuracy and update rate.

Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C. Karen Liu. 2024. DexCap: Scalable and portable mocap data collection system for dexterous manipulation. In *Robotics: Science and Systems (RSS)*.

Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. 2017. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*.

Mengda Xu, Han Zhang, Yifan Hou, Zhenjia Xu, Linxi Fan, Manuela Veloso, and Shuran Song. 2025. DexUMI: Using human hand as the universal manipulation interface for dexterous manipulation. *arXiv preprint arXiv:2505.21864*.

Ziv Yaniv, Emmanuel Wilson, David Lindisch, and Kevin Cleary. 2009. Electromagnetic tracking in the clinical environment. *Medical Physics*, 36(3):876–892.

Brent Yi, Vickie Ye, Maya Zheng, Yunqi Li, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. 2025. Estimating body and hand motion in an ego-sensed world. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jessica Yin, Haozhi Qi, Youngsun Wi, Sayantan Kundu, Mike Lambeta, William Yang, Changhao Wang, Tingfan Wu, Jitendra Malik, and Tess Hellebrekers. 2024. OSMO: Open-source tactile glove for human-to-robot skill transfer. In *arXiv preprint*.

Han Zhang, Songbo Hu, Zhecheng Yuan, and Huazhe Xu. 2025a. DOGlove: Dexterous manipulation with a low-cost open-source haptic force feedback glove. In *Robotics: Science and Systems (RSS)*.

Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. 2025b. HaWoR: World-space hand motion reconstruction from egocentric videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.