AVI-HT: Adaptive Vision-IMU Fusion for 3D Gloved 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 images with on-glove 6-DoF IMU signals. AVI-HT 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. A hand-object contact evaluation shows that AVI-HT keeps finger contacts with the manipulated object more accurately and consistently over time, reducing contact losses by 8.5%. Ablation studies further show that our fusion design outperforms simpler schemes and that the gains grow with occlusion and hold across subjects.

1 Introduction

Consider the set of egocentric images in the first row of Figure 1, where a human hand manipulates the note book 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,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

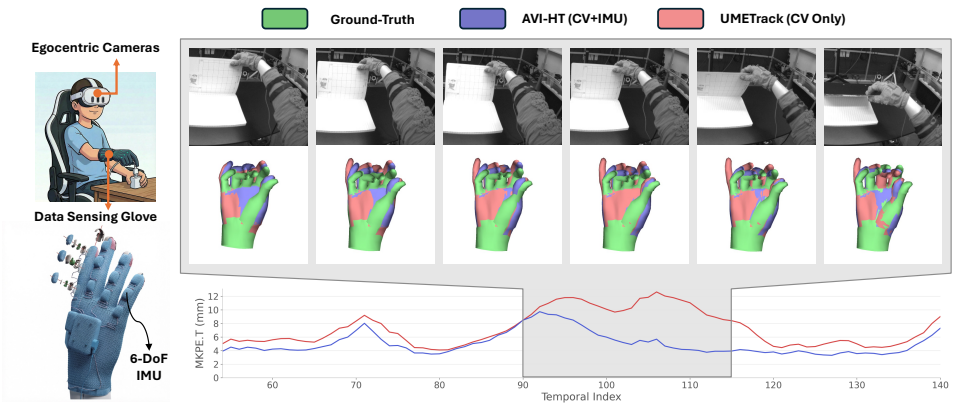

Figure 1: **3D hand tracking from egocentric vision and on-glove IMUs.** *(Left)* Our multi-modal data capture setup that consists of an egocentric camera and a data sensing glove with 12 6-DoF IMU sensors. *(Right)* Qualitative comparison on 4 egocentric frames. For each frame, the top row shows an 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 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 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 DexGloveHOI, 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 cross-sensor attention for handling occlusion and achieve state-of-the-art results on DexGloveHOI. Beyond keypoint accuracy, AVI-HT keeps hand-object contacts more accurately and consistently over time. Extensive ablations show that our fusion design outperforms simpler schemes, the gains grow with occlusion and hold across subjects, and each IMU sensor contributes most to its own finger, providing practical insights for deploying multi-modal hand tracking in AR/VR and robotic teleoperation.

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 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

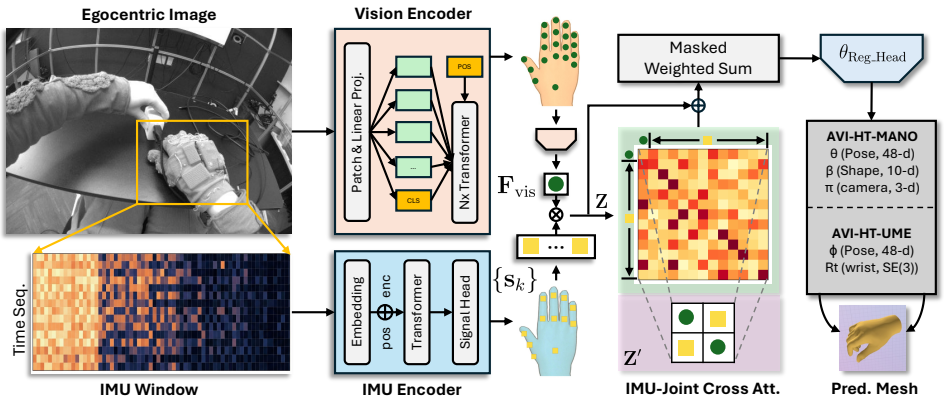

Figure 2: **Overview of AVI-HT.** A *Vision Encoder* converts the egocentric image into a global visual token $\mathbf{F}_{\text{vis}}$. An *IMU Encoder* embeds a 14 timestamp temporal window from 12 on-glove sensors into per-sensor tokens $\{\mathbf{s}_k\}$ via a transformer encoder and signal head. 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.

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 B.

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



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.

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

Table 1: Ablation of fusion design for AVI-HT-UME on DexGloveHOI.

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
AVI-HT (early concat)	11.928	16.935	8.649	16.545	76.208	82.629
AVI-HT (late fusion)	12.139	16.492	7.054	13.233	75.818	85.825
AVI-HT (pooled IMU token)	11.956	14.571	7.940	14.427	76.213	84.027
AVI-HT (fusion-1 only)	11.113	14.088	7.456	13.740	77.818	84.994
AVI-HT (w/o k-prior)	10.901	13.898	7.406	13.611	78.238	85.090
AVI-HT (full)	**10.359**	**13.253**	**7.021**	**12.785**	**79.315**	**85.859**

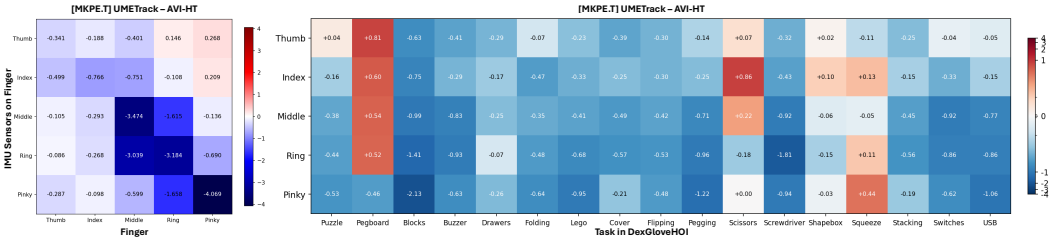

Figure 4: **Per-finger IMU contribution.** *(Left)* MKPE.T Gap when only the corresponding IMU sensor group (row) is provided during training. *(Right)* MKPE.T Gap broken down by activity type.

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.

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 first ablate the fusion design and the contribution of the glove IMU sensors across fingers, activities, occlusion levels, and subjects (5.1). We then compare 3D hand tracking accuracy against single- and multi-modal baselines under both UMETrack and MANO hand representations (5.2). We

Table 2: Per-subject results of AVI-HT-UME on DexGloveHOI.

Subject	**MKPE** ↓	**F.MKPE** ↓	**MKPE.T** ↓	**F.MKPE.T** ↓	**P.A** ↑	**P.A.T** ↑
Subject 1	10.847	13.841	7.069	12.792	78.324	85.762
Subject 2	9.624	12.190	6.548	12.469	80.797	86.811
Subject 3	11.266	14.448	7.722	13.814	77.486	84.450
Subject 4	8.852	11.512	6.266	11.311	82.333	87.371

Table 3: Fingertip error binned by the number of occluded keypoints per frame.

	F.MKPE ↓			**F.MKPE.T** ↓		
Occluded keypoints	UMETrack	AVI-HT	Gap	UMETrack	AVI-HT	Gap
$[0, 5)$	16.525	13.034	3.491	16.532	12.605	3.927
$[5, 10)$	16.704	13.226	3.478	17.383	12.745	4.638
$[10, 15)$	17.835	13.373	4.462	18.744	12.936	5.808
$[15, 20)$	21.741	15.539	6.202	21.067	13.585	7.482

further evaluate whether the tracked hand preserves hand-object contacts accurately and consistently over time (5.3). We finally study the model's sensitivity to IMU signal quality in Appendix D.

Implementation Details AVI-HT is trained on a separate ∼8M-image glove dataset, and DexGloveHOI is used for evaluation only, with subjects, sessions, and tasks disjoint from training. AVI-HT-UME and AVI-HT-MANO are initialized from pretrained UMETrack and HaMeR, and these baselines are fine-tuned on the same training set with the same backbones (vision-only). Training details are in Appendix A.

5.1 Ablation Study for IMU Sensors

We ablate AVI-HT along four axes to understand which components of the multi-modal design drive its improvements: the fusion architecture, the contribution of individual IMU sensors and activity types, the severity of visual occlusion, and generalization across subjects.

Fusion design. Table 1 compares AVI-HT-UME with simple fusion baselines and component removals: **early concat** (concatenated backbone features with a linear projection, no attention), **late fusion** (a separate IMU decoder whose joint angles are fused at the output), **pooled IMU token** (average-pooling over IMU timesteps instead of the last one), **fusion-1 only** (no second-level attention), and **w/o k-prior** (no kinematic prior mask $\mathbf{M}$). Naive fusion barely improves the global MKPE over UMETrack, since indiscriminately mixing in noisy IMU features hurts when vision alone suffices, whereas attention raises the IMU weight only when occlusion makes it beneficial. Removing the kinematic prior or the second-level attention degrades all metrics (MKPE $+0.54$ mm and $+0.75$ mm), showing that the gains stem from the hierarchical, topology-aware design rather than from added IMU features or parameters.

Per-subject generalization. Table 2 reports AVI-HT-UME results for each of the four held-out DexGloveHOI subjects. AVI-HT performs consistently across subjects with small variance, and every subject outperforms the aggregate vision-only UMETrack result (Table 4) on all six metrics. Since no DexGloveHOI subject is seen during training, this indicates that the gains come from vision-IMU fusion rather than subject-specific overfitting.

Per-finger and per-activity contribution. Figure 4 reports the *MKPE.T Gap* (AVI-HT − UMETrack; negative is better). The left heatmap shows **per-finger IMU contribution**: when only one IMU sensor group (row) is provided, the largest improvement consistently falls on its own finger (diagonal). Near-diagonal cells also improve notably due to kinematic coupling, as neighboring fingers share tendons and move together. It matches our cross-sensor attention design, which routes each finger information mainly through its sensor while still exploiting cross-finger cues.

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

Table 4: 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 5: 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**

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.

Occlusion severity. To directly quantify the benefit under occlusion, we label each keypoint in each frame as occluded or visible using the camera parameters and the 3D ground-truth pose, and bin the frames by the number of occluded keypoints. Table 3 reports F.MKPE and F.MKPE.T, the metrics most sensitive to the finger-mounted IMUs. As occlusion increases, the UMETrack error grows by 5.2 mm in F.MKPE, whereas AVI-HT grows by only 2.5 mm. Consequently, the AVI-HT advantage grows monotonically with occlusion severity, from a 3.5–3.9 mm gap in the least-occluded bin to 6.2–7.5 mm in the most-occluded bin. This confirms that the gains are concentrated where visual evidence is lost and scale with the severity of occlusion.

5.2 3D Hand Tracking Accuracy

UMETrack hand model. We compare AVI-HT on DexGloveHOI with three baselines that are finetuned on the same vision dataset: **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 4 reports the mean keypoint position error (MKPE), fingertip-only MKPE (F.MKPE), their variants aligned at the GT hand root (**T**ransformed; MKPE.T, F.MKPE.T), and PCK AUC (P.A, P.A.T). AVI-HT outperforms all baselines on every metric. It achieves the lowest MKPE and F.MKPE, improving over UMETrack by 16.1% and 24.2%. AVI-HT also outperforms UMETrack + EKF, showing that learned cross-sensor attention fuses the signals more effectively than a hand-crafted post-hoc filter. IMU-Tracker achieves competitive fingertip accuracy but cannot estimate the global wrist position from 6-DoF IMUs, which highlights the complementary vision-IMU.

Figure 5 (top) breaks down the absolute joint angle error over all 22 degrees of freedom. AVI-HT reduces the error relative to both UMETrack and UMETrack + EKF, with the largest gains on the MCP flexion joints that are most prone to occlusion. The cumulative error distribution (bottom) shows that AVI-HT places a larger fraction of samples below every threshold, with the largest gap at strict thresholds such as $<5^\circ$, which matters for high-precision dexterous manipulation.

MANO hand model. We compare AVI-HT with the finetuned HaMeR [30], a state-of-the-art vision-only hand mesh recovery method. Table 6 reports Procrustes-aligned mean per-joint and per-vertex position errors (PA-MPJPE, PA-MPVPE) and F-scores at 5 mm and 15 mm. AVI-HT reduces PA-MPJPE from 13.754 mm to 10.519 mm (23.5%) and PA-MPVPE from 12.736 mm to 9.265 mm (27.3%), and raises F@5 from 0.516 to 0.628 and F@15 from 0.882 to 0.936, showing that the fusion substantially improves even a strong transformer-based baseline.

Qualitative results. Figure 6 compares the MoCap ground truth, AVI-HT and UMETrack on two egocentric hand-object interaction sequences in which finger occlusion varies during manipulation. Both methods perform similarly when all fingers are visible. As fingers become occluded, AVI-HT still tracks the ground truth closely, whereas UMETrack deviates substantially on the unseen fingers,

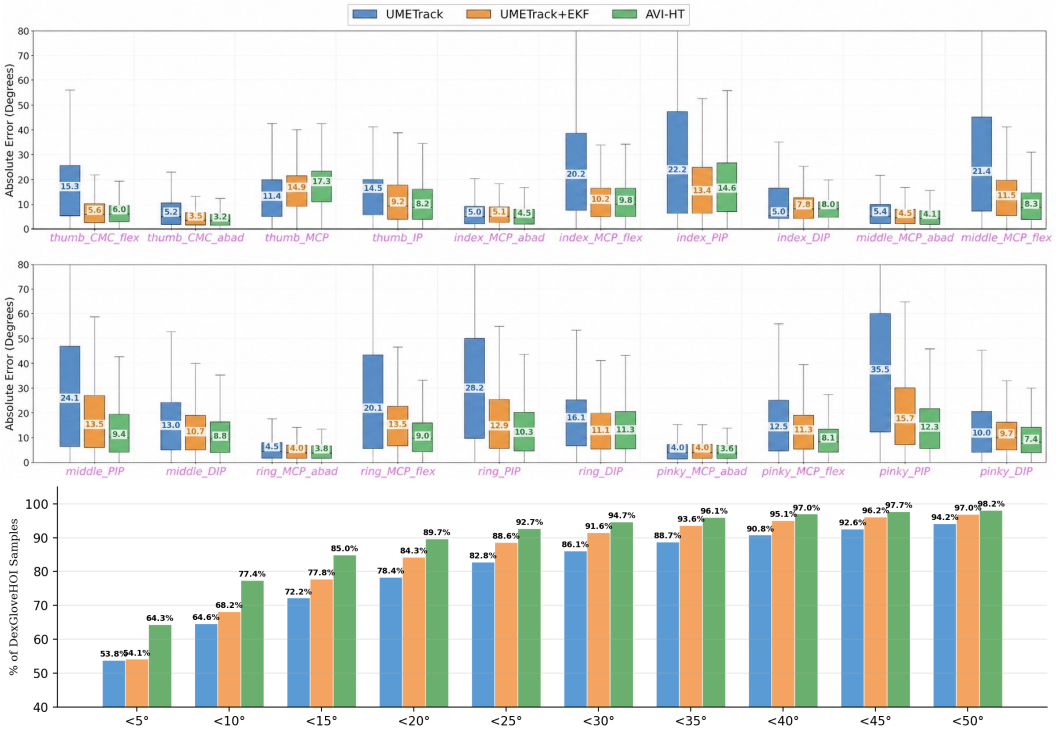

Figure 5: **Breakdown evaluation for AVI-HT.** *(Top)* Absolute angle error across all 22 degrees of freedom for UMETrack, UMETrack + EKF, and AVI-HT. AVI-HT reduces error on most joints, with the largest gains on MCP flexion joints. *(Bottom)* Cumulative distribution of joint angle errors over all DexGloveHOI samples. AVI-HT achieves a higher fraction of samples below each error threshold.

Table 6: 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**

showing that the cross-sensor attention leverages IMU cues when visual evidence is lost. More examples and tracking videos with attention scores are in Appendix C.

5.3 Hand-Object Contact Evaluation

Keypoint errors do not show whether the fingers stay on a grasped object. We therefore evaluate hand-object contact on 15 wooden-block sessions from the DexGloveHOI dataset, with the block pose from MoCap. With the predicted wrist aligned to the ground truth, a hand vertex within 2 mm of the block is a ground-truth contact, and it is kept if the predicted vertex lies within τ of the block. **Spatial** metrics are per frame: contact displacement (Disp.) on the block surface and retention (Ret@τ), the fraction of contacts kept. **Temporal** metrics follow each contact lasting at least 0.5 s: the fraction of time kept (Capt.), the longest uninterrupted fraction (Longest), and losses per contact-vertex second. As shown in Table 7, AVI-HT improves every metric, reducing displacement by 0.95 mm and contact losses by 8.5% (29.7% at $\tau = 10$ mm). Figure 7 shows that UMETrack repeatedly loses occluded fingertip contacts that AVI-HT keeps.

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

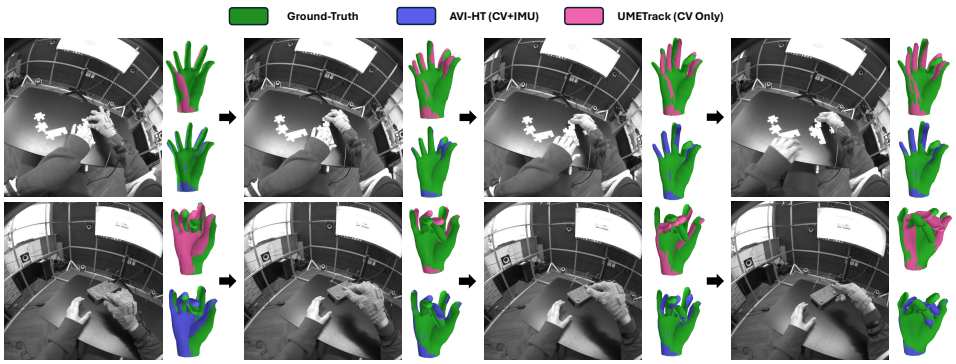

Figure 6: **Qualitative tracking comparison.** Two hand-object interaction sequences with finger occlusion moments. AVI-HT closely tracks the ground truth under heavy occlusion, while UMETrack deviates at unseen fingers from egocentric views.

Table 7: Hand-object contact evaluation. Temporal metrics use $\tau = 5$ mm. Disp. is in mm and Loss/s in losses per second; others in %.

	Spatial			Temporal		
Model	**Disp.** ↓	**Ret@5** ↑	**Ret@10** ↑	**Capt.** ↑	**Longest** ↑	**Loss/s** ↓
UMETrack	9.91	56.60	84.12	58.09	50.13	0.911
AVI-HT	**8.96**	**59.41**	**86.73**	**61.20**	**53.84**	**0.834**

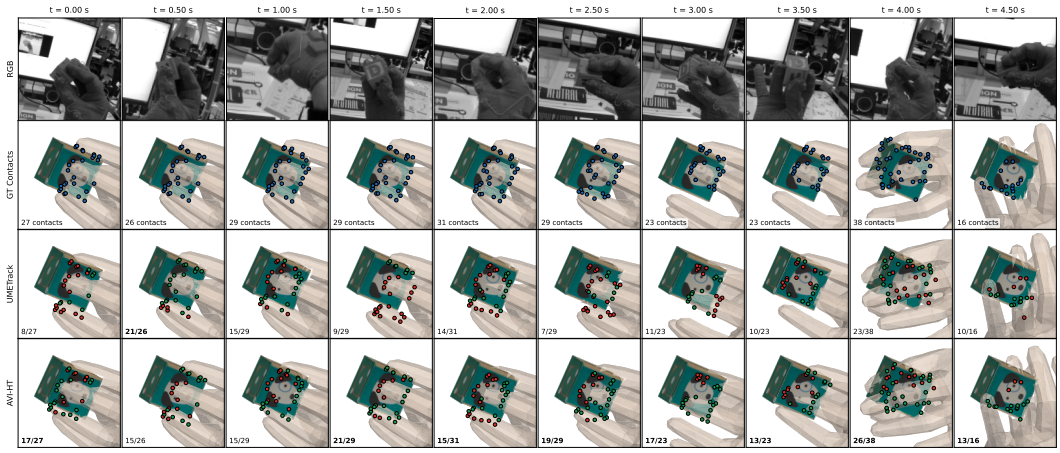

Figure 7: **Hand-object contact over time** (0.5 s apart). Ground-truth contacts (blue) are green in each method's row if kept within 5 mm and red if missed; X/Y counts kept contacts.

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 changes the hand's appearance, which may cause a domain gap for downstream pipelines that expect bare-hand images. Hand avatar methods [6, 40] could mitigate this by rendering a photorealistic bare hand from the tracked pose. In addition, we use one type of glove, so generalization to other sensor layouts, form factors, and IMU specifications remains to be validated, motivating future work on glove-agnostic fusion.

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 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 $\epsilon = 10^{-8}$, with a step-decay scheduler that reduces the learning rate by $10\times$ at epoch 30. Data augmentation includes fast-motion augmentation, per-frame image noise injection ($\sigma = 10$), and gamma jittering from 0.5 to 1.7. Training takes around 24.5 hours for AVI-HT-UME and 42 hours for AVI-HT-MANO.

B 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.

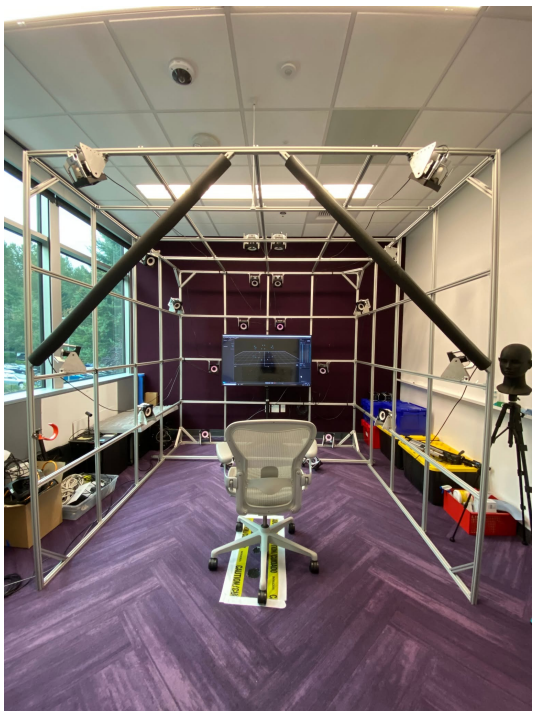

Figure 8: Marker Based Mocap System Setup

C 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. We also include a video in supplementary.zip that demonstrates AVI-HT's advantage in a sequential setting. As shown in the video, when the gloved hand becomes occluded, UMETrack's tracking error increases substantially, whereas AVI-HT largely maintains accurate pose estimates. In the right sub-plot, we visualize the 2nd vision-IMU attention weights: the importance assigned to

IMU signals rises markedly during occlusion and recedes as hand visibility is restored. The results confirm that the model learns to dynamically leverage inertial cues when visual evidence varies.

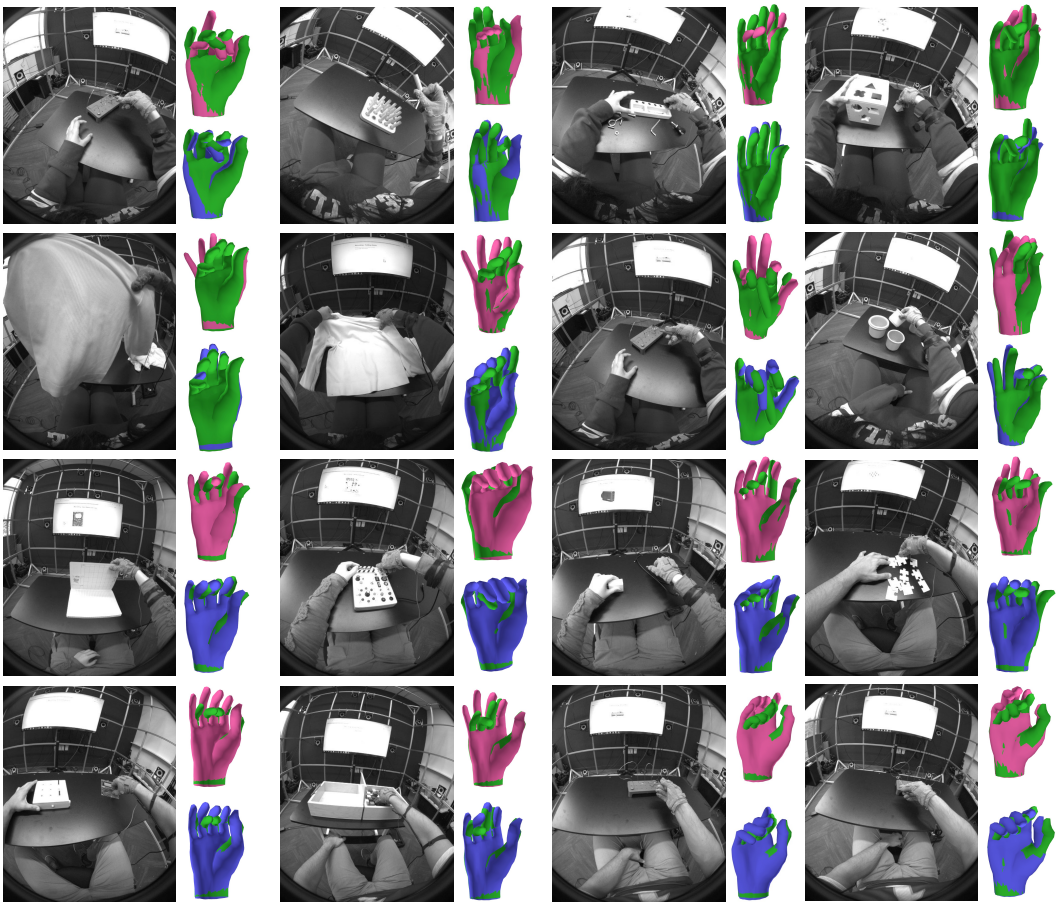

Figure 9: More qualitative comparison between AVI-HT and UMETrack.

D Sensitivity Study

To assess the robustness of AVI-HT to degraded IMU signals, we conduct two sensitivity analyses shown in Figure 10. 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.

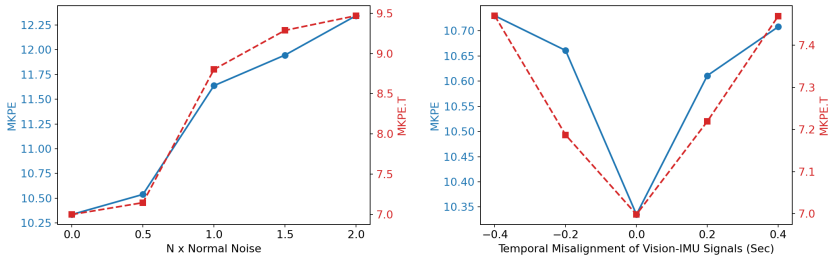

Figure 10: Sensitivity to IMU signal quality. (Left) MKPE and MKPE.T under increasing additive Gaussian noise. (Right) MKPE and MKPE.T under temporal misalignment between vision and IMU.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discuss the limitations of the work performed by the authors.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This paper does not include theoretical results; our contributions are empirical.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides complete architectural details including the IMU encoder, vision encoders, the per-joint cross-sensor attention mechanism with kinematic prior mask , and all training losses and objectives (Section 3.3). Input preprocessing, IMU temporal windowing, and synchronization procedures are described in Section 3.2. The DexGloveHOI benchmark, including capture setup, gesture protocol, and ground-truth generation, is fully specified in Section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Both the code and datasets are subject to institutional privacy and security review requirements. We are unable to provide open access at this time but are actively pursuing the necessary internal review processes to enable public release in the future. We provide detailed descriptions of our model architecture, training procedures, and evaluation protocol in the paper and supplementary material to support reproducibility.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We specified the details of the model architecture in Section 3.3 and training/testing implemenetation details in the second paragraph of Section 5..

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: we do not report statistical confidence intervals over multiple training runs since evaluation is deterministic on a fixed test set, the main source of variability would be training initialization, which we did not sweep.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: we provide sufficient information on the computer resources in the second paragraph of Section 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: there is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: the paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respectedthe creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: the paper does not release new assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: the paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: the core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.