

Beyond Dense Sensing: Biomechanically-Guided Hand Pose Reconstruction from Reduced Observations

Chao Yao*
University of Science and Technology
Beijing
Beijing, China
yaochao@ustb.edu.cn

Changsheng Li
University of Science and Technology
Beijing
Beijing, China
m202320825@xs.ustb.edu.cn

Yeting Song
University of Science and Technology
Beijing
School of Computer and
Communication Engineering
Beijing, China
m202421069@xs.ustb.edu.cn

Hongjun Liu
University of Science and Technology
Beijing
School of Intelligence Science and
Technology
Hai Dian, Beijing, China
ustb_lhj@163.com

Xiaojuan Ban*
University of Science and Technology
Beijing
Beijing, China
banxj@ustb.edu.cn

Abstract

Accurate full-hand dynamic capture in XR remains challenging because existing approaches largely depend on dense and layout-specific observations, including optical keypoints and high-density wearable sensing. Although such dense inputs can provide rich motion evidence under favorable conditions, they also introduce important practical limitations. Vision-based tracking becomes unreliable under self-occlusion, tool manipulation, and challenging illumination, while dense gloves increase hardware burden and often require reconstruction models tightly coupled to a specific sensing layout. As a result, existing hand-pose reconstruction pipelines remain strongly dependent on increasingly dense input observations. In this paper, we present PIRO-Pose, a biomechanically-guided sensing-model co-design framework for full-hand pose reconstruction from sparse wearable measurements. Instead of directly mapping sensor signals to poses, PIRO-Pose first transforms heterogeneous sensing inputs into a canonical anatomy-anchored sparse representation, where sensor reduction is guided by hand kinematic topology and inter-joint dependencies. Based on this structured representation, PIRO-Pose reconstructs dense hand articulation through anatomically structured graph expansion, multi-scale temporal modeling, and kinematic decoding with physiological constraints. Experiments show that PIRO-Pose achieves accurate real-time reconstruction on our in-house sparse-sensing dataset, remains competitive on a public high-density glove benchmark after compressing the original dense sensor streams into the same

reduced interface, and maintains robust continuous tracking under severe visual degradation. Together, these results show that reliable full-hand reconstruction can be achieved not only through dense sensing itself, but also through structured reduction and reconstruction on a normalized sparse observation space.

CCS Concepts

• **Human-centered computing** → **Interaction paradigms**; • **Computing methodologies** → *Artificial intelligence*.

Keywords

Reduced Observations; Biomechanically-Guided; Hand Pose Reconstruction

ACM Reference Format:

Chao Yao, Changsheng Li, Yeting Song, Hongjun Liu, and Xiaojuan Ban. 2026. Beyond Dense Sensing: Biomechanically-Guided Hand Pose Reconstruction from Reduced Observations. In *The 39th Annual ACM Symposium on User Interface Software and Technology (UIST '26)*, November 02–05, 2026, Detroit, MI, USA. ACM, New York, NY, USA, 15 pages. <https://doi.org/10.1145/3830398.3830490>

1 INTRODUCTION

Accurate and continuous full-hand dynamic capture is a fundamental enabler for immersive interaction in eXtended Reality (XR). Existing hand-tracking systems are dominated by two major paradigms: optical vision-based tracking and wearable sensing gloves. Vision-based approaches have achieved impressive performance under favorable line-of-sight conditions, while wearable systems can directly measure intrinsic hand articulation without relying on visibility. However, despite their different sensing mechanisms, both paradigms remain strongly dependent on dense observations.

For vision-based tracking, dense observations typically take the form of visible hand keypoints, meshes, or multi-view image evidence. Such approaches can degrade sharply once visual observability is compromised by self-occlusion, tool manipulation, poor

*Corresponding author.



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
UIST '26, Detroit, MI, USA

© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2856-3/26/11
<https://doi.org/10.1145/3830398.3830490>

illumination, or field-of-view truncation. For glove-based tracking, the dominant trend has been to increase sensing density so that individual joints or local deformation regions are measured more directly. While this strategy improves observability, it also increases bulk, wiring complexity, calibration burden, and hardware specificity. In practice, many reconstruction pipelines are therefore closely tied to a particular dense input configuration and transfer poorly across sensing layouts.

This dependence on dense, layout-specific observations motivates a different perspective: instead of continuously increasing sensing density, can we identify a structured intermediate representation that preserves essential hand articulation information while decoupling reconstruction from the original sensing configuration? Instead of treating the original sensing topology as the mandatory input space for pose estimation, we ask whether heterogeneous and potentially complex inputs can first be reduced into a shared sparse observation interface, and whether full-hand motion can then be reconstructed from that interface in a unified way. Under this view, sparse observation is not merely a hardware constraint; it becomes an intermediate formulation that decouples reconstruction from the original input layout.

To this end, we present PIRO-Pose, a biomechanically-guided sensing-model co-design framework for full-hand pose reconstruction from sparse wearable measurements. The framework contains two tightly connected stages. First, we introduce a Biomechanically-Guided observation reduction module that maps heterogeneous raw inputs into an anatomy-anchored 6-dimensional canonical representation. Here, “Biomechanically-Guided” refers to the fact that the reduction is guided by anatomical topology, joint articulation patterns, and inter-finger kinematic dependencies, rather than by unconstrained latent compression alone. Rather than serving as a fixed sensor-to-joint correspondence, this 5+1 architecture acts as a versatile kinematic bottleneck that absorbs heterogeneous sensing layouts into a shared latent space. It adapts to sensor disruptions and normalizes dense external hardware into our latent space, serving as a generalized intermediate framework rather than a hardcoded lookup table. Second, we reconstruct full-hand motion from this canonical sparse interface through graph-based spatial expansion, multi-scale temporal modeling, and kinematic decoding under physiological constraints.

In this sense, the reduced observation interface is not introduced as a replacement for all complex sensing, but as a normalized reconstruction space that reduces dependence on any single dense hardware layout. Rather than casting sparse-input hand-pose estimation as direct black-box regression, PIRO-Pose decomposes the problem into observation reduction and structured reconstruction. The reduction stage translates diverse sensing layouts into a consistent sparse representation, while the reconstruction stage expands this reduced interface into full-hand joint-level motion using anatomical structure, biomechanical coordination, and temporal consistency. This decomposition allows the same reconstruction pipeline to operate on both native sparse sensing and compressed dense observations.

We evaluate PIRO-Pose from three complementary perspectives. First, on our in-house dataset, we show that the proposed framework achieves accurate real-time hand-pose reconstruction from sparse wearable observations. Second, on a public high-density

glove benchmark, we show that the same framework remains competitive even after the original dense input streams are reduced to our canonical sparse interface. Third, under degraded visual conditions such as occlusion, dim lighting, and boundary-of-view motion, we demonstrate that the system maintains continuous and stable tracking when a near-field optical baseline becomes unreliable.

In summary, the main contributions of this work are three-fold:

- We introduce a Biomechanically-Guided observation reduction strategy that transforms heterogeneous hand-tracking inputs into a canonical sparse observation interface, enabling reconstruction beyond specific sensing layouts.
- We propose an anatomically structured reconstruction framework that expands the canonical sparse representation into full-hand motion through graph-based spatial reasoning, multi-scale temporal modeling, and kinematic decoding.
- We conduct extensive evaluations showing that the proposed formulation supports accurate real-time reconstruction on native sparse sensing, competitive transfer from compressed dense inputs, and robust continuous tracking under degraded visual conditions and generalization across heterogeneous sensing configurations.

2 RELATED WORK

2.1 Vision-Based Tracking and Its Limitations

Vision-based hand tracking remains the dominant non-contact paradigm in XR, benefiting from the ubiquity of commodity RGB cameras and headset-integrated optical sensors [24, 30, 52]. Over the past several years, substantial progress has been made in improving the robustness and fidelity of vision-based estimation. Representative directions include occlusion-aware mesh recovery, synthetic-to-real adaptation, and multi-view egocentric capture pipelines designed to improve spatial coverage and annotation quality [13, 14, 16, 17, 25–27, 50, 51]. In parallel, larger training corpora and stronger biomechanical priors have improved the plausibility and stability of reconstructed hand motion [1, 15, 31, 42, 48, 49].

Despite these advances, vision-based tracking remains fundamentally constrained by line-of-sight availability. In realistic HCI scenarios, severe self-occlusion, tool interaction, bimanual manipulation, degraded illumination, and field-of-view truncation can all lead to substantial temporal jitter or outright tracking loss. These limitations are especially problematic in continuous-feedback settings, where interaction reliability is often more critical than peak accuracy under ideal conditions. This motivates the use of wearable proprioceptive interfaces that can remain observable even when visual evidence becomes unreliable.

2.2 Wearable Proprioceptive Interfaces and the Density Trade-off

Wearable interfaces address optical occlusion by measuring intrinsic body motion directly. Prior work has explored a range of form factors, including wristbands, forearm-worn systems, and fully instrumented gloves [20, 23, 45]. These devices employ diverse sensing modalities such as IMUs, sEMG, and flexible strain sensors [3, 22, 39, 56]. Among them, glove-based systems are particularly

attractive for articulated hand tracking because they can capture finger-level deformation at the source.

A common design trend in this literature is to increase sensor density in order to improve observability of the hand’s highly coupled articulation. Many glove systems therefore deploy 15–25 or more sensing elements so that individual joints or local motion regions are measured as directly as possible [10, 21, 33, 35, 37, 44]. High-quality flexible sensors can indeed provide conformal contact and favorable signal characteristics such as low hysteresis and good repeatability [29, 36]. However, aggressively increasing sensing density also brings practical costs, including greater bulk, wiring complexity, calibration overhead, and wearing burden, all of which can reduce comfort and interfere with natural manipulation [8, 37].

A second challenge is algorithmic coupling to hardware topology. Models trained on dense wearable layouts are often strongly tied to the exact placement and semantics of the underlying sensors, which limits transfer across devices and typically necessitates layout-specific retraining [2, 28]. Although a few ultra-sparse systems have been proposed to reduce this burden [32], they mainly target discrete gesture recognition rather than continuous, high-fidelity 3D hand-pose reconstruction. Our work addresses this gap by asking whether structured reconstruction can compensate for aggressively reduced observations, thereby reducing dependence on dense and layout-specific sensing.

2.3 Structured Reconstruction from Reduced Observations

Beyond conventional cameras and dense sensing gloves, recent work has also investigated a broader family of less intrusive modalities for hand-pose estimation, including active acoustic sensing, thermal sensing, electrical impedance tomography, and multimodal wearable combinations built around watches, rings, and other lightweight form factors [4, 7, 9, 11, 12, 18, 19, 34, 40, 46, 47, 53–55]. These approaches broaden the design space for wearable interaction, but they also introduce their own trade-offs in terms of signal bandwidth, environmental sensitivity, and pose observability.

On the modeling side, hand-pose estimation has increasingly moved away from naive direct regression toward structured learning formulations that incorporate anatomical topology and biomechanical constraints. Representative examples include topology-aware Transformers, graph-based reasoning architectures, and models that explicitly regularize hand kinematics for improved physiological plausibility and temporal smoothness [5, 6, 41, 48]. However, most of these methods still assume access to high-dimensional and relatively information-rich observations.

Our work differs in problem formulation. Rather than improving hand-pose reconstruction by further increasing sensing density, we investigate whether a strongly reduced observation space can be made sufficient through structured spatiotemporal inference. Specifically, we combine a layout-normalized reduced representation with anatomically grounded graph expansion, multi-scale temporal modeling, and kinematic constraints. In this way, the proposed method treats sparse-input hand-pose estimation not as a direct black-box regression problem, but as a constrained reconstruction problem guided by hand biomechanics.

3 METHOD

3.1 Biomechanically-Guided Observation Reduction into a Canonical Sparse Interface

Existing hand-pose reconstruction pipelines are often tightly coupled to specific sensing layouts, such as dense strain arrays or optical multi-point tracking. To reduce this dependence, we first introduce a Biomechanically-Guided observation reduction module that maps heterogeneous raw observations into a shared 6-dimensional canonical representation anchored to hand structure. Here, “Biomechanically-Guided” does not refer to full rigid-body dynamics or contact modeling; instead, it means that the reduction is guided by the physical organization of the hand, including finger-wise articulation and transverse coupling around the MCP region, rather than by purely unconstrained latent compression.

For each time step t , the raw observation is projected into a standardized reduced interface, denoted by

$$\mathbf{r}_t = [r_{\text{thumb}}, r_{\text{index}}, r_{\text{middle}}, r_{\text{ring}}, r_{\text{little}}, c_t] \in \mathbb{R}^6.$$

In this 5+1 design, the first five dimensions summarize finger-wise longitudinal motion evidence, while the sixth dimension captures transverse coupling across the palm/MCP region. Rather than serving as a fully modality-agnostic latent code, this representation acts as an anatomy-anchored intermediate interface that normalizes diverse sensing layouts into a common sparse reconstruction space.

We implement this reduction as a dynamic, context-aware spatial routing operation. For a raw sensor input vector $\mathbf{x}_t \in \mathbb{R}^N$ ($N \geq 6$) at time step t , we map the inputs into Key and Value representations, $\mathbf{K}_t \in \mathbb{R}^{N \times d}$ and $\mathbf{V}_t \in \mathbb{R}^{N \times 1}$, respectively. The six target canonical nodes are parameterized by a fixed set of learnable queries, $\mathbf{Q} \in \mathbb{R}^{6 \times d}$.

Purely unconstrained attention-based routing may produce spatially implausible associations in this highly underdetermined reduction problem. To make the assignment more stable and physically interpretable, we introduce a lightweight Coarse Anatomical Prior (CAP) mask,

$$\mathbf{M} \in \{0, -\infty\}^{6 \times N},$$

which restricts routing to coarse, functionally valid hand regions. The soft assignment matrix $\mathbf{S}_t \in \mathbb{R}^{6 \times N}$ is then computed as

$$\mathbf{S}_t = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}_t^\top}{\sqrt{d}} + \mathbf{M}\right).$$

Here, $\mathbf{M}$ groups raw sensing channels into broad functional zones. If a raw sensor j is clearly mismatched with canonical node i (for example, a distal little-finger signal assigned to the index-finger node), we set $M_{i,j} = -\infty$ to suppress that route; otherwise, $M_{i,j} = 0$. In this way, the model still performs adaptive aggregation, but does so within a hand-structured routing space rather than over arbitrary cross-region assignments.

This design makes the reduction substantially less dependent on the original sensing topology. The model does not require the downstream reconstruction stage to preserve the exact number of sensors, their dense wiring pattern, or their original hardware-specific arrangement. Instead, it aggregates heterogeneous observations into a normalized sparse interface through constrained routing. The

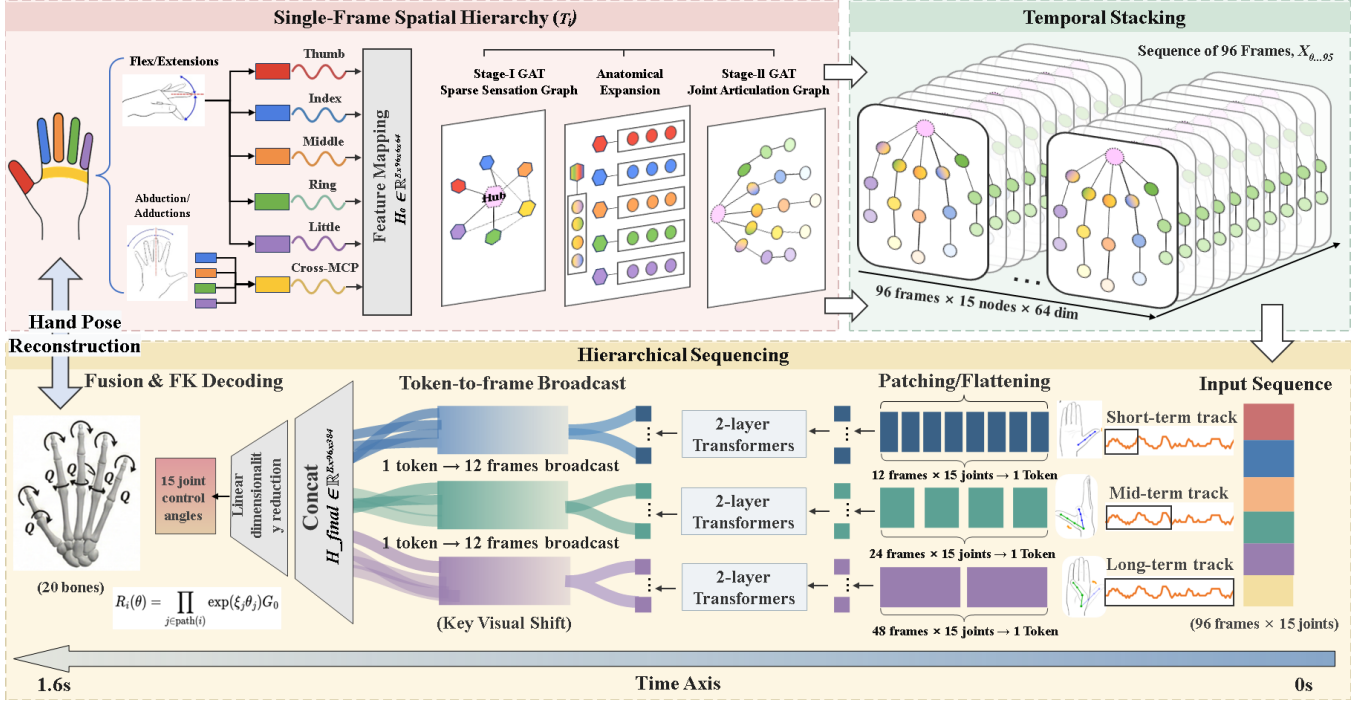


Figure 1: The proposed reduced-observation reconstruction framework. Top: The Single-Frame Spatial Hierarchy reduces high-dimensional inputs to a canonical 6-node sparse interface, which is then expanded to a 15-node joint articulation graph. Bottom: The Hierarchical Sequencing module stacks these representations temporally, applying multi-scale patching (short-, mid-, and long-term) and Transformers to ensure temporal consistency before Forward Kinematics (FK) decoding.

resulting canonical representation is obtained as

$$\mathbf{r}_t = \mathbf{S}_t \mathbf{V}_t,$$

which forwards the reduced 6-node features to the subsequent reconstruction modules.

To prevent the six canonical nodes from collapsing into redundant or overly overlapping responses, we further impose a structural orthogonality penalty during joint training:

$$\mathcal{L}_{\text{ortho}} = \lambda \|\mathbf{S}_t \mathbf{S}_t^\top - \mathbf{I}\|_F^2, \quad (1)$$

where $\mathbf{I}$ denotes the 6×6 identity matrix. This Frobenius-norm penalty encourages the aggregated signals to preserve distinct functional meanings, reducing excessive overlap between independent finger chains and the global MCP-coupling channel.

The reduced-observation sequence

$$\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_T]$$

then serves as the standardized input to the subsequent reconstruction stage, effectively decoupling raw sensing hardware from downstream pose estimation logic.

3.2 Structured Reconstruction from the Canonical Sparse Interface

Having established the canonical 6-dimensional sparse interface in Sec. 3.1, we now reconstruct these compact observations into a complete and biomechanically plausible 3D hand pose. Starting from the reduced-observation sequence $\mathbf{R}$, the reconstruction module

proceeds in four stages: part-level graph initialization, anatomically structured graph expansion, multi-scale temporal modeling, and kinematic decoding with physiological constraints. Importantly, this stage operates on the normalized sparse interface rather than on the original sensing layout, allowing the same reconstruction pipeline to be shared across native sparse inputs and compressed dense observations.

3.2.1 Bio-Synergistic Graph Initialization. At each time step, the reduced observation $\mathbf{r}_t$ is converted into a sparse perception graph $G_s = (V_s, E_s)$. The node set $V_s = \{v_1, v_2, v_3, v_4, v_5, v_6\}$ contains five longitudinal finger nodes and one transverse node. The first five nodes correspond to motion evidence from the thumb, index, middle, ring, and little fingers, respectively, while v_6 represents shared transverse evidence from the palm/MCP region.

The purpose of this graph is to inject coarse hand structure into the reduced observation before joint-level reconstruction. Specifically, the structural prior used here has two parts. First, the graph follows the part-level organization of the hand: each reduced dimension is assigned to a specific finger or to the shared MCP-coupling region, rather than being treated as an anonymous latent feature. Second, the graph encodes coarse coordination patterns in natural hand motion: neighboring fingers often exhibit correlated motion trends, while the transverse MCP region provides shared coupling information for the four non-thumb fingers.

Accordingly, neighboring finger nodes are connected to capture local inter-finger coordination, and the transverse node is connected

to the four non-thumb finger nodes to provide shared context for MCP-related coupling. Each scalar observation is first lifted into a node feature through

$$\mathbf{h}_i^{(0)} = \sigma(W_i r_{t,i} + b_i),$$

where W_i and b_i are learnable parameters.

A standard Graph Attention Network (GAT) [38] is then applied over this sparse graph to aggregate features only along these predefined edges. In this way, the first-stage graph encoder does not simply smooth features; instead, it produces a part-structured representation in which finger-wise evidence and transverse MCP coupling are already partially organized before expansion to the joint level. This coarse graph initialization also establishes the thumb/non-thumb asymmetry that is further developed in the next stage.

3.2.2 Anatomically Structured Graph Expansion. Although the sparse perception graph captures coarse coordination, full-hand reconstruction requires a joint-resolved representation. We therefore expand the six perception nodes into 15 joint-control nodes. The target joint-control space is denoted by $\hat{\Theta}_t \in \mathbb{R}^{15}$, comprising three thumb controls (TM, MCP, and IP) and the MCP/PIP/DIP controls for each of the other four fingers.

This expansion is structured rather than unconstrained. The joint organization of the hand determines how reduced nodes are unfolded: the thumb is expanded into its own three-joint chain, whereas each non-thumb finger is expanded into an MCP–PIP–DIP chain. At the same time, different joints are not inferred from the same source in the same way. For the thumb, the longitudinal thumb node is unfolded into three thumb-specific joint controls. For each non-thumb finger k , the MCP node is derived jointly from the finger-local feature and the transverse hub, whereas the PIP and DIP nodes are expanded from the same longitudinal finger feature:

$$\begin{aligned} [\mathbf{h}_{\text{TM}}, \mathbf{h}_{\text{MCP}}^{\text{th}}, \mathbf{h}_{\text{IP}}] &= \psi_{\text{th}}(\mathbf{h}_{\text{thumb}}^{(1)}), \\ \mathbf{h}_{\text{MCP}}^k &= \psi_{\text{MCP}}([\mathbf{h}_k^{(1)} \parallel \mathbf{h}_6^{(1)}]), \\ [\mathbf{h}_{\text{PIP}}^k, \mathbf{h}_{\text{DIP}}^k] &= \psi_{\text{chain}}(\mathbf{h}_k^{(1)}). \end{aligned} \quad (2)$$

This asymmetric expansion rule encodes the key coordination prior used in our model: MCP motion across the four non-thumb fingers is partly cooperative and is therefore informed by both finger-local and transverse MCP-region evidence, whereas PIP and DIP motion is more strongly governed by the local flexion chain of each finger. The expansion step therefore acts as a structured mapping from reduced observations to joint-level semantics, rather than as arbitrary dimensionality lifting.

After expansion, we build a joint graph over the 15 control nodes and apply a second graph-attention stage. The joint graph contains within-finger chain edges (e.g., MCP–PIP–DIP) and lateral MCP couplings across fingers. This stage refines joint-level dependencies and produces a joint-aware spatial representation for subsequent temporal modeling.

3.2.3 Multi-scale Temporal Modeling. Reduced observations are inherently noisy and may exhibit delayed or hysteretic responses under dynamic motion. Moreover, hand actions unfold across multiple temporal scales: short intervals capture rapid flexion changes

and local transients, whereas longer intervals encode motion preparation, transition dynamics, and broader contextual trends. To address this, we build a multi-scale temporal encoder on top of the joint-level spatial features.

Let the spatial module output be $\mathbf{H} \in \mathbb{R}^{T \times 15 \times d}$. We partition $\mathbf{H}$ into three temporal scales with patch lengths $p \in \{12, 24, 48\}$. For each temporal scale p , let $\mathcal{P}_p(\mathbf{H})$ denote the operator that partitions the feature sequence into non-overlapping temporal patches and flattens each patch into a token-level representation. Let $\mathbf{W}^{(p)}$ denote the learnable projection matrix associated with scale p . The scale-specific token embeddings are then written as

$$\mathbf{Z}^{(p)} = \text{PE}(\mathcal{P}_p(\mathbf{H}) \mathbf{W}^{(p)}), \quad (3)$$

where $\text{PE}(\cdot)$ denotes positional encoding.

Under a 96-frame input window, the three branches produce 8, 4, and 2 tokens, respectively. These branches correspond to short-term motion detail, mid-term motion-unit dynamics, and long-term temporal context. Each token sequence is then processed by an independent Transformer encoder, yielding $\hat{\mathbf{Z}}^{(p)} = \text{Transformer}_p(\mathbf{Z}^{(p)})$. After encoding, the patch-level tokens are lifted back to frame-level representations and fused across scales. Let $\mathcal{U}_p(\cdot)$ denote the scale-specific token-to-frame lifting operator. The fused representation is given by

$$\mathbf{F} = [\mathcal{U}_{12}(\hat{\mathbf{Z}}^{(12)}) \parallel \mathcal{U}_{24}(\hat{\mathbf{Z}}^{(24)}) \parallel \mathcal{U}_{48}(\hat{\mathbf{Z}}^{(48)})] \mathbf{W}_f, \quad (4)$$

where $\mathbf{W}_f$ is a learnable fusion projection.

The short scale preserves rapid local motion cues, such as quick finger flexion and subtle transient changes. The middle scale captures more complete motion units, while the long scale models broader temporal context and transition trends. In this sense, the temporal module acts not only as a feature extractor, but also as a learned consistency filter that improves stability under delayed or noisy reduced observations.

3.2.4 Kinematic Decoding and Physical Constraints. From the fused frame-wise feature $\mathbf{F}_t$, a prediction head outputs the 15 joint controls $\hat{\Theta}_t$. These controls are then decoded by a forward kinematics (FK) module into 20 skeletal bone rotations represented as quaternions, namely $\hat{\mathbf{Q}}_t \in \mathbb{R}^{20 \times 4}$. By predicting joint controls before skeletal rotations, the model avoids directly regressing a high-dimensional pose from the reduced observation interface.

Training is driven by both reconstruction and physical-consistency objectives. The primary supervision is a quaternion pose loss L_{quat} . To improve physiological plausibility, we further introduce a range-of-motion constraint L_{rom} and a soft PIP–DIP coupling constraint L_{couple} . To improve dynamic stability, we add temporal regularization terms on velocity and acceleration, denoted by L_{vel} and L_{acc} , respectively. Finally, a quaternion normalization term L_{norm} encourages valid rotations. The overall training objective is

$$\mathcal{L} = \lambda_1 L_{\text{quat}} + \lambda_2 L_{\text{rom}} + \lambda_3 L_{\text{couple}} + \lambda_4 L_{\text{vel}} + \lambda_5 L_{\text{acc}} + \lambda_6 L_{\text{norm}}. \quad (5)$$

Together, these constraints transform the ill-posed mapping from reduced observations to full-hand motion into a structured optimization problem governed by joint organization, cross-joint coordination, and temporal continuity. As a result, PIRO-Pose reconstructs

temporally stable and physiologically plausible hand motion from the canonical sparse observation interface.

4 DATASETS AND PREPROCESSING

We evaluate the proposed framework on two complementary datasets: an in-house sparse-sensing dataset collected under controlled, occlusion-free conditions to assess reconstruction performance under our native sensing setup, and a public high-density wearable dataset to examine whether the same reduced-observation formulation remains effective across substantially different input layouts.

4.1 Our Sparse Sensing Dataset

Hardware Setup. We constructed a sparse stretch-sensing prototype using six flexible strain sensors. Following the sensing design described in Sec. 3, five sensors were distributed longitudinally along the fingers, and one hub sensor was arranged transversely across the MCP region. To obtain continuous optical reference trajectories, we used an Ultraleap Leap Motion Controller 2 (LMC2) mounted 20 cm above the workspace to synchronously record 3D hand poses under controlled line-of-sight conditions. Under such strict, occlusion-free conditions (C_0), prior validation by Weichert et al.[43] bounds the LMC spatial precision at ≤ 1.2 mm. Via kinematic error propagation on a typical 40 mm phalanx, this uncertainty bounds the visual reference’s angular error at approximately 1.72° . Thus, the LMC2 acts as a highly reliable Teacher providing clean kinematic priors for training. As illustrated in Fig. 2, this setup minimizes external mechanical disturbance during capture and allows the sensor response to be driven primarily by hand articulation.

Dataset Overview. We recruited 12 participants to perform a diverse set of hand motions. Each participant completed 20 action categories, and each category was recorded continuously for 3 minutes at 60 Hz, yielding approximately 2.59 million paired frames in total. The action set was designed to cover a broad motion range, including isolated single-finger degrees of freedom (DoF), foundational motion patterns (e.g., flexion/extension and adduction/abduction), and daily functional interactions. In addition, continuous random-motion sequences were recorded to capture transitional states between more canonical poses. Representative poses and interaction segments are shown in Fig. 2.

Evaluation Protocol. To evaluate generalization to unseen users, we adopted a Leave-One-Subject-Out (LOSO) cross-validation protocol. In each fold, data from 11 participants were used for training and validation, while the remaining participant was held out exclusively for testing. Within the 11-subject pool, trials were further split at a 7:3 ratio to form the training and validation sets for model selection and early stopping. [If true: The split was performed at the trial level before window extraction to avoid leakage between adjacent temporal segments.]

4.2 Data Alignment and Preprocessing

Because the 6-channel strain signals and the optical reference stream were not perfectly synchronized, we applied an offline multi-stage alignment procedure. First, cross-correlation analysis was used to estimate the global time offset between the two streams. We then applied Dynamic Time Warping (DTW) for finer non-linear

alignment of paired recordings. The aligned sequences were subsequently resampled to a uniform 60 Hz grid using cubic spline interpolation. Finally, individualized normalization was applied to the raw sensor values using extreme flexion and extension calibrations for each participant, reducing inter-user variation in signal scale.

4.3 Public High-Density Benchmark

To evaluate whether the proposed reduced-observation formulation transfers beyond our native sparse prototype, we selected the publicly available *SmartTextileGlove* dataset [37] for cross-dataset validation. Unlike our 6-channel prototype, this benchmark was collected using a substantially denser wearable setup with 25 stretchable textile sensors and more than 3 million motion frames. We use this benchmark as an external testbed for a denser and structurally different sensing topology. By projecting its high-dimensional inputs into the same canonical 6-node sparse interface, we examine whether the proposed framework remains effective without redesigning the downstream reconstruction architecture for the original hardware layout.

5 EVALUATION

We evaluate PIRO-Pose along four axes: runtime performance, reconstruction accuracy on our in-house dataset, cross-dataset transfer under reduced observations, and robustness under degraded visual conditions. We further conduct ablation experiments to isolate the contribution of the major architectural components. On both the in-house dataset and the public *Smart Textile Glove* benchmark, we report joint-angle RMSE (degrees) as the primary reconstruction metric; on the in-house dataset, we additionally report skeletal position error (mm). For the public benchmark, we follow the original paper and adopt the inter-participant protocol. Together, these experiments assess not only reconstruction accuracy, but also transferability, practical deployability, and robustness under conditions where vision-based tracking becomes unreliable.

5.1 Runtime Performance

A practical hand-pose reconstruction system must satisfy real-time constraints in addition to achieving good reconstruction quality. We therefore first report the runtime performance of PIRO-Pose. In terms of deployability, PIRO-Pose is highly lightweight, requiring only 0.84M parameters and 12.5 MB of memory with a per-frame latency of 3.2 ms. On a desktop RTX 3090 GPU, the complete pipeline runs at **115 FPS**. Remarkably, even on a CPU-only laptop (Intel Core i7-12700H, no dedicated GPU), it maintains **65 FPS**. This level of efficiency meets or approaches the 90 Hz refresh-rate budgets of typical untethered XR headsets. Although this is not the fastest configuration among the compared baselines, it indicates that the proposed structured reconstruction strategy remains suitable for real-time deployment. This level of efficiency also supports downstream continuous-feedback applications such as manipulation guidance and skill coaching.

Runtime alone, however, does not indicate whether the reconstructed poses are sufficiently accurate for fine-grained interaction. We therefore next evaluate reconstruction fidelity on our in-house

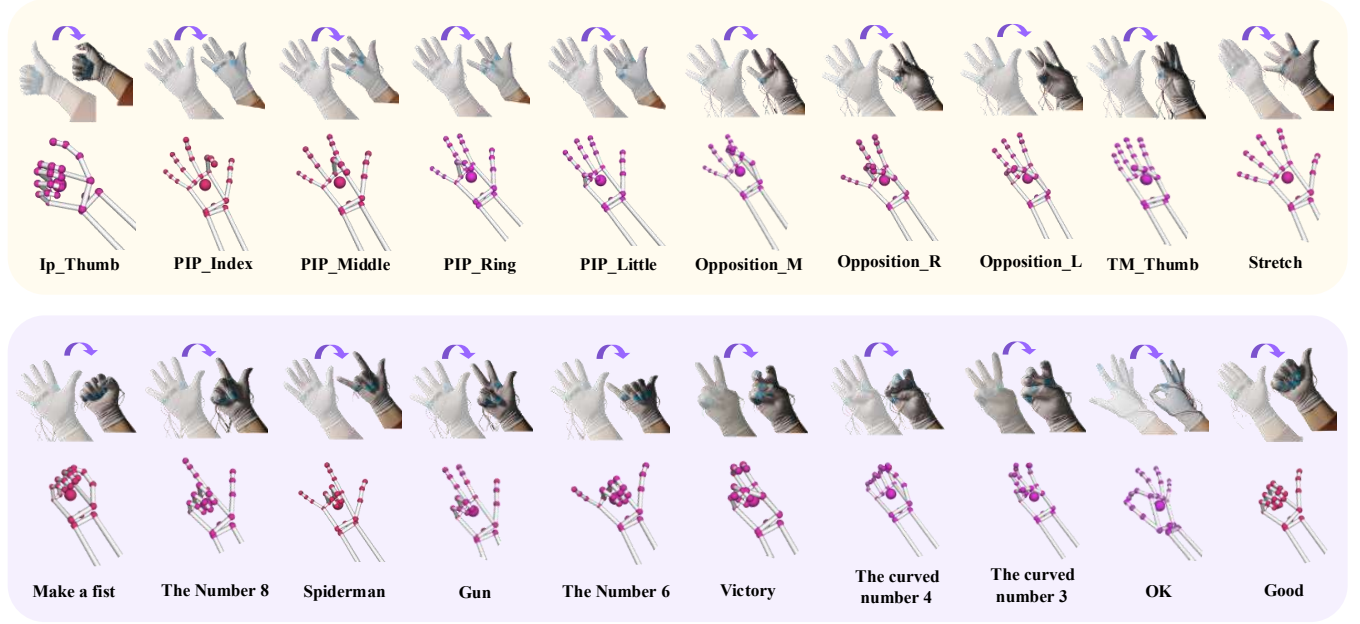


Figure 2: Our in-house data acquisition setup and dataset overview. The yellow examples correspond to isolated single-finger degrees of freedom (DoF) and foundational motion patterns, such as flexion/extension and adduction/abduction, while the purple examples correspond to daily functional interactions.

dataset under controlled comparison with standard learning baselines.

5.2 Intra-Dataset Reconstruction Accuracy

We first compare PIRO-Pose against several representative baselines on the same in-house dataset in order to verify whether the proposed structured reconstruction framework is more effective than direct regression from reduced observations. The baselines include: (1) a frame-wise MLP, which processes each time step independently; (2) an LSTM, which models temporal recurrence over flattened inputs; (3) modern temporal architectures, specifically a Temporal Convolutional Network (TCN) and a pure Transformer; and (4) a Direct Mapping baseline, which regresses the high-dimensional hand pose directly from reduced observations without explicit anatomical factorization. To ensure a fair comparison, all baseline models utilized identical train/test splits, preprocessing, and normalization protocols. We additionally performed an exhaustive hyperparameter grid search for all models (the optimal configuration for our model yielded: learning rate $5e-4$, weight decay $1e-4$, batch size 256, 128 hidden dimensions, and 8 attention heads). The complete model is denoted as **PIRO-Pose**.

As shown in Table 1, PIRO-Pose achieves a joint-angle RMSE of **0.984°** and a skeletal position error of **3.024 mm** . When combined orthogonally with the LMC2’s baseline uncertainty (1.72°), this yields a theoretical anatomical error bound of $\sim 1.98^\circ$. Therefore, the reported RMSE should be strictly interpreted as relative agreement with the LMC2 reference rather than absolute anatomical accuracy. Nevertheless, these results indicate accurate and geometrically stable hand-pose reconstruction under reduced observations

Table 1: Intra-dataset comparison on the in-house hand-pose reconstruction benchmark. Lower is better for joint-angle RMSE and position error; higher is better for FPS.

Architecture	Joint RMSE ($^\circ$)↓	Position Error (mm)↓	FPS
MLP	2.452	7.815	185
LSTM	1.837	5.642	120
TCN	1.650	5.120	149
Transformer	1.424	4.550	87
Direct Mapping	1.415	4.380	155
PIRO-Pose (Ours)	0.984	3.024	115

relative to the visual teacher. These results should be interpreted together with our native sensing topology, which was co-designed with the proposed reconstruction pipeline. Such hardware–algorithm co-design improves the observability of the reduced representation and likely contributes to the strong performance observed in this setting.

The relative behavior of the baselines remains informative. The frame-wise MLP captures only shallow nonlinear correspondences and cannot model either cross-finger structure or temporal dependency. Because the MLP relies on dense feature entanglement, it often suffers from cross-talk (e.g., thumb motions inappropriately distorting the pinky). The LSTM improves temporal continuity, but still operates on flattened inputs and does not explicitly encode anatomical topology or joint-chain relations. Conversely, the Direct Mapping baseline forces a hardcoded 1-to-1 routing (e.g., sensor 1 strictly to the thumb, sensor 6 to the MCP). While this prevents cross-talk, it completely ignores natural biomechanical

synergies, making it brittle under motion patterns involving thumb opposition, coupled MCP motion, and multi-finger coordination. Finally, even modern sequence models like the TCN (1.650°) and pure Transformer (1.424°) struggle. Although they possess strong temporal capacity, attempting sparse-to-dense mapping without explicit spatial priors leads to suboptimal reconstruction.

By contrast, PIRO-Pose better matches the structure of the task. Reduced observations are first lifted into a sparse perception graph and then expanded into 15 joint-control nodes through an anatomically grounded graph-expansion mechanism. Temporal modeling is subsequently performed using multi-scale Transformer branches, which capture both short-range motion detail and longer-range temporal intent while compensating for hysteresis and transient noise. Together with physical constraints, this turns sparse-to-high-dimensional pose recovery into a structured reconstruction problem rather than a purely direct regression problem.

From the efficiency perspective, PIRO-Pose runs at 115 FPS. Notably, although our architecture incorporates a temporal Transformer, it runs substantially faster than the standard pure Transformer baseline (87 FPS). This efficiency stems directly from our multi-scale temporal patching (3.2.3), which aggressively compresses the raw input sequence into a highly compact set of tokenized motion units. Although the full model is slower than the lightest baselines, the runtime remains firmly within the real-time regime and is accompanied by a substantial gain in reconstruction quality. Overall, these results indicate that PIRO-Pose achieves a favorable trade-off between accuracy and deployability under its native sensing setup.

The in-house study above establishes effectiveness under our co-designed acquisition topology. A more stringent question is whether the proposed reduced-observation formulation remains useful when transferred to a substantially different sensing layout. We examine this next on a public dense-glove benchmark.

5.3 Cross-Dataset Generalization

A major limitation of existing hand-pose reconstruction methods is that they are often tightly coupled to a particular sensing layout. To evaluate whether the reduced-observation formulation introduced in Sec. 3.1 supports transfer across substantially different hardware configurations, we test PIRO-Pose on the public *Smart Textile Glove* dataset released by Tashakori et al. [37]. This benchmark was collected using a substantially denser sensing setup than ours—25 helical sensor yarns (HSYs) and two 9-axis IMUs sampled at 20 Hz—with motion-capture supervision. Its original estimator, *GlovePoseML*, is a task-specific recurrent model composed of a two-layer bidirectional LSTM followed by two fully connected layers. The original paper reports an inter-participant joint-angle RMSE of 1.45° .

In our evaluation, we do not redesign the PIRO-Pose backbone for this dataset. Instead, we first map the dense textile-glove signals into the same anatomy-anchored 5+1 reduced-observation interface used throughout this work, namely five finger-wise longitudinal motion factors together with one transverse MCP-synergy factor. Full-hand reconstruction is then performed using the same Biomechanically-Guided spatiotemporal reconstruction pipeline

as in our native setting. To rigorously test architectural transferability without carrying over any dataset-specific calibrations, our model is retrained from scratch on this external benchmark. This experiment is intended as a targeted transfer test rather than an exhaustive benchmark sweep: the central question is whether the proposed reduced representation remains effective when moved to a dense and heterogeneous sensing platform.

As shown in Table 2, PIRO-Pose achieves an inter-participant joint-angle RMSE of 1.76° on this external benchmark. Although this is slightly higher than the 1.45° reported by the original dense-input model, it remains competitive given the substantially stronger information bottleneck imposed at the input stage. Importantly, our method neither preserves the original dense sensing topology nor relies on the benchmark’s task-specific recurrent architecture. Instead, it reconstructs full-hand motion from a topology-aware reduced representation while keeping the same structured reconstruction pipeline used in our own system.

Beyond architectural transferability across distinct hardware layouts, we also verified robust generalization across unseen users. On our native 6-sensor dataset, the standard deviation across all 12 Leave-One-Subject-Out (LOSO) cross-validation folds remained remarkably stable at $\pm 0.250^\circ$. This low variance confirms that the model generalizes reliably to new individuals without succumbing to user-specific overfitting.

Figure 3 provides a qualitative comparison on the same benchmark. Even after compressing the original dense textile-glove signals into the proposed reduced-observation interface, PIRO-Pose still produces stable and visually plausible joint regression results.

The significance of this result lies not in setting a new benchmark best, but in demonstrating that the proposed formulation is not tied to our native hardware design alone. As explicitly acknowledged earlier, the strong in-house results benefit from hardware-algorithm co-design, yet the external benchmark shows that the same reduced-observation representation can also normalize heterogeneous inputs to support accurate reconstruction on a very different sensing platform. This suggests that the central contribution of PIRO-Pose lies in the structured sparse-to-pose formulation rather than in a single device-specific topology.

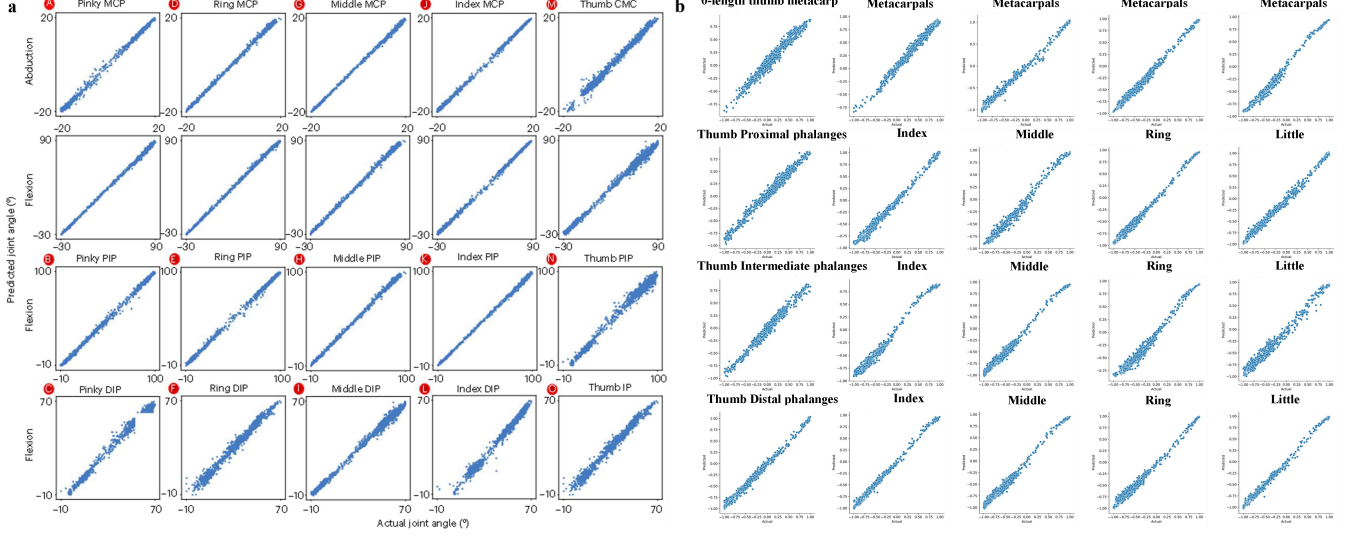
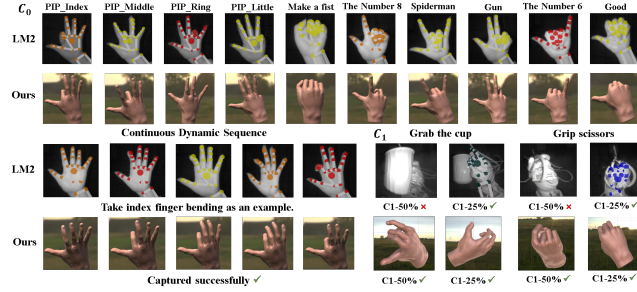
Cross-dataset transfer addresses one practical challenge, namely dependence on sensing layout. Another challenge concerns deployment under conditions where visual evidence becomes unreliable. We therefore next evaluate robustness under degraded visual conditions.

5.4 Robustness Under Degraded Visual Conditions

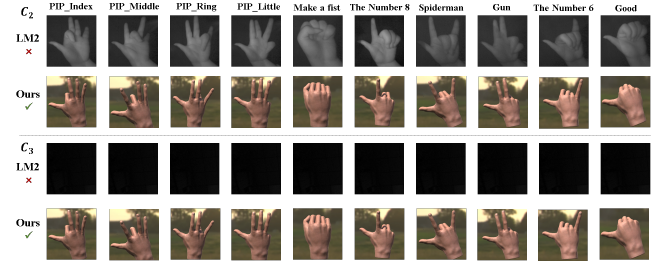
We next evaluate robustness under four controlled visual conditions (C0–C3). To avoid any perceived “modality tautology” when comparing a wearable against a vision-based system, we frame this evaluation through a Teacher-Student paradigm. Under the ideal C0 condition (as evaluated in Sec. 5.2), the optical tracker acts as a high-fidelity Teacher providing clean kinematic priors to train our wearable Student. However, the purpose of the C1–C3 experiments is not to claim higher absolute accuracy over the Teacher, but to

Table 2: Cross-dataset transfer on the public Smart Textile Glove benchmark under the inter-participant protocol. Lower is better.

Method	Input Representation	Backbone	Protocol	Joint-angle RMSE ($^{\circ}$)↓
GlovePoseML [37]	25 HSYs + 2×9-axis IMUs	2-layer Bi-LSTM + 2 FC	Inter-participant	1.45
PIRO-Pose (Ours)	Reduced observations	Graph + Multi-scale Transformer	Inter-participant	1.76

**Figure 3: Qualitative cross-dataset transfer on the public Smart Textile Glove benchmark. (a) Example regression result from the original dense-input benchmark of Tashakori et al. [37]. (b) Result obtained by PIRO-Pose on the same dataset after compressing the dense sensor streams into the proposed reduced-observation representation. Although our method no longer uses the original dense sensing topology, it still produces stable and plausible joint regression, indicating that the proposed representation and reconstruction strategy are not tied to a specific hardware layout.****Figure 4: C0 Normal and C1 Occlusion. Under C0, both systems maintain stable tracking. Under C1 (e.g., holding a mug or scissors), LMC2 exhibits severe skeleton dropouts at 50% coverage, whereas our system preserves a continuous pose stream.**

test whether tracking remains *usable* when the Teacher’s visual observability collapses. As a representative near-field optical baseline, we use the Ultraleap Leap Motion Controller 2 (LMC2).

**Figure 5: C2 Challenging Illumination and C3 FoV Boundary. Under C2 (dim/backlit) and C3 (edge/partial-out), LMC2 frequently loses tracking, whereas our system remains comparatively stable because it relies on intrinsic skin deformation rather than line-of-sight visibility.**

Experimental Protocol and Metrics. We recruited 12 participants to perform both static holds and dynamic object manipulations across the four conditions shown in Fig. 6(a). Because the optical baseline itself becomes unreliable under severe degradation, absolute pose error with respect to LMC2 is no longer a reliable

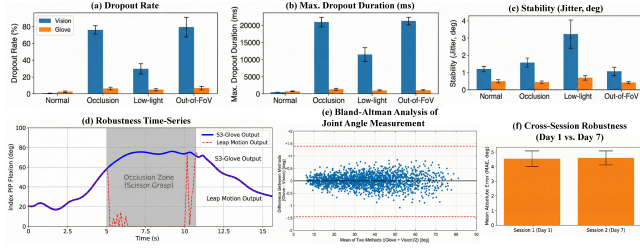


Figure 6: Robustness benchmark setup and results. (a) Experimental setup with four conditions (C0–C3). (b) Dropout rate. (c) Longest uninterrupted dropout duration. (d) Static-hold jitter. (e) Continuous time-series comparison in C1. (f) Cross-session robustness showing re-donning penalties.

or fair evaluation target in these conditions. Accordingly, this robustness benchmark focuses on availability- and stability-oriented metrics that remain well-defined even when optical tracking fails: dropout rate, longest uninterrupted dropout duration, and static jitter. In other words, these metrics strictly measure tracking *continuity* rather than accuracy, directly capturing whether a tracker can continue to provide usable and temporally stable feedback during manipulation, which is the central question of this experiment.

Results. As shown in Figs. 4–6, both systems behave similarly under normal conditions, with C0 dropout remaining below 0.8%. However, their behavior diverges sharply once visual conditions degrade. Under C1 (occlusion), LMC2 suffers frequent tracking failures, with a median dropout rate of **48.5%** (IQR: [15.2%, 82.4%]), whereas our system remains essentially unaffected with a median dropout of **0.0%**. Relying on skin stretch, our wearable deploys undisturbed, retaining the kinematics learned from the C0 Teacher. Repeated-measures analysis further reveals a significant System $\times$ Condition interaction for dropout rate ($p < .001$).

A similar pattern is observed under challenging illumination (C2) and field-of-view boundary conditions (C3). In both cases, LMC2 exhibits substantially larger static jitter and longer uninterrupted dropout durations (Wilcoxon, $p < .05$), whereas our system continues to provide continuous, low-jitter motion estimates. For interactive applications, these failures are particularly consequential: long dropout intervals interrupt feedback and break action continuity even if average reconstruction quality under ideal conditions is high.

Overall, these results show that intrinsic deformation sensing provides substantially higher practical reliability than pure vision-based tracking when line-of-sight is compromised. Importantly, the claim of this section is not that PIRO-Pose achieves lower *absolute* pose error under conditions where the optical reference itself fails; rather, it is that PIRO-Pose better preserves tracking availability and temporal stability under degraded observability. This complements the accuracy-oriented evaluations above by demonstrating that the system remains usable in realistic tool-based and mobile XR scenarios where visibility cannot be guaranteed.

Having established practical robustness under degraded visual conditions, we next isolate the contributions of the major components in PIRO-Pose through controlled ablations.

Table 3: Ablation study of the key components and biomechanical priors in PIRO-Pose.

Variant	Joint RMSE ($^{\circ}$) $\downarrow$	Δ	Position Error (mm) $\downarrow$	Δ
w/o Graph	1.348	+0.364	4.156	+1.132
w/o Temporal	1.172	+0.188	3.590	+0.566
w/o CAP	1.159	+0.175	3.520	+0.496
w/o PIP-DIP	1.106	+0.122	3.385	+0.361
w/o ROM	1.080	+0.096	3.290	+0.266
PIRO-Pose (Full)	0.984	—	3.024	—

Table 4: Topological robustness under sensor masking and hardware rearrangement (measured in Joint RMSE $^{\circ}$).

Condition	PIRO-Pose (Ours)	MLP Baseline
Full Sensing (Normal)	0.984	2.452
Masked Index	1.454	4.809
Masked Middle	1.402	4.751
Masked Index & Middle	2.805	>8.516
Masked Thumb	3.206	6.103
Thumb/Pinky Swapped (5-epoch FT)	1.117	>12.176

5.5 Ablation Study

To quantify the contribution of PIRO-Pose’s design choices, we ablate major architectural components and biomechanical priors (Table 3). We define two structural variants: w/o Graph (removes spatial expansion, retaining temporal modeling) and w/o Temporal (keeps graph expansion, removing the temporal Transformer). To isolate physiological priors, we systematically remove the CAP Mask (w/o CAP), PIP-DIP coupling (w/o PIP-DIP), and Range-of-Motion constraints (w/o ROM).

Table 3 confirms the necessity of both architectural components. The graph-expansion module (w/o Graph) yields the largest error increase (+0.364 $^{\circ}$). By structuring the expansion into 15 joint-control nodes, this module injects anatomical topology into the inverse problem. Without it, the network degrades toward unconstrained regression, struggling to resolve coupled motion patterns like MCP coordination and thumb opposition. Similarly, removing the multi-scale temporal encoder (w/o Temporal) destabilizes dynamic reconstruction. While the spatial graph ensures static anatomical plausibility, the temporal module is indispensable for modeling long-range dependencies and compensating for sensor hysteresis (+0.188 $^{\circ}$ penalty when removed).

Crucially, ablating specific physiological priors confirms their individual necessity: omitting the CAP Mask (w/o CAP) permits unphysical feature entanglement across independent functional zones; removing PIP-DIP coupling induces noticeable distal jitter; and dropping ROM constraints allows biomechanically impossible hyper-extensions during complex articulations.

Furthermore, to verify that PIRO-Pose truly internalizes these kinematic synergies rather than merely memorizing a hardcoded sensor-to-joint lookup table, we additionally evaluate its topological adaptability. Specifically, we test the framework’s robustness under sensor masking and hardware rearrangement against the MLP baseline (Table 4).

Under single-sensor dropouts, PIRO-Pose exhibits strong fault tolerance (e.g., 1.454° vs. 4.809° for Masked Index). Even under a severe double-dropout scenario (masking both Index and Middle), our model degrades gracefully (2.805°) by leveraging spatial compensation from the transverse MCP hub and adjacent finger synergies, whereas the MLP collapses entirely ($> 8.516^\circ$).

Notably, masking the Thumb yields a higher penalty (3.206°). Because the thumb possesses high degrees of freedom and lacks strong inter-finger synergies, our network correctly struggles to reconstruct it without valid evidence. This demonstrates that the model relies on learned physical priors rather than unconstrained hallucination. Finally, when simulating a hardware rearrangement by swapping the Thumb and Pinky sensor streams, PIRO-Pose rapidly restores accuracy after a brief 5-epoch fine-tuning (1.117°), while the MLP fails to adapt ($> 12.176^\circ$). These results confirm that our 5+1 architecture functions as an adaptable kinematic bottleneck rather than a fixed mapping.

6 Application & User Evaluation

To evaluate the practical HCI value of PIRO-Pose beyond reconstruction accuracy, we instantiated the system in a VR fine-motor coaching scenario centered on continuous bimanual paper-cutting. This task stresses the exact failure modes of a vision-based hand-tracking baseline—persistent self-occlusion, tool interference, and rapid grasp transitions—while requiring temporally stable pose feedback. We therefore conducted a within-subjects user study comparing PIRO-Pose against a state-of-the-art vision baseline. The full application pipeline, interface design, task protocol, measures, and statistical analysis are reported in Appendix A. In brief, PIRO-Pose yielded significantly higher joint-level kinematic similarity to expert reference trajectories and significantly lower NASA-TLX workload, while participants reported that the uninterrupted, occlusion-robust feedback provided a clearer sense of spatial progress during manipulation.

7 LIMITATIONS AND FUTURE WORK

While PIRO-Pose demonstrates robust continuity under severe occlusion and promising application-level utility, several limitations remain.

Sensor hysteresis and drift. The resistive strain sensors still exhibit hysteresis and slow baseline drift during prolonged wear. We mitigate these effects through brief per-session calibration and focus our evaluation on short-bout interactions and cross-session re-donning robustness. Future work should investigate online drift compensation and lower-hysteresis sensing materials.

Reference fidelity. Our training targets and reference trajectories are derived from LMC2 under nominal visual conditions. Accordingly, the reported errors under C_0 (e.g., 0.984° RMSE) should be interpreted as relative agreement with a commercial optical tracker rather than absolute anatomical ground truth. Future work should validate the system against higher-fidelity motion-capture references to isolate absolute anatomical error.

Dependence on coordination priors. PIRO-Pose reconstructs full-hand motion from reduced observations by leveraging structured assumptions about typical hand coordination. While these priors

regularize the inverse mapping, performance may degrade for atypical hand mechanics or passive deformations that deviate from common coordination patterns. Future work could incorporate adaptive personalization and user-specific kinematic constraints.

System scope. The current prototype focuses on intrinsic joint articulation and does not yet provide full global root translation or active haptic feedback. Future systems could combine the reduced-observation glove with external tracking for global motion capture and integrate lightweight haptic actuation for force-aware guidance.

8 CONCLUSION

In this paper, we challenge the prevailing trend of relying on either fragile optical tracking or increasingly dense sensing arrays for continuous full-hand dynamic capture in XR. We present PIRO-Pose, a biomechanically-guided sensing-model co-design framework built around a canonical sparse observation interface. Rather than binding reconstruction to a specific dense hardware layout, our approach first projects heterogeneous sensory inputs into a shared, anatomy-anchored canonical representation. Full-hand motion is then reconstructed from this reduced interface through graph-based joint expansion, multi-scale temporal modeling, and kinematic decoding with physiological constraints.

Extensive evaluations demonstrate the effectiveness and transferability of our framework. On our native dataset, the system achieves a mean joint-angle relative agreement of 0.984° with the optical reference and a positional error of 3.024 mm, while running in real time at 115 FPS. More importantly, we demonstrate cross-layout transfer by applying the same reduced-observation formulation to a public high-density textile glove dataset, where it remains highly competitive with existing hardware-specific models despite the stronger input bottleneck.

Furthermore, controlled ablation and sensor-masking experiments confirm that our framework effectively internalizes natural kinematic synergies rather than over-fitting to fixed hardware mappings, degrading gracefully even under severe sensor dropouts. Finally, our robustness study highlights the practical advantage of the proposed wearable approach in degraded scenarios: under severe tool occlusion, where a state-of-the-art near-field optical tracker suffered a 48.5% dropout rate, our system maintained uninterrupted tracking availability with a 0.0% dropout rate.

Overall, this work shows that robust and high-fidelity hand tracking does not require reconstruction pipelines to remain tied to increasingly dense and layout-specific observations. Instead, reducing heterogeneous inputs into a normalized sparse observation space and reconstructing full-hand motion through structured spatial, temporal, and kinematic modeling provides a more practical and deployable path toward next-generation wearable XR interfaces.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (Grant No. 2022ZD0118001) and the National Natural Science Foundation of China (Grant No. 62332017).

References

- [1] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. 2025. 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 (CVPR)*.
- [2] M. Chen, J. Ouyang, A. Jian, et al. 2022. Imperceptible, designable, and scalable braided electronic cord. *Nature Communications* 13, 1 (2022), 7097.
- [3] W. Chen, C. Yu, C. Tu, et al. 2020. A survey on hand pose estimation with wearable sensors and computer-vision-based methods. *Sensors* 20, 4 (2020), 1074.
- [4] Weiya Chen, Chenchen Yu, Chenyu Tu, Zehua Lyu, Jing Tang, Shiqi Ou, Yan Fu, and Zhidong Xue. 2020. A Survey on Hand Pose Estimation with Wearable Sensors and Computer-Vision-Based Methods. *Sensors* 20, 4 (2020), 1074. doi:10.3390/s20041074
- [5] Wencan Cheng and Jong Hwan Ko. 2023. HandR2N2: Iterative 3D Hand Pose Estimation Using a Residual Recurrent Neural Network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 20847–20856. doi:10.1109/ICCV51070.2023.01911
- [6] Lin Huang, Jianchao Tan, Ji Liu, and Junsong Yuan. 2020. Hand-Transformer: Non-Autoregressive Structured Modeling for 3D Hand Pose Estimation. In *Computer Vision – ECCV 2020*. doi:10.1007/978-3-030-58595-2_2
- [7] Daehwa Kim and Chris Harrison. 2022. EtherPose: Continuous Hand Pose Tracking with Wrist-Worn Antenna Impedance Characteristic Sensing. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22)*. Article 58, 12 pages. doi:10.1145/3526113.3545665
- [8] D. Kim, B. Kim, B. Shin, et al. 2022. Actuating compact wearable augmented reality devices by multifunctional artificial muscle. *Nature Communications* 13, 1 (2022), 4155.
- [9] Jiwan Kim, Chi-Jung Lee, Hohurn Jung, Tianhong Catherine Yu, Ruidong Zhang, Ian Oakley, and Cheng Zhang. 2026. WatchHand: Enabling Continuous Hand Pose Tracking On Off-the-Shelf Smartwatches. *arXiv preprint arXiv:2602.21610* (2026).
- [10] M. K. Kim, R. N. Parasuraman, L. Wang, et al. 2019. Soft-packaged sensory glove system for human-like natural interaction and control of prosthetic hands. *NPG Asia Materials* 11, 1 (2019), 43.
- [11] Alexander Kyu, Hongyu Miao, Junyi Zhu, Mayank Goel, and Karan Ahuja. 2024. EITPose: Wearable and Practical Electrical Impedance Tomography for Continuous Hand Pose Estimation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Article 402, 10 pages. doi:10.1145/3613904.3642663
- [12] Chi-Jung Lee, Ruidong Zhang, Devansh Agarwal, Tianhong Catherine Yu, Vipin Gunda, Oliver Lopez, James Kim, Sicheng Yin, Boao Dong, Ke Li, Mose Sakashita, Francois Guimbretiere, and Cheng Zhang. 2024. EchoWrist: Continuous Hand Pose Tracking and Hand-Object Interaction Recognition Using Low-Power Active Acoustic Sensing On a Wristband. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Article 403, 21 pages. doi:10.1145/3613904.3642910
- [13] Jihyun Lee, Minhyuk Sung, Honggyu Choi, and Tae-Kyun Kim. 2023. Im2Hands: Learning Attentive Implicit Representation of Interacting Two-Hand Shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/CVPR52729.2023.02028
- [14] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. 2022. Interacting Attention Graph for Single Image Two-Hand Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/CVPR52688.2022.00278
- [15] Mengcheng Li, Hongwen Zhang, Yuxiang Zhang, Ruizhi Shao, Tao Yu, and Yebin Liu. 2024. HHMR: Holistic Hand Mesh Recovery by Enhancing the Multimodal Controllability of Graph Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/CVPR52733.2024.00068
- [16] Fanqing Lin and Tony R. Martinez. 2024. Ego2HandsPose: A Dataset for Egocentric Two-hand 3D Global Pose Estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 4363–4371.
- [17] Qiuxia Lin, Linlin Yang, and Angela Yao. 2023. Cross-Domain 3D Hand Pose Estimation with Dual Modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 17184–17193. doi:10.1109/CVPR52729.2023.01648
- [18] Yang Liu, Chengdong Lin, and Zhenjiang Li. 2021. WR-Hand: Wearable Armband Can Track User's Hand. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3, Article 118 (2021), 27 pages. doi:10.1145/3478112
- [19] Yilin Liu, Shijia Zhang, and Mahanth Gowda. 2021. NeuroPose: 3D Hand Pose Tracking using EMG Wearables. In *Proceedings of the Web Conference 2021 (WWW '21)*. 1471–1482. doi:10.1145/3442381.3449890
- [20] Y. Luo, Y. Li, P. Sharma, et al. 2021. Learning human–environment interactions using conformal tactile textiles. *Nature Electronics* 4, 3 (2021), 193–201.
- [21] Y. Luo, C. Liu, Y. J. Lee, et al. 2024. Adaptive tactile interaction transfer via digitally embroidered smart gloves. *Nature Communications* 15, 1 (2024), 868.
- [22] A. Moin, A. Zhou, A. Rahimi, et al. 2018. An EMG gesture recognition system with flexible high-density sensors and brain-inspired high-dimensional classifier. In *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 1–5.
- [23] A. Moin, A. Zhou, A. Rahimi, et al. 2021. A wearable biosensing system with in-sensor adaptive machine learning for hand gesture recognition. *Nature Electronics* 4, 1 (2021), 54–63.
- [24] Gyeongsik Moon, Shouu-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. 2020. InterHand2.6M: A Dataset and Baseline for 3D Interacting Hand Pose Estimation from a Single RGB Image. In *Computer Vision – ECCV 2020*.
- [25] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. 2023. AssemblyHands: Towards Egocentric Activity Understanding via 3D Hand Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [26] Joonkyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. 2022. HandOccNet: Occlusion-Robust 3D Hand Mesh Estimation Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [27] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. 2024. Reconstructing Hands in 3D with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [28] Y. Qiu, S. Sun, X. Wang, et al. 2022. Nondestructive identification of softness via bioinspired multisensory electronic skins integrated on a robotic hand. *npj Flexible Electronics* 6, 1 (2022), 45.
- [29] Y. Qiu, Z. Zou, Z. Zou, et al. 2023. Deep-learning-assisted printed liquid metal sensory system for wearable applications and boxing training. *npj Flexible Electronics* 7, 1 (2023), 37.
- [30] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. 2021. FrankMocap: A Monocular 3D Whole-Body Pose Estimation System via Regression and Integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. 1749–1759.
- [31] Muhammad Usama Saleem, Ekkasit Pinyoanuntapong, Mayur Jagdishbhai Patel, Hongfei Xue, Ahmed Helmy, Srijan Das, and Pu Wang. 2025. MaskHand: Generative Masked Modeling for Robust Hand Mesh Reconstruction in the Wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 8372–8383.
- [32] D. Sengupta, J. Romano, and A. G. P. Kottapalli. 2021. Electrospun bundled carbon nanofibers for skin-inspired tactile sensing, proprioception and gesture tracking applications. *npj Flexible Electronics* 5, 1 (2021), 29.
- [33] B. Shao, M. H. Lu, T. C. Wu, et al. 2024. Large-area, untethered, metamorphic, and omnidirectionally stretchable multiplexing self-powered triboelectric skins. *Nature Communications* 15, 1 (2024), 1238.
- [34] Lin Song, Guanya Zhou, Changyunkun Xiao, Qiyu Jiang, Daquan Yang, and He Yu. 2025. CrossHand: Multimodal 3D Hand Reconstruction via Vision and Wearable Sensor Data Fusion. In *Proceedings of the 2025 International Conference on Multimedia Retrieval (ICMR '25)*. 2033–2037. doi:10.1145/3731715.3733472
- [35] Z. Sun, M. Zhu, X. Shan, et al. 2022. Augmented tactile-perception and haptic-feedback rings as human-machine interfaces aiming for immersive interactions. *Nature Communications* 13, 1 (2022), 5224.
- [36] Z. Sun, M. Zhu, X. Shan, et al. 2022. Augmented tactile-perception and haptic-feedback rings as human-machine interfaces aiming for immersive interactions. *Nature Communications* 13, 1 (2022), 5224.
- [37] A. Tashakori, Z. Jiang, A. Servati, et al. 2024. Capturing complex hand movements and object interactions using machine learning-powered stretchable smart textile gloves. *Nature Machine Intelligence* 6, 1 (2024), 106–118.
- [38] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *International Conference on Learning Representations*.
- [39] M. Wang, Z. Yan, T. Wang, et al. 2020. Gesture recognition using a bioinspired learning architecture that integrates visual data with somatosensory data from stretchable sensors. *Nature Electronics* 3, 9 (2020), 563–570.
- [40] Shiyang Wang, Xingchen Wang, Wenjun Jiang, Chenglin Miao, Qiming Cao, Haoyu Wang, Ke Sun, Hongfei Xue, and Lu Su. 2024. Towards Smartphone-based 3D Hand Pose Reconstruction Using Acoustic Signals. *ACM Transactions on Sensor Networks* (2024). doi:10.1145/3677122
- [41] Yintong Wang, Lili Chen, Jiamao Li, and Xiaolin Zhang. 2023. HandGCNFormer: A Novel Topology-Aware Transformer Network for 3D Hand Pose Estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 5675–5684.
- [42] Yinqiao Wang, Hao Xu, Pheng-Ann Heng, and Chi-Wing Fu. 2025. UniHOPE: A Unified Approach for Hand-Only and Hand-Object Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12231–12241.
- [43] Frank Weichert, Daniel Bachmann, Bartholomäus Rudak, and Denis Fisseler. 2013. Analysis of the accuracy and robustness of the leap motion controller. *Sensors* 13, 5 (2013), 6380–6393.

- [44] F. Wen, Z. Zhang, T. He, et al. 2021. AI enabled sign language recognition and VR space bidirectional communication using triboelectric smart glove. *Nature Communications* 12, 1 (2021), 5378.
- [45] E. Wu, Y. Yuan, H. S. Yeo, et al. 2020. Back-hand-pose: 3D hand pose estimation for a wrist-worn camera via dorsum deformation network. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 1147–1160.
- [46] Yingjing Xiao, Zhichao Huang, and Yang Gao. 2025. From Wrist to Finger: Hand Pose Tracking Using Ring-Watch Wearables. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Article 294, 7 pages. doi:10.1145/3706599.3720220
- [47] Yingjing Xiao, Zhichao Huang, Junbin Ren, Haichuan Song, Yang Gao, Yuting Bai, and Zhanpeng Jin. 2025. Wrist2Finger: Sensing Fingertip Force for Force-Aware Hand Interaction with a Ring-Watch Wearable. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. doi:10.1145/3746059.3747767
- [48] Pengfei Xie, Wenqiang Xu, Tutian Tang, Zhenjun Yu, and Cewu Lu. 2024. MS-MANO: Enabling Hand Pose Tracking with Biomechanical Constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2382–2392. doi:10.1109/CVPR52733.2024.00231
- [49] Hao Xu, Haipeng Li, Yinqiao Wang, Shuaicheng Liu, and Chi-Wing Fu. 2024. HandBooster: Boosting 3D Hand-Mesh Reconstruction by Conditional Synthesis and Sampling of Hand-Object Interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [50] Hao Xu, Tianyu Wang, Xiao Tang, and Chi-Wing Fu. 2023. H2ONet: Hand-Occlusion-and-Orientation-Aware Network for Real-Time 3D Hand Mesh Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 17048–17058.
- [51] Ziwei Yu, Chen Li, Linlin Yang, Xiaoxu Zheng, Michael Bi Mi, Gim Hee Lee, and Angela Yao. 2023. Overcoming the Trade-off Between Accuracy and Plausibility in 3D Hand Shape Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [52] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. 2020. MediaPipe Hands: On-device Real-time Hand Tracking. *arXiv preprint arXiv:2006.10214* (2020).
- [53] Qiang Zhang, Yuanqiao Lin, Yubin Lin, and Szymon Rusinkiewicz. 2023. UltraGlove: Hand Pose Estimation with MEMS-Ultrasonic Sensors. In *SIGGRAPH Asia 2023 Conference Papers*. Article 79.
- [54] Xiao Zhang, Griffin Klevering, Juexing Wang, Li Xiao, and Tianxing Li. 2023. RoFin: 3D Hand Pose Reconstructing via 2D Rolling Fingertips. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services (MobiSys '23)*. 330–342. doi:10.1145/3581791.3596838
- [55] Xie Zhang, Chengxiao Li, and Chenshu Wu. 2025. TAPOR: 3D Hand Pose Reconstruction with Fully Passive Thermal Sensing for Around-Device Interactions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, Article 63 (2025). doi:10.1145/3729499
- [56] M. Zhu, Z. Sun, and C. Lee. 2022. Soft modular glove with multimodal sensing and augmented haptic feedback enabled by materials' multifunctionalities. *ACS Nano* 16, 9 (2022), 14097–14110.

A Application Details and User Study Protocol

A.1 Application Scenario and Design Goals

To validate the practical HCI value of PIRO-Pose beyond reconstruction accuracy, we instantiated the system in a VR fine-motor coaching scenario centered on paper-cutting. We selected paper-cutting because it is a representative dexterous manipulation task involving persistent self-occlusion, tool interference, grasp transitions, and fine joint-level coordination. These characteristics make it an appropriate downstream testbed for reduced-observation hand-pose reconstruction.

The application was designed around four goals:

- (1) **Occlusion-free visibility of expert technique.** The system should expose otherwise hidden finger motion through a shared 3D workspace.
- (2) **Replayable expert reference.** Expert demonstrations should be captured as reusable kinematic trajectories for online comparison and replay.

- (3) **Real-time, interpretable discrepancy feedback.** The interface should indicate where the learner deviates, by how much, and in which direction correction is needed.
- (4) **Robustness to tempo variation.** Since novices and experts often move at different speeds, the feedback mechanism should remain semantically meaningful under timing differences.

Accordingly, the application implements the following loop:

expert demonstration capture → temporal alignment → real-time coaching interface

A.2 Expert Demonstration Capture and Temporal Alignment

Expert demonstration capture. An expert performs each practice task while wearing the PIRO-Pose system. For each trial, the system records the full-hand motion sequence together with task meta-data such as task type, repetition index, and start/end time. These demonstrations are stored as independently replayable reference trajectories for both online ghost-hand overlay and offline similarity analysis.

Unified representation. For online comparison, both the expert sequence and the learner sequence are represented in the same kinematic space:

$$s_t = \{q_t, r_{w,t}\}, \quad (6)$$

where $q_t \in \mathbb{R}^{15}$ denotes the intrinsic joint-angle vector (thumb TM/MCP/IP; index–little MCP/PIP/DIP), and $r_{w,t}$ denotes wrist global rotation used for rendering consistency.

Online phase alignment. Because learners may move slower or faster than the expert, direct timestamp matching is not appropriate. We therefore align the learner to the expert within a constrained local search window:

$$\hat{\tau}_t = \arg \min_{\tau \in [\hat{\tau}_{t-1}, \hat{\tau}_{t-1} + v_{\max}]} \|q_t^{\text{usr}} - q_\tau^{\text{exp}}\|_1, \quad (7)$$

where $v_{\max}$ constrains the forward advance of the reference pointer and preserves monotonic temporal consistency. The aligned expert pose at frame t is then

$$\hat{q}_t = q_{\hat{\tau}_t}^{\text{exp}}. \quad (8)$$

Offline trial-level alignment. For post-trial analysis, we align the full learner trajectory and expert trajectory using Dynamic Time Warping (DTW), which provides a stable sequence-level similarity score under variable execution speeds.

A.3 Real-Time Coaching Interface

The interface is organized around a shared 3D workspace and three auxiliary feedback channels:

- (1) **Main workspace:** renders the learner hand, the virtual cutting scene, and the aligned expert ghost hand in a common 3D frame.
- (2) **Visual overlay:** projects joint-level discrepancy heatmaps onto the hand model.
- (3) **Instruction and guidance panel:** displays step information together with signed, actionable textual corrections.
- (4) **Performance dashboard:** visualizes a global trend score to indicate current similarity to the expert trajectory.

A.3.1 Signed Joint Error and Heatmap. When temporal alignment is valid, the instantaneous signed joint deviation is computed as

$$e_t = q_t^{\text{usr}} - \hat{q}_t. \quad (9)$$

The sign indicates whether the learner is over-flexed or under-flexed relative to the expert, while the magnitude determines visual intensity.

To reduce display flicker, we apply short-window temporal smoothing:

$$\bar{e}_t = \frac{1}{W} \sum_{k=0}^{W-1} e_{t-k}, \quad (10)$$

with a default window of $W = 60$ frames at a 60 Hz runtime.

We further use dual thresholds to stabilize cue onset and release:

$$\delta_{\text{on}} = 5^\circ, \quad \delta_{\text{off}} = 3^\circ.$$

A.3.2 Global Trend Score. We first aggregate the joint-wise discrepancy as

$$E_t = \sum_j w_j |e_{t,j}|, \quad \sum_j w_j = 1, \quad (11)$$

where w_j optionally up-weights task-critical joints.

The aggregated discrepancy is then stabilized using an exponential moving average:

$$\tilde{E}_t = \beta \tilde{E}_{t-1} + (1 - \beta) E_t, \quad (12)$$

where β is determined by the runtime update interval and a smoothing constant. Finally, we map the discrepancy to an interpretable 0–100 score:

$$S_t = 100 \exp(-\alpha \tilde{E}_t), \quad (13)$$

where α is fixed from pilot calibration.

A.3.3 Post-trial Similarity Score. For post-trial reporting, the DTW path cost is defined as

$$c(t, \tau) = \|q_t^{\text{usr}} - q_\tau^{\text{exp}}\|_1. \quad (14)$$

Let π denote the DTW alignment path. The average path cost is

$$D = \frac{1}{|\pi|} \sum_{(t, \tau) \in \pi} c(t, \tau). \quad (15)$$

We map this value to a trial-level similarity score:

$$\text{Score} = 100 \exp(-\alpha D). \quad (16)$$

A.3.4 Runtime Gating and Stabilization Discipline. All error-driven feedback is gated by synchronization validity. The aligned ghost hand, heatmap, signed guidance, and trend score are updated only when the system determines that the learner is phase-synchronized with the expert reference. Otherwise, the interface explicitly indicates that the current state is not synchronized and freezes discrepancy-based cues to avoid misleading feedback caused by phase mismatch.

To preserve measurement validity, all stabilization heuristics described above are confined to the feedback layer and are not written back into the quantitative evaluation stream.

A.4 Apparatus and Materials

The application was implemented in a Unity-based VR environment operating at approximately 60 Hz. The same synchronized logging pipeline used in the technical evaluations was retained for the application study, enabling frame-level comparison between the reconstructed learner trajectory and the expert reference trajectory.

We used a hybrid task setup to preserve realistic hand–tool interaction while maintaining controllable measurement conditions. Participants manipulated real scissors and standardized task templates, while the coaching signals, expert overlay, and kinematic similarity measures were computed in the virtual environment. This setup preserved passive haptics and natural grasp behavior while allowing stable, replayable expert references and consistent trajectory logging.

A.5 Within-Subjects User Study

A.5.1 Participants. We recruited $N = 24$ participants (11 female, 13 male; mean age 23.6 ± 2.7 years). Handedness was recorded: 22 participants were right-handed and 2 were left-handed. Most participants had little to no prior paper-cutting experience. On a 7-point self-assessed proficiency scale (1 = no experience, 7 = expert), 12 participants rated 1, 8 rated 2, 3 rated 3, and 1 rated 4. Participants were recruited from a university campus, provided informed consent, and each received a \$15 USD gift card as compensation. Scissor-operation safety was emphasized, and participants could take breaks or withdraw at any time.

A.5.2 Conditions. We used a within-subjects design with two conditions:

- (1) **PIRO-Pose condition.** Participants received the full coaching interface driven by our wearable reconstruction system, including aligned expert ghost-hand overlay, joint-level discrepancy heatmaps, signed textual guidance, and the global trend score.
- (2) **Vision-baseline condition.** Participants performed the same tasks with a state-of-the-art vision-based hand-tracking baseline. The surrounding task environment and expert-reference structure were kept as similar as possible, such that the primary manipulation was the continuity and reliability of the underlying hand-pose estimate.

Condition order was counterbalanced across participants to reduce sequence and fatigue effects.

A.5.3 Tasks and Pacing. We used an easy-to-hard task set spanning straight-line control, continuous turning, and multi-corner contour control:

- **T1:** straight-line cutting,
- **T2:** arc (or single S-curve) cutting,
- **T3:** a multi-corner contour task.

A trial started when the participant began cutting and ended when the participant reported completion. Trials were self-paced. To reduce fatigue and improve comparability, conservative time caps were imposed: 120 s for T1/T2 and 180 s for T3. Trials exceeding the cap were marked as incomplete.

A.5.4 Procedure. Each participant completed both sensing conditions. Each condition block followed the same structure:

- (1) **Calibration and familiarization.** Participants completed the standard calibration routine and received a short introduction to the interface.
- (2) **Baseline assessment.** Participants completed two baseline trials per task without joint-level corrective cues, providing an estimate of initial performance under the current condition block.
- (3) **Training.** Participants practiced under fixed time-matched budgets: T1 = 2 min, T2 = 2 min, and T3 = 6 min. They were instructed to complete as many full attempts as possible within the allotted time.
- (4) **Post-test.** Participants then completed evaluation trials used for objective and subjective analysis.

This design ensured comparable task exposure across conditions while allowing us to evaluate whether PIRO-Pose yielded more effective and less cognitively demanding guidance than the vision baseline.

A.6 Measures

A.6.1 Primary objective measure. Our primary behavioral outcome was joint-level similarity between learner motion and expert reference motion. Let

$$\Delta\theta_{t,j} = \theta_{t,j}^{\text{usr}} - \theta_{t,j}^{\text{exp}} \quad (17)$$

denote the deviation at joint j and aligned frame t . We summarize this as a joint mean absolute error:

$$\text{JMAE} = \sum_j w_j \mathbb{E}_t [|\Delta\theta_{t,j}|], \quad (18)$$

and report the DTW-based trial similarity score in Eq. (16) as a complementary process-level metric.

A.6.2 Secondary measures. We additionally logged task completion status and completion time as secondary indicators of operational efficiency. Where applicable, qualitative post-trial summaries of dominant error contributors were also computed from the aligned joint trajectories.

A.6.3 Subjective measures. After each condition block, participants completed NASA-TLX to assess perceived workload. We also collected short post-study comments regarding feedback interpretability, perceived continuity, and whether the system provided a clear sense of spatial progress during cutting.

A.7 Statistical Analysis

We analyzed repeated measurements using Linear Mixed-Effects Models (LMMs) to account for within-participant dependence and task-level variability. For objective outcomes such as JMAE and DTW similarity, fixed effects included *Condition* (PIRO-Pose vs. Vision), *Task* (T1–T3), and *Phase* where applicable, with participant modeled as a random intercept. For subjective measures such as NASA-TLX, we used the same repeated-measures framework with participant as a random intercept.

When post-hoc comparisons were required, multiple comparisons were corrected using a Holm-style procedure. In addition to significance testing, effect sizes and confidence intervals were reported wherever appropriate.

A.8 Relationship to the Main Text

Section 6 of the main paper reports only the core application-level findings due to page constraints. This appendix supplements that summary by documenting the application design, expert-reference alignment mechanism, real-time feedback computation, participant demographics, task protocol, measurement definitions, and statistical analysis plan in sufficient detail for interpretation and replication.