arXiv:2605.21714v1 [cs.CV] 20 May 2026

AVI-HT: Adaptive Vision-IMU Fusion for 3D Hand Tracking

Ziyi Kou **Ankit Kumar** **Mia Huang** **Taylor Niehues**

Vatsal Mehta **Ergys Ristani** **Li Guan**

Meta Reality Labs

Abstract

We present **AVI-HT**, an adaptive visual-IMU fusion approach for tracking 3D hand poses by jointly modeling the egocentric image with on-glove 6-DoF IMU signals. AVI-HT achieves significantly improved accuracy and availability, particularly in hand-object interaction (HOI) scenarios involving heavy visual occlusion. Two complementary ingredients underpin its success: (1) synchronized multi-modal training data pairing on-body vision-IMU sensor streams with ground-truth 3D hand poses from a motion-capture system, and (2) a cross-sensor deep attention mechanism that adaptively modulates the trust assigned to the vision and individual IMU sensors. To evaluate AVI-HT in real-world settings, we conduct extensive experiments on our DexGloveHOI dataset that consists of 100K+ pairwise vision-IMU samples with synchronized 3D annotated poses, in which users manipulate a variety of objects during daily tasks. We compare against multiple single- and multi-modal tracking approaches under two hand models (UmeTrack, MANO). The results show that AVI-HT reduces mean keypoint error by 16.1% and its wrist-aligned variant by 24.2% over the baselines. Ablation studies further reveal the per-finger contribution of IMU sensors across activity types, and the model's sensitivity to IMU noise and temporal misalignment in vision-IMU fusion.

1 Introduction

Consider the set of egocentric images in the first row of Figure 1, where a human hand dexterously manipulates various objects on the table, which is common in AR/VR [38, 25] and robotic teleoperation [31, 1] applications. As the grasp tightens, the fingers are progressively occluded. These interactions are contact-rich and happening in 3D, yet the very act of grasping hides the hand from the camera. Therefore, accurately estimating hand pose from egocentric vision alone becomes extremely difficult under occlusion, as the set of kinematically plausible configurations that explain the visible evidence could be prohibitively large [42]. To mitigate such limitations, we introduce a framework with a data sensing glove that can accurately perceive hand pose in 3D from joint visual and on-glove IMU inputs, particularly when the hand is heavily occluded by the object it is manipulating.

Recent developments in computer vision and robotics point to the direction where advances are achieved by leveraging joint visual and sensor inputs, powered by multi-modal fusion models. This emerging insight has been demonstrated in the context of computer vision by ImageBind [8], MultiMAE [2]. In the context of wearable sensing and human motion capture, we observe this with systems like [29] and [4]. In the area of 3D hand pose tracking, prior work has largely relied on a single modality: either RGB images [30, 33] or inertial sensors [34]. Each of these approaches suffers from fundamental limitations when applied to dexterous hand-object interaction (HOI) scenarios,

Preprint.

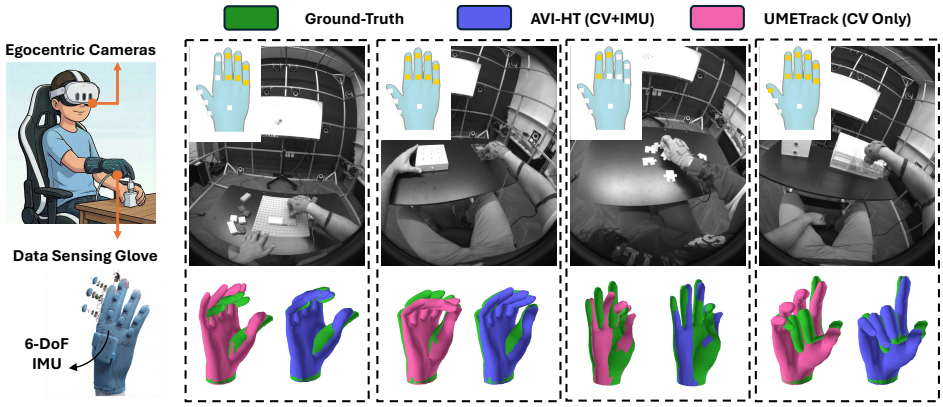

Figure 1: **3D hand tracking from egocentric vision and on-glove IMUs.** *(Left)* Our multi-modal data capture setup that consists of an egocentric camera and a data sensing glove with 12 6-DoF IMU sensors. *(Right)* Qualitative comparison on 4 egocentric frames. For each frame, the top row shows an inset hand diagram. The bottom row shows the corresponding 3D hand pose tracking overlaid for UMETrack and AVI-HT against ground truth. The tracking results by AVI-HT are closely aligned with the ground truth, while the vision-only baseline deviates substantially. The top yellow patches highlight the activated IMU sensors by the adaptive attention module of AVI-HT, indicating the regions with greater reliance on IMU signals over egocentric vision. *Best viewed in color.*

such as inaccurate hand reconstruction due to heavy occlusion from vision signals [26], magnetic interference from 9-DoF IMUs [24], or lack of absolute orientation from 6-DoF IMUs [34].

In this paper, we take the philosophy of multi-modal fusion and apply it to the problem of 3D hand pose tracking. *Our motivation is that visual signals and 6-DoF IMUs naturally complement each other: vision provides absolute position and global orientation from wrist but degrades in finger tracking under occlusion, while 6-DoF IMUs capture high-frequency local joint dynamics that are invariant to visual obstruction but lack a global spatial reference. Therefore, fusing the two recovers what each modality alone cannot.* With the above motivation, we propose AVI-HT, an **a**daptive **v**isual-**I**MU fusion approach for 3D **h**and **t**racking from joint egocentric images and on-glove 6-DoF IMU signals. AVI-HT captures faithful 3D hand poses in a variety of dexterous interaction scenarios, especially when evaluated on challenging HOI scenarios with heavy occlusion. As illustrated in the second row of Figure 1, a vision-only approach UMETrack [11] produces hand pose estimates that deviate substantially from the ground truth under heavy occlusion, whereas our joint vision-IMU approach recovers poses that are more closely aligned.

The key to the success of AVI-HT lies in two complementary innovations. Firstly, for training data, we build a synchronized multi-modal capture setup by integrating egocentric cameras with a data sensing glove, and deploy a motion capture (MoCap) system for 3D ground-truth pose collection. This yields a dataset of paired egocentric video and per-finger IMU measurements with accurate 3D annotations. Secondly, for the model, we design a deep cross-sensor attention-based architecture that adaptively fuses visual features with on-glove IMU signals. As visualized in Figure 1, there are 2 to 3 IMU sensors mounted at each finger of the glove. When a finger joint is occluded, indicated in egocentric images, AVI-HT automatically places greater reliance on the IMU signal at that location (with its attention score threshold as a yellow patch) to compensate for the loss of visual evidence. Conversely, the wrist remains largely visible to the egocentric cameras even during heavy object manipulation, which provides a consistent estimate of global wrist rotation and translation that anchors the entire hand in 3D space. The combination of these two ingredients leads to significant improvements compared to vision-only and IMU-only works.

Evaluating multi-modal hand tracking methods in realistic conditions remains difficult: existing datasets are largely captured in controlled, low-occlusion settings and rarely include inertial sensor data [43, 3]. To enable rigorous evaluation under the dexterous manipulation scenarios that matter most, we collect DexGloveHOI, a multi-subject evaluation dataset focusing on hand-object-interaction (HOI). The dataset pairs egocentric video with time-synchronized on-glove IMU signals and marker-based 3D ground-truth hand poses across a diverse range of activities. Using DexGlove-

HOI, we evaluate AVI-HT with two hand representations, UMETrack [12] and MANO [32], and conduct extensive experiments comparing AVI-HT with single- and multi-modal approaches. Results demonstrate that IMU signals provide complementary cues that yield more stable and accurate hand tracking performance, with the largest gains in heavily occluded regions.

In summary, we contribute AVI-HT, an approach for 3D hand tracking from egocentric images and on-glove IMU signals. We demonstrate the key effect of the cross-sensor attention mechanism for handling occlusion, and we achieve state-of-the-art results on DexGloveHOI by comparing AVI-HT with various tracking schemes. Beyond accuracy improvements, we conduct extensive ablation and sensitivity studies to reveal how vision and IMU signals interact: we analyze the effect of partial IMU sensor availability on per-finger tracking, and study robustness to temporal misalignment between visual and inertial streams, which we believe provides practical insights for deploying multi-modal hand tracking in real-world AR/VR and robotic teleoperation systems.

2 Related Work

3D Hand Pose Tracking. In recent years, significant progress has been made in estimating 3D hand pose and shape from a single RGB image [20, 9]. Earlier methods often relied on convolutional neural networks to regress parameters of parametric hand models [5] or directly estimate 3D joint locations [33]. More recently, transformer-based architectures have emerged as the dominant paradigm. Methods such as METRO [21] and MeshGraphormer [22] leverage self-attention mechanisms to jointly model vertex-vertex and vertex-joint interactions for end-to-end hand mesh reconstruction. Building on this, HaMeR [30] demonstrated that scaling up both the training data and the vision transformer capacity leads to substantial improvements for in-the-wild hand reconstruction. However, while these vision-based methods use large-scale data to improve generalization, they fundamentally suffer from severe performance degradation under heavy occlusion, which is a common occurrence during dexterous hand-object interactions [19]. On the other hand, wearable inertial measurement unit (IMU) sensors offer an occlusion-free alternative. Existing IMU-based hand tracking systems typically rely on dense arrays of 9-DoF sensors [27]. However, the magnetometer in 9-DoF IMUs is highly susceptible to electromagnetic interference in indoor environments [15]. While more compact 6-DoF IMUs avoid magnetometer-related distortions, they lack of an absolute orientation reference and suffer from cumulative gyroscope drift [34]. In this paper, we propose AVI-HT, an adaptive visual-IMU fusion framework that overcomes the occlusion limitations of vision-only methods while leveraging sparse, low-cost 6-DoF IMUs for high-fidelity global hand pose reconstruction.

Visual-Sensor Fusion for Pose Estimation. Fusing visual data with inertial sensor signals has proven highly effective for resolving ambiguities in human poses. For full-body tracking, methods like DIP-IMU [14] and TransPose [37] have successfully utilized sparse IMU configurations, while subsequent works have combined these inertial priors with multi-view or egocentric cameras to achieve drift-free, occlusion-robust body tracking [41, 18]. Despite its success, visual-sensor fusion remains a relatively unexplored domain for fine-grained 3D hand tracking. This gap is primarily due to the lack of robust, unobtrusive IMU-equipped data gloves suitable for natural manipulation, as well as the absence of high-quality multi-modal datasets that provide synchronized egocentric vision, dense IMU signals, and ground-truth 3D hand poses during complex object interactions [40]. In this paper, we address this gap by leveraging a data sensing glove covered by per-finger IMU sensors, and DexGloveHOI, a comprehensive vision-IMU dataset for multi-modal evaluation. We demonstrate how our cross-sensor attention mechanism adaptively fuses egocentric visual features with on-glove IMU signals to achieve state-of-the-art hand tracking under heavy occlusion.

3 Technical Approach

We describe AVI-HT, our approach for 3D hand tracking from egocentric images and on-glove IMU signals. We show the overview of AVI-HT in Figure 2 and illustrate each component below.

3.1 Hand Model Representations

We evaluate AVI-HT with two hand representations to cover both offline 3D reconstruction and real-time AR/VR tracking scenarios. **MANO** [32] is a parametric model widely used in offline hand mesh recovery [30, 21, 22]. It defines a differentiable function $\mathcal{M}(\theta, \beta)$ that maps pose $\theta \in \mathbb{R}^{48}$ and

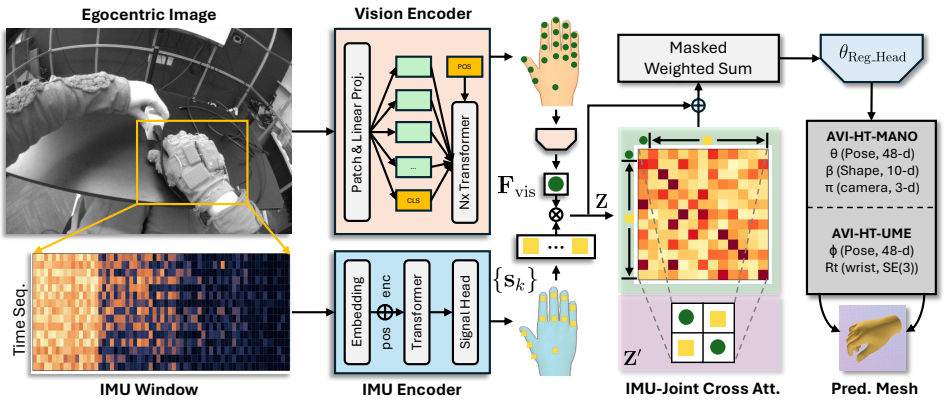

Figure 2: **Overview of AVI-HT.** A *Vision Encoder* converts the egocentric image into a global visual token $\mathbf{F}_{\text{vis}}$. An *IMU Encoder* embeds a 14 timestamp temporal window from 12 on-glove sensors into per-sensor tokens $\{\mathbf{s}_k\}$ via a transformer encoder and signal head. The two token sets are fused through a *hierarchical cross-sensor attention* module. A *kinematic prior mask* aggregates attended features based on physical sensor-joint proximity while the attention scores averaged for each IMU sensor represent its adaptive importance. The fused representation is decoded by a regression head into either **MANO** or **UMETrack** parameters for hand tracking.

shape $\beta \in \mathbb{R}^{10}$ to a mesh $M \in \mathbb{R}^{778\times3}$ and 21 joint locations. **UMETrack** [12] is an articulated model designed for real-time egocentric hand tracking in AR/VR [11, 28]. It represents pose through 22 scalar joint angles $\phi \in \mathbb{R}^{22}$, i.e., flexion/extension and abduction/adduction per joint, and recovers 21 landmark positions via forward kinematics with linear blend skinning [17].

3.2 Input Signals

Egocentric image. We use a Meta Quest headset that contains a monochrome egocentric camera. The camera captures egocentric images at 60 Hz with 512x640 raw image resolution.

On-glove IMU signals. We use a full-hand covered data sensing glove equipped with 12 6-DoF IMU sensors. Each sensor consists of 3-axis accelerometer and 3-axis gyroscope. The sensors are arranged across the back of the glove with one sensor on each proximal and distal phalanx of the four fingers, with one additional sensor on the thumb and hand back each. All IMU sensors sample at 200 Hz. Since the visual and IMU modalities operate at different rates, we synchronize them by aligning each 60 Hz video frame to the nearest IMU timestamp. For each aligned frame, we follow Sarker et al. [34] to extract a temporal window of 14 consecutive IMU samples before the synchronized timestamp. For each sample, we calculate gravity vector from 6-DoF data and append it with 3-axis gyroscope, which yields $14 \times 23 \times 3$ data window for each visual sample.

3D Mocap system. We collect 3D hand pose from a marker-based MoCap system as ground-truth. The marker positions are firstly solved to UMETrack hand pose and further converted to MANO using a Levenberg–Marquardt optimizer [23]. More description about the system is in Appendix A.

3.3 Model Architecture

AVI-HT is a general vision-IMU fusion framework that can be applied to different vision backbones and hand representations. We instantiate it with two models: **AVI-HT-UME** for real-time AR/VR tracking and **AVI-HT-MANO** for offline 3D reconstruction. Both models share the same IMU encoder and cross-attention fusion design, differing only in the vision encoder and output head.

IMU encoder. The IMU encoder is a transformer-based network [35] that processes the $14 \times 23 \times 3$ IMU input window. It firstly applies linear mapping from 3 to 69 dimensions, then sinusoidal positional encoding over the 14 timesteps, followed by a 2-layer transformer encoder with multi-head self-attention over each IMU feature. The output at the last timestep is selected and projected through a two-layer MLP to produce an IMU feature vector $\mathbf{F}_{\text{imu}} \in \mathbb{R}^{23\times d}$.

Vision encoder. AVI-HT-UME uses a ResNet-based encoder [13] that processes 96×96 monochrome egocentric images from two similar camera views. A feature transform layer (FTL) [12] lifts the 2D feature maps into a camera-disentangled 3D representation by applying SE(3) transformations derived from camera intrinsics and extrinsics. A convolutional RNN then aggregates temporal context across frames. **AVI-HT-MANO** uses a ViT-Huge [7] backbone pre-trained on body and hand pose datasets [36], which encodes a 256×192 RGB crop into a sequence of patch tokens $\mathbf{F}_{\text{vis}} \in \mathbb{R}^{N \times 1280}$. A transformer decoder with 6 cross-attention layers then attends from a learnable query token to the patch tokens to regress MANO parameters.

Hierarchical cross-sensor attention. The core of AVI-HT is a cross-attention module that fuses vision and IMU features at two hierarchical levels. We generate a global visual token $\mathbf{F}^*_{\text{vis}} \in \mathbb{R}^d$ from the visual encoder and aggregate the gravity and gyroscope features per IMU sensor, yielding a set of N_s=12 sensor tokens $\{\mathbf{s}_k\}_{k=1}^{N_s}$. Both token sets are projected to a shared embedding dimension d and concatenated into a unified sequence $\mathbf{Z} = [\mathbf{F}^*_{\text{vis}}, \mathbf{s}_1, \ldots, \mathbf{s}_{N_s}] \in \mathbb{R}^{(1+N_s)\times d}$. Since the human hand is a kinematic chain with known topology, not all sensor–sensor interactions are equally informative: an IMU mounted on the ring finger's proximal phalanx is highly relevant to the distal phalanx but carries little direct information about the thumb. We encode this inductive bias via a kinematic prior mask $\mathbf{M} \in \mathbb{R}^{(1+N_s)\times(1+N_s)}$, where $M_{ij} = -\alpha, d_{\text{geo}}(i,j)$ is the negated geodesic distance between tokens i and j on the hand skeleton graph, measured as the shortest-path hop count between two sensor mounting sites. The first-level attention is:

$$\text{Attention}^{[1]}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} + \mathbf{M}\right)\mathbf{V}, \tag{1}$$

where $\mathbf{Q} = \mathbf{Z}\mathbf{W}^Q$, $\mathbf{K} = \mathbf{Z}\mathbf{W}^K$, $\mathbf{V} = \mathbf{Z}\mathbf{W}^V$. When a hand joint is visually occluded, its attention weight shifts toward the spatially corresponding IMU sensor, which provides adaptive, occlusion-aware fusion that is anatomically grounded by the kinematic prior. To assess the relative importance of each IMU sensor, we extract the first row of the attention matrix, $\mathbf{a}_{\text{vis}} = \text{softmax}(\cdot)\{1,:\} \in \mathbb{R}^{N_s}$, which captures how strongly the visual representation attends to each sensor under the current observation. However, one potential limitation of the first-level attention is that the single visual token is outnumbered by N_s=12 IMU tokens, and this imbalance can cause the softmax mass to concentrate on the more numerous inertial modality. To mitigate this, we introduce a second-level self-attention that operates on a balanced token pair. Specifically, we average-pool the N_s sensor tokens from the first level into a single aggregated IMU token $\bar{\mathbf{s}} \in \mathbb{R}^d$ and form a compact sequence $\mathbf{Z}' = [\tilde{\mathbf{F}}^*_{\text{vis}}, \bar{\mathbf{s}}] \in \mathbb{R}^{2\times d}$. A standard self-attention layer then re-calibrates the relative contribution of vision and inertial sensing. Finally, the two resulting tokens are mean-pooled into a global hand representation in d dimension, which is passed to the respective regression heads.

Output and Losses. For **AVI-HT-UME**, the regression head decodes joint angles $\phi \in \mathbb{R}^{22}$, wrist transformation $(\mathbf{R}, \mathbf{t}) \in SE(3)$, and per-landmark uncertainty σ_ℓ. 3D landmarks are recovered via differentiable forward kinematics. The primary loss is a landmark NLL that jointly supervises position accuracy and uncertainty calibration supplemented by a joint angle loss [12]. For **AVI-HT-MANO**, the regression head decodes the fused joint tokens into MANO pose θ, shape β, and camera π via linear projections. Training uses a combination of 3D and 2D losses and an adversarial loss from a discriminator trained on hand shape, pose, and per-joint rotations separately [30].

4 Evaluation Dataset

A key challenge in advancing multi-modal hand tracking is the lack of suitable evaluation data. Existing 3D hand pose datasets [43, 10] are predominantly vision-only, captured in controlled studio settings [25], and limited to a narrow range of hand poses [39]. They do not include inertial sensor data, nor do they reflect the diversity of hand behaviors encountered in real-world AR/VR use, such as such as fine-grained tool manipulation. Without such data, it is difficult to assess whether fusing vision with IMU signals genuinely improves tracking, or merely adds complexity.

To address this gap, we collect a multi-modal evaluation dataset, **DexGloveHOI**, that pairs synchronized egocentric video with dense on-glove IMU signals and high-fidelity marker-based motion capture ground truth. DexGloveHOI is designed with two goals: (1) to cover the full spectrum of dexterous hand use that an AR/VR or robotic system must handle, and (2) to provide the multi-modal signals needed to evaluate vision-IMU fusion approaches.

Figure 3: **Overview of DexGloveHOI dataset.** *(Left)* Sample egocentric images showing diverse activities captured with the data sensing glove. *(Top right)* Distribution of activity categories in DexGloveHOI. *(Bottom right)* Key statistics of dataset. Each frame is paired with synchronized on-glove IMU signals and ground truth from MoCap system in both MANO and UMETrack representations.

We design the capture protocol to span the breadth of hand behaviors relevant to AR/VR input that are organized into 17 distinct dexterous manipulation tasks per session across several complementary categories. We show the full list of the task names and sampled egocentric images in Figure 3. Task instructions are recorded on video before each activity, and task ordering is determined per session. Each session lasts approximately 45-55 minutes of usable capture across 4 participants. Before the formal data collection, we conduct calibration for MoCap markers and IMU orientation. We finally collected 3.5 hours of quality-controlled multi-modal data. This yields over 100K+ egocentric camera frames, 2.4M IMU samples, and 720 GB of synchronized sensor recordings.

5 Experiments

We firstly compare 3D hand tracking accuracy against single- and multi-modal baselines under both hand representations (5.1). We then ablate the contribution of individual IMU sensors and activity types (5.2), and study the model's sensitivity to IMU signal quality (5.3).

Implementation Details We train AVI-HT for 40 epochs with a batch size of 48 on $8\times$ NVIDIA H100 GPUs. We use the Adam optimizer with a learning rate of 7.89×10^{-4}, weight decay of 10^{-6}, momentum of 0.9, and $\epsilon = 10^{-8}$, combined with a step-decay learning rate scheduler that reduces the learning rate by $10\times$ at epochs 30. Data augmentation includes fast-motion augmentation, per-frame image noise injection ($\sigma = 10$), and gamma jittering ranging from 0.5 to 1.7. Training is initialized from pretrained UMETrack checkpoint for AVI-HT-UME and HaMeR for AVI-HT-MANO. The total training time for both are around 24.5 and 42 hours, respectively.

5.1 3D Hand Tracking Accuracy

The two hand representations use related but non-interchangeable evaluation metrics. We therefore report results under each representation separately to enable fair comparison.

UMETrack hand model. We evaluate AVI-HT on DexGloveHOI against three baselines that use UMETrack hand models: **UMETrack** [12] (vision-only), **UMETrack + EKF** (vision-IMU fusion by a post-hoc Extended Kalman Filter) [16], and **IMU-Tracker** (6-DoF IMU-only) [34]. Table 1 reports mean keypoint position error (MKPE), fingertip-only MKPE (F.MKPE), and their hand-root GT oriented (**T**ransformed) variants (MKPE.T, F.MKPE.T), as well as PUK AUC (P.A) and transformed variant (P.A.T). We observe that AVI-HT outperforms all other baselines across all evaluation metrics. For example, AVI-HT achieves the lowest MKPE (10.359 mm) and F.MKPE (13.253 mm), improving over the vision-only UMETrack baseline by 16.1% and 24.2% respectively. AVI-HT also outperforms UMETrack + EKF, which applies multi-modal fusion as a post-processing filtering step, which

Table 1: 3D Hand Tracking Accuracy for UMETrack Pose

Model	**MKPE** ↓	**F.MKPE** ↓	**MKPE.T** ↓	**F.MKPE.T** ↓	**P.A** ↑	**P.A.T** ↑
UMETrack	12.342	17.478	9.359	17.798	76.059	81.433
UMETrack + EKF	15.406	18.658	7.891	14.232	70.903	84.210
IMU-Tracker	-	-	7.151	12.923	-	85.769
AVI-HT	**10.359**	**13.253**	**7.021**	**12.785**	**79.315**	**85.859**

Table 2: 3D Hand Tracking Accuracy for MANO Pose

	PA-MPJPE ↓	**PA-MPVPE** ↓	**F@5** ↑	**F@15** ↑
HaMeR	13.754	12.736	0.516	0.882
AVI-HT	**10.519**	**9.265**	**0.628**	**0.936**

demonstrates that the learned cross-sensor attention provides more effective signal fusion than a hand-crafted filter. The IMU-only tracker achieves competitive metrics, showing that IMU signals carry strong fingertip information. However, it cannot estimate global wrist position due to 6-DoF IMUs. Such limitation highlights the complementary nature of vision-IMU that AVI-HT exploits.

Figure 4-top provides a per-joint breakdown of absolute joint angle error across all 22 degrees of freedom. AVI-HT consistently reduces error compared to UMETrack and UMETrack + EKF, with the largest improvements on the MCP flexion joints that are most prone to visual occlusion. Similarly, Figure 4-bot shows the cumulative distribution of joint angle errors across all DexGloveHOI samples. AVI-HT consistently achieves a higher percentage of samples below each error threshold, with the gap being most pronounced at stricter thresholds like $<5°$. This confirms that AVI-HT improves more samples towards low-error range which is essential for high-precision dexterous manipulation.

MANO hand model. We evaluate AVI-HT against HaMeR [30], one of the state-of-the-art vision-only hand mesh recovery methods, on DexGloveHOI. Table 2 reports Procrustes-aligned mean per-joint position error (PA-MPJPE), Procrustes-aligned mean per-vertex position error (PA-MPVPE), and F-scores at 5 mm and 15 mm thresholds. AVI-HT reduces PA-MPJPE from 13.754 mm to 10.519 mm (23.5% improvement) and PA-MPVPE from 12.736 mm to 9.265 mm (27.3% improvement), while increasing F@5 from 0.516 to 0.628 and F@15 from 0.882 to 0.936. These results demonstrate that the adaptive vision-IMU fusion provides substantial improvements even for a strong transformer-based baseline operating on the MANO representation.

Qualitative results. Figure 5 presents qualitative comparisons on two egocentric hand-object interaction sequences from DexGloveHOI. Each row depicts a temporal progression of frames where finger occlusion varies as the hand manipulates an object. We compare 3D tracked hand pose from three sources: ground truth from MoCap system, AVI-HT and UMETrack. For both rows, we observe that AVI-HT and UMETrack achieve competitive performance when all fingers are visible from the egocentric camera. However, as more fingers are occluded during the manipulation, AVI-HT is still able to track the ground truth closely while UMETrack deviates substantially, particularly for the unseen fingers. The results demonstrate that the cross-sensor attention mechanism of AVI-HT adaptively leverages IMU cues to maintain accurate tracking when visual evidence is lost to occlusion. We show more visualization examples and hand tracking videos with attention scores in Appendix B.

5.2 Ablation Study

We ablate the contribution of individual IMU sensors and activity types to understand which components of the multi-modal input drive AVI-HT's improvements. Figure 6 presents two heatmaps in terms of *MKPE.T Gap* which is AVI-HT − UMETrack, thus negative values indicate AVI-HT is better). The left heatmap shows **per-finger IMU contribution**: for each IMU sensor group (rows) and the finger being evaluated (columns), we report the MKPE.T Gap when only that sensor group is provided. The diagonal entries, where the sensor and the evaluated finger correspond to the same finger, consistently show the largest improvement. The results confirm that each IMU sensor contributes most to its own finger's tracking. Off-diagonal entries are generally smaller. However, several

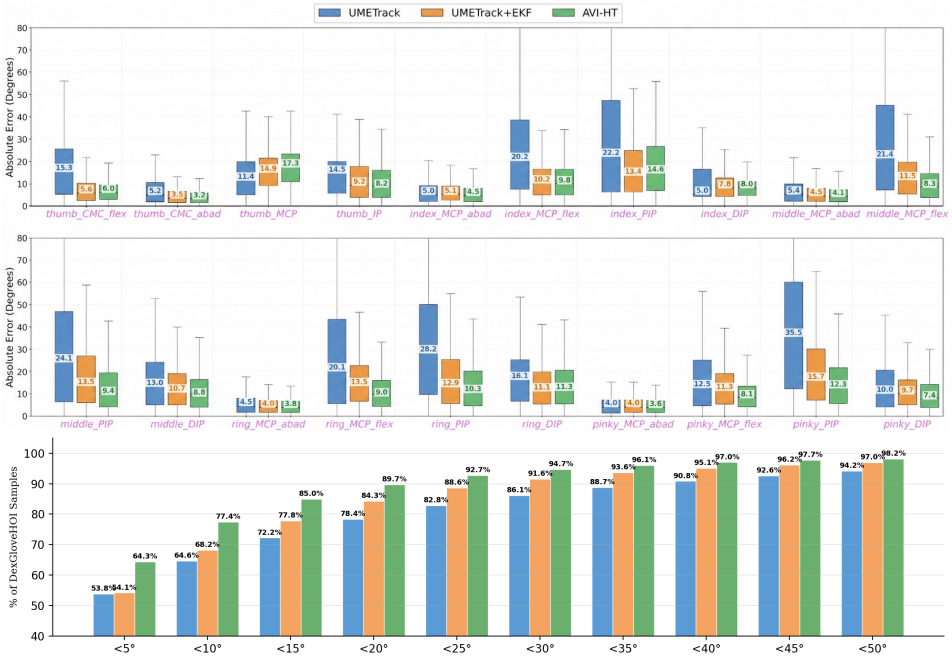

Figure 4: **Breakdown evaluation for AVI-HT.** *(Top)* Absolute angle error across all 22 degrees of freedom for UMETrack, UMETrack + EKF, and AVI-HT. AVI-HT reduces error on most joints, with the largest gains on MCP flexion joints. *(Bottom)* Cumulative distribution of joint angle errors over all DexGloveHOI samples. AVI-HT achieves a higher fraction of samples below each error threshold.

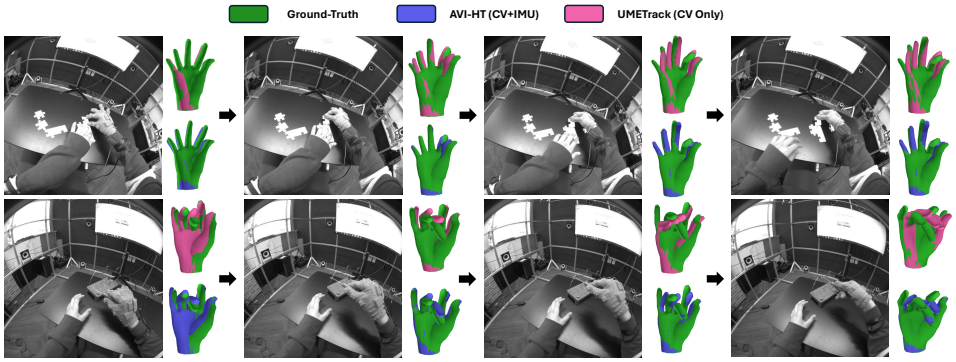

Figure 5: **Qualitative tracking comparison.** Two hand-object interaction sequences with finger occlusion moments. AVI-HT closely tracks the ground truth under heavy occlusion, while UMETrack deviates at unseen fingers from egocentric views.

off-diagonal cells also show notable negative values, particularly the ones closer to the diagonal. We attribute this to the kinematic coupling of the hand. That is, neighboring fingers share tendons and move in a correlated fashion, so an IMU on the ring finger captures motion that is predictive of the middle and pinky fingers as well. This validates our cross-sensor attention module design: the model learns to route each finger's information primarily through its spatially corresponding sensor, while still leveraging cross-finger cues when they are available.

The right heatmap shows **per-activity-type breakdown**: for each IMU sensor group (rows) and activity category (columns), we report MKPE.T improvement with the same AVI-HT − UMETrack metric. Activities involving heavy occlusion and fine-grained manipulation (e.g., cutting, scissors, screwdriver) show the largest improvements, which confirms that IMU fusion provides the greatest benefit precisely in the scenarios where vision alone struggles most.

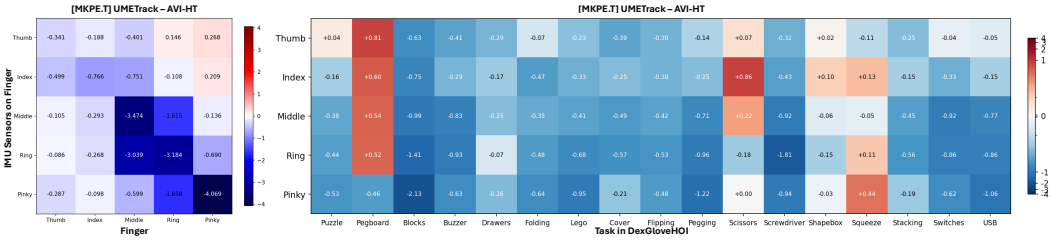

Figure 6: **Per-finger IMU contribution.** *(Left)* MKPE.T Gap when only the corresponding IMU sensor group (row) is provided during training. *(Right)* MKPE.T Gap broken down by activity type.

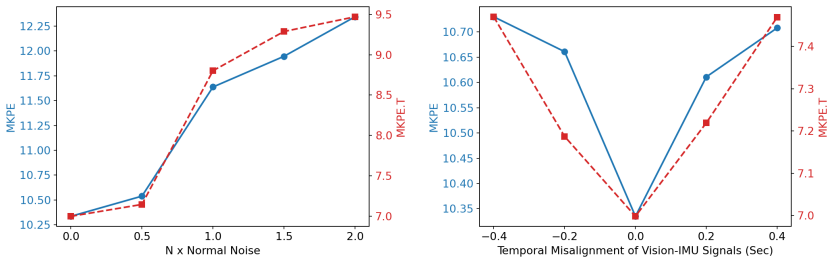

Figure 7: Sensitivity to IMU signal quality. (Left) MKPE and MKPE.T under increasing additive Gaussian noise. (Right) MKPE and MKPE.T under temporal misalignment between vision and IMU.

5.3 Sensitivity Study

To assess the robustness of AVI-HT to degraded IMU signals, we conduct two sensitivity analyses shown in Figure 7. We firstly inject additive Gaussian noise into the IMU input at increasing levels from $\times 0$ to $\times 2$ of the native noise floor and train and evaluate AVI-HT in terms of MKPE and MKPE.T metrics. We observe that AVI-HT degrades with MKPE rising from $\sim$10.335 mm to $\sim$12.344 mm at $2\times$ noise, while MKPE.T increases from $\sim$6.998 mm to $\sim$9.469 mm. The results indicate that AVI-HT is robust to moderate perturbations (e.g., $< \times 0.5$) but suffers more with higher perturbations. We then artificially shift the IMU window relative to the visual frame by -0.4 to $+0.4$ seconds and measure MKPE and MKPE.T for AVI-HT-UME. Both error curves form a clear V-shape centered at zero shift, which confirms that the temporal alignment between modalities is critical. We observe that the performance is relatively more stable than the noise injection, which suggests that moderate temporal misalignment preserves the overall motion dynamics within the IMU window.

6 Conclusion

We presented AVI-HT, an adaptive vision-IMU fusion approach for 3D hand tracking that introduces cross-sensor attention to fuse egocentric visual features with on-glove IMU signals. The design is general across vision backbones and hand models. Experiment results on DexGloveHOI show consistent improvements over various baselines on both UMETrack and MANO hand models, with the largest gains under heavy occlusion. Ablation study reveals that each IMU sensor primarily benefits its own finger with additional cross-finger transfer from kinematic coupling, and sensitivity studies confirm certain robustness to IMU noise and temporal misalignment.

Limitations. The data sensing glove alters the visual appearance of the hand, which may introduce a domain gap when the egocentric image is also consumed by additional downstream pipelines that expect bare-hand input. A potential mitigation is to use hand avatar methods [6, 40] that can render a photorealistic bare-hand image conditioned on the tracked pose, which effectively removes the glove from the visual stream. In addition, since our experiments use a specific type of data sensing glove, the generalization to other gloves with different sensor layouts, form factors, or IMU specifications remains to be validated. We hope this work motivates further research into glove-agnostic fusion methods and cross-device generalization for multi-modal hand tracking.

References

[1] Sridhar Pandian Arunachalam, Sneha Silwal, Ben Evans, and Lerrel Pinto. Dexterous imitation made easy: A learning-based framework for efficient dexterous manipulation. In *2023 ieee international conference on robotics and automation (icra)*, pages 5954–5961. IEEE, 2023.

[2] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European conference on computer vision*, pages 348–367. Springer, 2022.

[3] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, et al. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7071, 2025.

[4] Yiming Bao, Xu Zhao, and Dahong Qian. Fusepose: Imu-vision sensor fusion in kinematic space for parametric human pose estimation. *IEEE Transactions on Multimedia*, 25:7736–7746, 2022.

[5] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10843–10852, 2019.

[6] Xingyu Chen, Baoyuan Wang, and Heung-Yeung Shum. Hand avatar: Free-pose hand animation and rendering from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8683–8693, 2023.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023.

[9] Shaoxiang Guo, Eric Rigall, Yakun Ju, and Junyu Dong. 3d hand pose estimation from monocular rgb with feature interaction module. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5293–5306, 2022.

[10] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020.

[11] Shangchen Han, Beibei Liu, Randi Cabezas, Christopher D Twigg, Peizhao Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Zheng Wang, et al. Megatrack: monochrome egocentric articulated hand-tracking for virtual reality. *ACM Transactions on Graphics (ToG)*, 39(4):87–1, 2020.

[12] Shangchen Han, Po-chen Wu, Yubo Zhang, Beibei Liu, Linguang Zhang, Zheng Wang, Weiguang Si, Peizhao Zhang, Yujun Cai, Tomas Hodan, et al. Umetrack: Unified multi-view end-to-end hand tracking for vr. In *SIGGRAPH Asia 2022 conference papers*, pages 1–9, 2022.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics (TOG)*, 37(6):1–15, 2018.

[15] Daniel Laidig and Thomas Seel. Vqf: Highly accurate imu orientation estimation with bias estimation and magnetic disturbance rejection. *Information Fusion*, 91:187–204, 2023.

[16] Yu Lei, Yi Deng, Lin Dong, Xiaohui Li, Xiangnan Li, and Zhi Su. A novel sensor fusion approach for precise hand tracking in virtual reality-based human—computer interaction. *Biomimetics*, 8(3):326, 2023.

[17] John P Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 811–818. 2023.

[18] Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17142–17151, 2023.

[19] Jiawen Li, Fei Jiang, Dandan Zhu, and Aimin Zhou. Handnet: Occlusion-robust 3d hand mesh reconstruction with prior information. *Available at SSRN 5244153*, 2025.

[20] Fanqing Lin, Connor Wilhelm, and Tony Martinez. Two-hand global 3d pose estimation using monocular rgb. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2373–2381, 2021.

[21] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1954–1963, 2021.

[22] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12939–12948, 2021.

[23] Manolis LA Lourakis and Antonis A Argyros. Is levenberg-marquardt the most efficient optimization algorithm for implementing bundle adjustment? In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1526–1531. IEEE, 2005.

[24] Andualem T Maereg, Emanuele L Secco, Tayachew F Agidew, David Reid, and Atulya K Nagar. A low-cost, wearable opto-inertial 6-dof hand pose tracking system for vr. *Technologies*, 5(3):49, 2017.

[25] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision*, pages 548–564. Springer, 2020.

[26] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *Proceedings of the IEEE international conference on computer vision*, pages 1154–1163, 2017.

[27] Chaithanya Kumar Mummadi, Frederic Philips Peter Leo, Keshav Deep Verma, Shivaji Kasireddy, Philipp M Scholl, Jochen Kempfle, and Kristof Van Laerhoven. Real-time and embedded detection of hand gestures with an imu-based glove. In *Informatics*, volume 5, page 28. MDPI, 2018.

[28] Takehiko Ohkawa, Ryosuke Furuta, and Yoichi Sato. Efficient annotation and learning for 3d hand pose estimation: A survey. *International Journal of Computer Vision*, 131(12):3193–3206, 2023.

[29] Shaohua Pan, Qi Ma, Xinyu Yi, Weifeng Hu, Xiong Wang, Xingkang Zhou, Jijunnan Li, and Feng Xu. Fusing monocular images and sparse imu signals for real-time human motion capture. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023.

[30] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2024.

[31] Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system. *arXiv preprint arXiv:2307.04577*, 2023.

[32] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), November 2017.

[33] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1749–1759, 2021.

[34] Anik Sarker, Ziyi Kou, Ergys Ristani, Li Guan, and Taylor Niehues. Real-time hand pose tracking using 6-axis imus. In *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction*, pages 1182–1191, 2026.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[36] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022.

[37] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions On Graphics (TOG)*, 40(4):1–13, 2021.

[38] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020.

[39] Haoran Zhang, Shuanghao Bai, Wanqi Zhou, Yuedi Zhang, Qi Zhang, Pengxiang Ding, Cheng Chi, Donglin Wang, and Badong Chen. Vcot-grasp: Grasp foundation models with visual chain-of-thought reasoning for language-driven grasp generation. *arXiv preprint arXiv:2510.05827*, 2025.

[40] Xinyu Zhang, Ziyi Kou, Chuan Qin, Mia Huang, Ergys Ristani, Ankit Kumar, Lele Chen, Kun He, Abdeslam Boularias, and Li Guan. Glove2hand: Synthesizing natural hand-object interaction from multi-modal sensing gloves. *arXiv preprint arXiv:2603.20850*, 2026.

[41] Zhe Zhang, Chunyu Wang, Wenhu Qin, and Wenjun Zeng. Fusing wearable imus with multi-view images for human pose estimation: A geometric approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2200–2209, 2020.

[42] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*, pages 4903–4911, 2017.

[43] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 813–822, 2019.

A Marker Based Mocap System Setup

To obtain ground-truth 3D hand pose annotations, we construct a dedicated marker-based optical motion capture environment. As shown in Figure 8, the capture volume is defined by a modular aluminum frame structure. We mount 12 infrared cameras at varying heights and angles around the frame to ensure dense, multi-view coverage of the capture volume and minimize marker occlusion. The back wall is covered with a dark matte backdrop to reduce infrared reflections that could introduce spurious marker detections. Subjects are seated in a fixed chair at the center of the capture volume, with the designated capture area demarcated on the floor. This seated configuration constrains the hand workspace to a consistent region within the calibrated volume, maximizing marker visibility across all cameras and ensuring reliable sub-millimeter tracking accuracy throughout data collection.

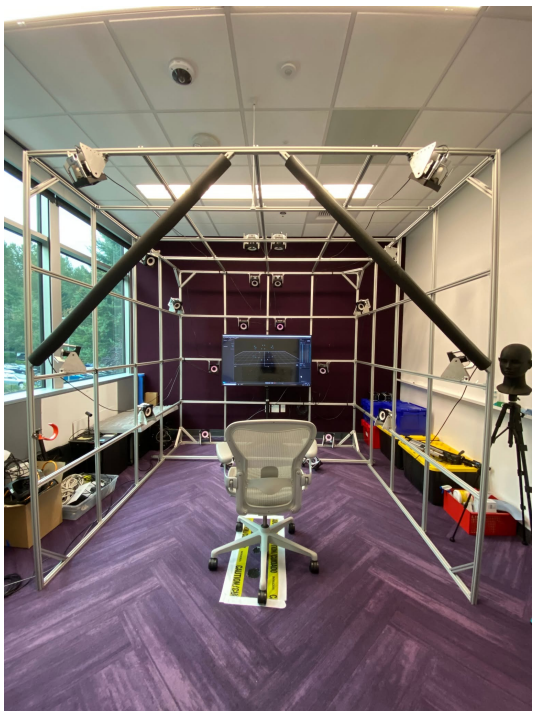

Figure 8: Marker Based Mocap System Setup

B More Visualization for AVI-HT

We provide additional qualitative comparisons between AVI-HT and UMETrack in Figure 9, illustrating that AVI-HT achieves more accurate hand tracking during hand-object interaction under visual occlusion.

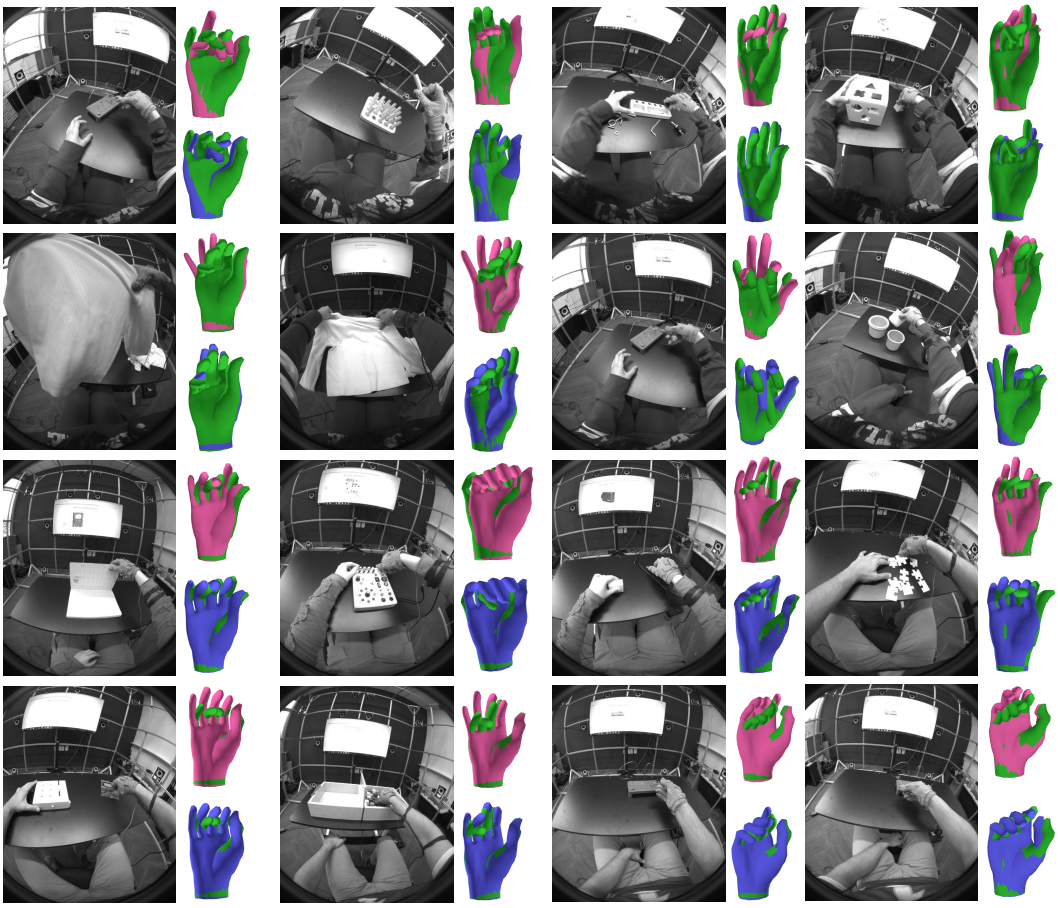

Figure 9: More qualitative comparison between AVI-HT and UMETrack.