ViPoser: Sparse-IMU Based Human Pose Estimation with Distilled Vision Foundation Priors

Anonymous Author(s)

ABSTRACT

Human pose estimation using inertial measurement units (IMUs) in daily life has gained increasing attention due to its ability to enable camera-free motion capture for applications such as healthcare monitoring, sports analysis, and immersive VR/AR systems. However, estimating full-body pose from only one to three daily carried consumer-grade IMUs (e.g., smartphones, smartwatches, and earphones) is highly under-constrained, as sparse sensor coverage and noisy signals often lead to unstable predictions and physically implausible body configurations. To address this challenge, we propose ViPoser, a lightweight pose estimation framework that transfers human-structure priors from a large vision foundation model into an IMU-based model through knowledge distillation. Specifically, we analyze the human-centric vision model Sapiens to identify layers rich in biomechanical and skeletal knowledge, distill these representations into a compact student Transformer, and integrate the distilled priors into a multi-stage pose estimation pipeline that performs local joint estimation, full-body masked completion, and global motion recovery. This design enables the model to infer anatomically plausible full-body poses even under sparse IMU observations while remaining suitable for real-time edge deployment. Experiments on multiple real-world datasets show that ViPoser consistently outperforms state-of-the-art IMU-based methods such as IMUPoser and MobilePoser, achieving 10~30% lower pose estimation error. In addition, ViPoser produces more anatomically realistic poses, significantly reducing failure cases observed in prior methods where predicted limbs intersect the body or violate biomechanical constraints.

1 INTRODUCTION

Human pose estimation (HPE) aims to infer the configuration of human body parts from sensor observations. While early methods predominantly relied on visual inputs such as images and videos, recent research has expanded to non-visual sensing modalities, including wireless/radio signals and inertial measurement units (IMUs) [1, 45, 50, 53, 54, 56]. These techniques are increasingly adopted across a wide range of domains, including healthcare (fall detection [20], which is reported to cause ~646,000 deaths annually [47]), sports and rehabilitation such as exercise form correction, and injury prevention[41], and immersive computing like VR/AR body avatar animation, with a market of $16B [21, 50].

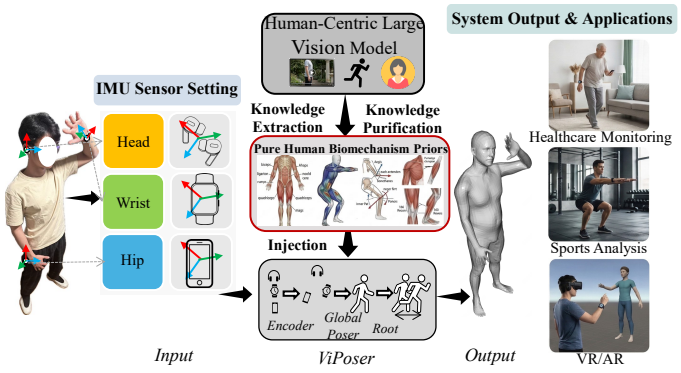

Figure 1: Overview of ViPoser. Sparse IMU signals from everyday devices are processed by a lightweight poser initialised with human priors distilled from a vision foundation model, enabling various applications.

Vision-based approaches for HPE have achieved high accuracy on benchmarks and in controlled settings [5, 37]. However, several practical limitations hinder their real-world deployment: (1) they require the user to remain within camera coverage, which is often unrealistic in daily scenarios; (2) occlusions are common and are frequently mitigated only by multi-view camera setups; and (3) continuous video capture raises privacy concerns [14, 37]. To overcome these issues, IMU-based approaches estimate full-body pose from wearable inertial sensors, enabling flexible, camera-free motion capture. Early systems typically rely on the fixed configuration of around six IMUs attached to specific body locations [19, 45, 53, 54, 56]. Yet such setups remain difficult to popularize because everyday users are unlikely to wear multiple dedicated sensors in prescribed positions [35, 50]. Motivated by this gap, recent work targets full-body pose estimation using IMUs from only one to three consumer devices such as smartphone, smartwatch, or earbuds [1, 35, 50]. However, relying purely on data-driven regression without explicitly acquiring human physiological priors (e.g., skeletal structure and biomechanical constraints [4]), these models often predict physically impossible body configurations. For safety-critical applications like fall detection, such physically implausible artifacts represent a fatal system failure.

By nature, IMU signals inherently carry limited information about human body structure and physiological movement patterns; in contrast, such information can be readily extracted from visual data. Large foundation models [3, 10, 24, 38], trained on massive pretraining corpora, can

internalize broad, reusable representations and factual associations that serve as general background knowledge for downstream tasks [3]. Among them, Sapiens [25] is a family of large vision models centered on humans that have been trained on more than 300 million in-the-wild human images, scaled from 0.3B to 2B parameters. It reports state-of-the-art performance across multiple human-centric downstream tasks, including body-part segmentation, depth estimation, and surface normal prediction [25]. These results suggest that large-scale human-focused visual pretraining may provide transferable priors related to human structure and biomechanical regularities [25]. Since our IMU-based HPE task also depends on such priors, a natural question arises: **Can we extract these priors from vision models and constructively inject them into our IMU HPE model in a lightweight way?**

While this idea sounds compelling, there are substantial challenges to realize it. First, vision foundation models are trained on images or videos [25], whereas our target task uses inertial measurement signals. Directly transferring such knowledge to IMU inputs may therefore introduce irrelevant priors and modality mismatch, potentially leading to redundancy or misinterpretation across modalities. Additionally, vision foundation models are often too large for real-time edge inference. Second, directly predicting full-body pose from only 1∼3 IMU sensors is a severely under-constrained problem: the sparse sensor coverage leaves the vast majority of joints unobserved, making end-to-end models prone to unstable and physically implausible outputs.

In this paper, we first explore the feasibility of realizing this idea. We decompose the Sapiens-0.3B model [25] and attach trainable, lightweight task heads to the latent features output by different layers, performing masked image recovery and pose estimation respectively. This allows us to investigate the extent to which different portions of the model encode general visual knowledge versus human structural and biomechanical regularities. Through comparative experiments, layers 18–21 are identified as containing the richest human body priors while retaining only a relatively small amount of low-level visual knowledge. However, these layers still remain too large for direct deployment on edge devices. We therefore distill them into a compact student model, which is then integrated into our redesigned full-body pose estimator. Due to the sparse IMU configurations, a key source of missing constraint is skeletal connectivity, i.e., neighboring joints are coupled by rigid limbs and anatomical range-of-motion limits, imposing rich implicit regularities that a vision model naturally absorbs from large-scale human imagery. We therefore abandon the end-to-end paradigm and reformulate whole-body inference as a *masked completion* problem: given local estimates at a sparse set of instrumented joints, recover the remaining unobserved ones. This mirrors the masked autoencoding objective that underlies the pretraining of most vision foundation models [13], making it a task our distilled student is naturally suited to solve.

We propose ViPoser, as illustrated in Fig. 1, a lightweight pose estimation framework that transfers human-centric priors from large vision models into an IMU-based model through knowledge distillation. Unlike prior methods such as MobilePoser [50] and IMUPoser [35] that collapse the IMU modeling into an end-to-end mapping, ViPoser consists of three stages: (1) a local module predicts joint poses at instrumented sites from raw IMU streams; (2) a masked-completion module leverages the distilled vision priors to infer the remaining joints; and (3) a translation module recovers global movement with foot–ground contact to suppress drift. Unlike prior methods that train the entire network in one stage, we adopt a two-phase training strategy. In the first phase, each of the three modules is trained independently until they converge. In the second phase, the three modules are connected and trained end-to-end, with the overall loss defined as a weighted sum of the individual module losses.

We evaluate the effectiveness of ViPoser using MobilePoser [50] and IMUPoser [35] as our primary baselines. All models are trained exclusively on synthetic data and evaluated on several real-world datasets, including DIP-IMU [19], IMUPoser [35], and TotalCapture [44]. ViPoser outperforms the baselines across all settings. For example, on the DIP-IMU dataset, ViPoser achieves a Mean Per Joint Vertex Error (MPJVE) of 13.94 cm, a Mean Per Joint Position Error (MPJPE) of 10.80 cm, and a Mean Per Joint Rotation Error (MPJRE) of 26.66°, outperforming both MobilePoser (15.32 cm / 11.89 cm / 29.43°) and IMUPoser (16.4 cm / 12.78 cm / 29.92°). Furthermore, compared with our base variant that does not incorporate Sapiens-derived priors, these metrics improve from 15.00 cm / 11.79° / 28.08 cm (MPJVE / MPJPE / MPJRE) to the reported values, indicating that transferring human-centric priors from vision models can meaningfully benefit IMU-based pose estimation.

Meanwhile, low quantitative error does not necessarily mean an estimation of plausible human pose since they can still violate the human joint angle limits of the human body [46]. To quantify this, we introduce *Joint Violation Rate* (JVR), a metric that measures the percentage of frames in which any predicted joint angle exceeds the clinical range-of-motion limits published by the American Academy of Orthopaedic Surgeons (AAOS) [11]. ViPoser achieves an extremely low JVR (0.03%, 0.03%, and 0.07% across three benchmarks), whereas IMUPoser (0.21%–0.72%) and MobilePoser (0.31%–1.10%) violate these limits far more frequently. Moreover, removing the distilled priors from ViPoser noticeably increases JVR, indicating that the human-centric structural knowledge transferred from the vision model actively helps the poser avoid anatomically implausible configurations.

In this paper, our main contributions are:

- We propose transferring human body priors learned by large-scale vision foundation models to compact IMU-based pose estimators. Through a layer-wise probing analysis of Sapiens 0.3B, we identify the layers that are richest in human structural and biomechanical knowledge yet carry relatively little general visual information, providing a principled basis for selecting what to transfer.
- We develop a lightweight distillation pipeline that extracts the identified prior-rich layers into a compact module deployable on resource-constrained devices. Ablation studies confirm that incorporating the distilled priors yields consistent improvements over a base variant without them, demonstrating their value for sparse-IMU pose estimation.
- We design ViPoser, a modular full-body pose estimation architecture using three stages including sparse-joint pose recovery from IMU signals, prior-guided full-body pose inference, and global movement prediction. We show that this decomposition, combined with the distilled priors, outperforms state-of-the-art methods on multiple real-world benchmarks.

2 BACKGROUND AND MOTIVATION

2.1 IMU-Based Human Pose Estimation

Human pose estimation (HPE) is central to various applications including healthcare [20], sports and rehabilitation such as exercise form correction, injury prevention, and remote physical therapy monitoring [41], and immersive computing like VR/AR body avatar animation. Improvements in HPE can benefit many of those downstream tasks [19, 35, 50].

Inertial measurement units (IMUs) measure linear acceleration, angular velocity, and sometimes magnetic field orientation. Being low-cost and widely embedded in everyday devices such as smartphones, smartwatches, and earbuds, IMUs serve as a versatile sensing backbone for mobile applications. However, full-body pose estimation is particularly challenging under sparse-IMU sensing, because the model must reconstruct a high-dimensional body configuration from signals at only a few body locations. Early IMU-based HPE methods such as SIP [45], DIP [19], TransPose [54], PIP [53], and DynaIP [56] recover full-body pose from six professional-grade IMUs and report strong accuracy on standard benchmarks. However, the six-sensor setup remains logistically burdensome for everyday use. To reduce instrumentation, IMUPoser [35] demonstrates that full-body pose can be estimated from the consumer-grade IMUs already present in a smartphone, smartwatch, and earbuds, making the setting broadly accessible with clear potential in fitness and health scenarios. MobilePoser [50] further incorporates physical cues such as foot–ground contact into its multi-stage estimation pipeline to improve temporal stability and global translation tracking. Other recent efforts explore augmenting sparse IMU signals with complementary modalities. For example, Ultra Inertial Poser [2] and UltraPoser [28] fuse IMU data with ultra-wideband ranging or ultrasound sensing to mitigate drift and improve coverage.

Despite the progress, sparse-IMU based HPE remains fundamentally underconstrained: the model must infer the configuration of 24 body joints from sensors at only a few locations (e.g., pelvis via a phone in a pocket, wrist via a watch, head via earbuds), leaving substantial information gaps, which makes existing methods tend to rely on pure data-driven regression rather than human body structure. This underscores the need for incorporating knowledge like skeletal structure and biomechanical regularities to compensate for the missing observations when inferring full-body pose from 1∼3 sensors.

2.2 HPE with Vision Foundation Model

In recent years, large vision models have achieved widespread adoption. Among them, Sapiens [25] is a Vision Transformer (ViT)–based family of models released by Meta that targets human-related tasks. Leveraging scaling and pretraining on about billions of image–text tokens, such models often achieve strong few-shot and even zero-shot transfer on downstream tasks after light adaptation [22, 31, 39]. Human centric tasks such as pose estimation are no exception. For example, after pretrained with millions of human images, Sapiens can achieve the state-of-the-art performance in several human centric tasks including HPE [25, 58]. Sapiens scales up ViT backbones and natively supports 1K-resolution inference, enabling the encoder to capture finer spatial details than typical 224px backbones [8, 25]. The models are pretrained in a self-supervised manner on a curated corpus of roughly 300 million in-the-wild human images and then lightly fine-tuned for downstream tasks [13, 25]. Sapiens involves 0.3b to 2b parameters, and after fine-tuning Sapiens achieves state-of-the-art on body-part segmentation, depth estimation, and surface-normal prediction, illustrating that human-related priors learned from large-scale visual data can substantially benefit human-centric applications [25].

2.3 Why Human-centric Priors Can Help

These observations naturally prompt the question: **can the knowledge learned by human-centric vision foundation models be repurposed for IMU-based tasks?** While many parameters in vision foundation models specialize in visual signal processing, a non-trivial portion implicitly encodes human-related priors such as skeletal structure, joint-angle limits, and everyday biomechanics, as evidenced by the

fact that mask-recovery task pretrained models transfer effectively to other diverse tasks including part segmentation, depth, and surface normal prediction [25].

A fundamental limitation of current IMU-based methods is that they treat pose estimation purely as a data-driven in-out mapping: given observations at 1∼3 body sites, predict the positions of remaining joints, without any explicit encoding of biomechanical feasibility. The model has no awareness that the predicted locations correspond to human joints governed by skeletal constraints and physiological range-of-motion limits. Consequently, predictions can violate basic anatomical plausibility, e.g., joints may twist beyond physically possible angles, or limbs may penetrate the body mesh entirely. Fig. 2 illustrates representative failure cases: IMUPoser (center) produces a pose in which the left forearm intersects the torso, while MobilePoser (right) folds the arm into an anatomically impossible configuration behind the back. Such errors are not merely aesthetic; for instance, safety-critical applications such as fall detection for older adults - where falls are leading cause of injury-related mortality [47], implausible pose estimates can lead to missed or false detections with serious consequences. This motivates explicitly endowing IMU models with human-structural knowledge to constrain predictions to the manifold of anatomically plausible poses.

A key challenge for IMU-based applications is that sensors observe only a few local body segments (e.g., a wrist-worn device) and must infer the global body state from sparse cues [45]. Despite recent progress in self-supervised representation learning for IMU (e.g., LIMU-BERT [49]) and studies documenting domain shift across placements, devices, and users in practical HAR deployments [48], current IMU-specific models rarely encode human biomechanics explicitly. In contrast, vision models benefit from holistic full- or half-body context and large-scale pretraining that enables effective transfer after lightweight adaptation [3, 39, 58]. We therefore advocate importing human-centric visual priors into lightweight IMU models to bridge the gap between local sensing and global human understanding.

Admittedly, it is attractive to lightweightly extract prior knowledge about human body, learned by vision foundation models, and use it to guide IMU-based tasks. However, two practical challenges remain. First, state-of-the-art vision Transformers contain hundreds of millions to billions of parameters (e.g., ViT-22B [6]), making direct edge deployment unrealistic; IMU-centric applications such as HAR and full-body pose estimation additionally require real-time processing, further tightening latency and energy budgets [29, 35]. Second, not all knowledge in a vision model is relevant for IMU inference: large portions of the parameters specialize in processing low-level image statistics, which risks introducing redundant or spurious cues when transferred cross-modally [9]. This motivates *selective* transfer—distilling only task-relevant representations via parameter-efficient methods such as knowledge distillation, adapters/LoRA, or pruning and quantization—rather than wholesale reuse of the entire vision model [12, 15, 17, 18].

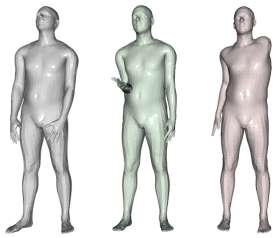

Figure 2: Failure cases of IMU-based HPE. Left to right: ground truth, IMUPoser [35], and MobilePoser [50].

3 LAYER-WISE ANALYSIS OF HUMAN-CENTRIC PRIORS IN VISION MODELS

We adopt Sapiens [25] as the vision teacher which is pretrained via masked autoencoders on over 300 million in-the-wild human images. We use the smallest variant, Sapiens-0.3B, which consists of L=24 Transformer blocks with hidden dimension d=1024. Formally, let $\mathcal{T} = \{T_1, T_2, \ldots, T_L\}$ denote the ordered sequence of Transformer blocks, and let $\mathbf{z}^{(l)} \in \mathbb{R}^{N \times d}$ be the token-level output of block T_l, where N is the number of patch tokens. Even this smallest variant is far too heavy for on-device IMU inference; our goal is therefore *not* to run the vision model at test time, but to compress its knowledge into a compact IMU student via distillation [12, 15].

Sapiens follows an encoder–head paradigm: a pretrained ViT encoder is shared across task-specific heads [25]. Prior analysis of ViTs has shown that, unlike CNNs, they aggregate global information already in shallow layers and maintain relatively uniform representations across depth [40]. For our downstream IMU-based pose estimation task, the most relevant knowledge concerns human body structure and biomechanical regularities (e.g., skeletal connectivity and joint-angle constraints), rather than appearance-level cues such as texture or lighting. We therefore adopt the *pose-finetuned* Sapiens checkpoint as the teacher. Our goal is to identify layers whose representations are rich in human-structure priors yet comparatively free of low-level, image-specific information—knowledge that would be redundant or even misleading when the downstream input modality is sparse IMU data rather than images.

3.1 Probing Protocol

To quantify the trade-off between human-structure content and image-specific content across encoder depth, we design two linear-probing tasks on the frozen Sapiens encoder. Throughout the probing study, all parameters of $\mathcal{T}$ are frozen;

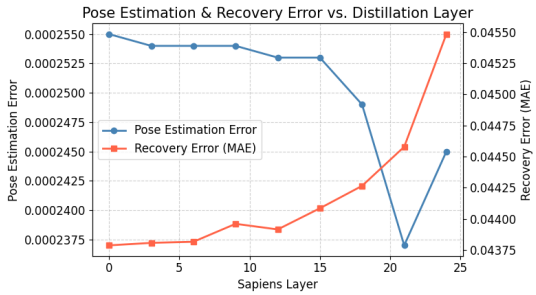

Figure 3: Layer-wise probing on COCO validation.

only the lightweight task heads described below are trainable. To prevent the adapter from learning strong task-specific representations on its own, we train all probes on a small random subset of 2,000 images from the COCO training split and evaluate on the full COCO validation set [30].

Probe 1: Pose estimation (human-structure content). Given a person crop $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, we feed it through the first l blocks of $\mathcal{T}$ to obtain $\mathbf{z}^{(l)}$. A lightweight deconvolutional head h_{pose} then predicts heatmaps for the 17 COCO keypoints [30]:

$$\hat{\mathbf{H}}^{(l)} = h_{\text{pose}}(\mathbf{z}^{(l)}), \tag{1}$$

where $\hat{\mathbf{H}}^{(l)} \in \mathbb{R}^{17\times H'\times W'}$ and $H'\times W'$ is the heatmap resolution. We train the probe with mean absolute error (MAE) between predicted and ground-truth heatmaps:

$$\mathcal{L}^{(l)}_{\text{pose}} = \frac{1}{17\,H'W'} \sum_{k=1}^{17} \sum_{i,j} \left|\hat{\mathbf{H}}^{(l)}_{k,i,j} - \mathbf{H}^{*}_{k,i,j}\right|, \tag{2}$$

where $\mathbf{H}^{*}$ is the ground-truth heatmap. We evaluate the converged test error $\mathcal{E}^{(l)}_{\text{pose}}$ for $l \in \{0, 3, 6, 9, 12, 15, 18, 21, 24\}$. A *lower* $\mathcal{E}^{(l)}_{\text{pose}}$ indicates that $\mathbf{z}^{(l)}$ encodes richer human-structure information.

Probe 2: Masked image reconstruction (image-specific content). Following the masked autoencoder protocol [13], we randomly mask 75% of the input patches and pass only the visible tokens through the first l blocks to obtain $\mathbf{z}^{(l)}_{\text{vis}}$. A single-layer Transformer decoder h_{rec} reconstructs the full image, and we measure the reconstruction error on masked patches:

$$\hat{\mathbf{I}}^{(l)} = h_{\text{rec}}(\mathbf{z}^{(l)}_{\text{vis}}), \tag{3}$$

$$\mathcal{L}^{(l)}_{\text{rec}} = \frac{1}{|\mathcal{M}|} \sum_{p\in\mathcal{M}} \left\|\hat{\mathbf{I}}^{(l)}_{p} - \mathbf{I}^{*}_{p}\right\|_1, \tag{4}$$

where $\hat{\mathbf{I}}^{(l)} \in \mathbb{R}^{H\times W\times 3}$, $\mathcal{M}$ denotes the set of masked patches, and $\mathbf{I}^{*}_{p}$ is the ground-truth patch. A *lower* $\mathcal{E}^{(l)}_{\text{rec}}$ indicates that $\mathbf{z}^{(l)}$ retains more low-level, image-specific information.

3.2 Results and Layer Selection

As shown in Fig. 3 the pose probe error $\mathcal{E}^{(l)}_{\text{pose}}$ remains relatively flat for $l \in \{0, \ldots, 15\}$ and then drops markedly between layers 15 and 21, with the most pronounced improvement occurring after layer 18. This suggests that layers 18–21 encode the richest human body-structure priors. Unexpectedly, tapping features from the final three layers ($l \in \{22, 23, 24\}$) causes $\mathcal{E}^{(l)}_{\text{pose}}$ to increase sharply. We attribute this to the role of the final layers: after task-specific fine-tuning, they primarily project intermediate representations into the space expected by the official pose head, rather than encoding generic body-structure knowledge. As supporting evidence, the original Sapiens encoder paired with its official pose adapter achieved a pose estimation error of 2.31×10^{-4}, substantially outperforming our lightweight probe trained on only 2,000 images, confirming that the final-layer features are co-adapted to a high-capacity head and do not transfer well through a minimal adapter. In Fig. 3 the reconstruction error $\mathcal{E}^{(l)}_{\text{rec}}$ shows an overall increasing trend with layer depth, despite minor fluctuations in the shallow layers. This indicates that shallower layers still retain substantial image-specific representations, whereas deeper layers progressively discard pixel-level appearance in favor of task-relevant abstractions [40, 55].

Combining both observations, layers 18–21 emerge as the optimal source for our purposes: they maximize human-structure content (lowest $\mathcal{E}_{\text{pose}}$) while minimizing image-specific content (high $\mathcal{E}_{\text{rec}}$). We therefore extract features in $\{T_{19}, T_{20}, T_{21}\}$ from the Sapiens encoder to serve as supervision targets for our subsequent compact IMU model.

4 ViPoser

Motivated by the analysis, we propose ViPoser, a lightweight cross-modal guidance method that transfers human-centric background knowledge from a vision foundation model pretrained on large-scale real-world human images to an IMU model. Concretely, we adopt Sapiens as the teacher which is pretrained with a masked-autoencoder objective on ~300M in-the-wild human images and then fine-tuned for various tasks [8, 13, 25]. We choose the smallest released variant, Sapiens-0.3B, to constrain compute; this checkpoint uses a 24-layer ViT encoder with 1024-dimensional embeddings [25]. Even so, naively invoking such a teacher during IMU inference on edge devices is impractical; thus model compression/distillation is required to meet real-time constraints typical for mobile sensing [12, 29, 35].

4.1 Model Distillation

As motivated in Section 3, we leverage human-centric priors from a vision teacher and empirically focus on deeper

Sapiens blocks that exhibit stronger task-relevant representations for human-structure cues. Specifically, we target blocks 18–21 of the Sapiens-0.3B encoder (24 layers total) [25].

Notation. Let $\mathbf{z}^{(l)} \in \mathbb{R}^{N\times 1024}$ denote the output of the l-th Transformer block of the frozen Sapiens encoder, where N is the number of image patch tokens. We use $\mathbf{z}^{(18)}$ as the *input target* and $\mathbf{z}^{(21)}$ as the *supervision target* for distillation.

Projection-based distillation design. The structure of the distillation model is shown in Fig 4. Since the student operates in a 64-dimensional space while the teacher operates in a 1024-dimensional space, we introduce a pair of linear projectors to bridge the dimensional gap. Specifically, let $\mathbf{W}_{\text{in}} \in \mathbb{R}^{64\times 1024}$ be the down-projection and $\mathbf{W}_{\text{out}} \in \mathbb{R}^{1024\times 64}$ be the up-projection. To bias the projectors toward pure dimension conversion rather than task-specific feature learning, we enforce the transpose constraint and define the forward pass as:

$$\mathbf{W}_{\text{out}} = \mathbf{W}_{\text{in}}^{\top}, \quad (5) \qquad \hat{\mathbf{z}}^{(21)} = \mathbf{W}_{\text{in}}^{\top} \cdot f_S\Big(\mathbf{W}_{\text{in}} \cdot \mathbf{z}^{(18)}\Big), \quad (6)$$

where f_S denotes the 2-layer, 64-dimensional Transformer student. Only f_S is retained after training; the projectors are discarded at inference time.

Training objective. The total loss comprises a distillation term $\mathcal{L}_{\text{distill}}$ and a regularisation term $\mathcal{L}_{\text{reg}}$. $\mathcal{L}_{\text{distill}}$ measures ℓ_1 error between the teacher's layer-21 outputs $\mathbf{z}^{(21)}$ and the student's reconstructed features $\hat{\mathbf{z}}^{(21)}$, encouraging the student to match the teacher's representation. $\mathcal{L}_{\text{reg}}$ applies ℓ_1 regularisation [43] to the projector weights $\mathbf{W}_{\text{in}}$, encouraging sparsity so the projectors learn a minimal, dimension-matching mapping instead of absorbing excessive task-specific knowledge from the teacher. The combined objective is:

$$\mathcal{L} = \mathcal{L}_{\text{distill}} + \mathcal{L}_{\text{reg}} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{\mathbf{z}}_i^{(21)} - \mathbf{z}_i^{(21)}\right\|_1 + \lambda \left\|\mathbf{W}_{\text{in}}\right\|_1, \quad (7)$$

where λ controls the regularisation strength. The full Sapiens encoder is frozen throughout; only f_S and $\mathbf{W}_{\text{in}}$ are optimised. After training, only the compact student f_S is retained for downstream pose estimation.

4.2 Pose Estimation

We consider full-body pose estimation as the primary downstream task for our distilled model, where a user carries only a small set of IMUs (smartphone, smartwatch, earbuds). Following IMUPoser and MobilePoser [35, 50], we define five candidate placement sites—right pocket, left pocket, right wrist, left wrist, and head/earbuds—and support one to three active IMU streams at inference by masking absent sites. This design allows a single model to operate robustly across common daily-use configurations.

Unlike prior work that directly regresses full-body pose from sparse IMUs with a single sequence model [35, 50],

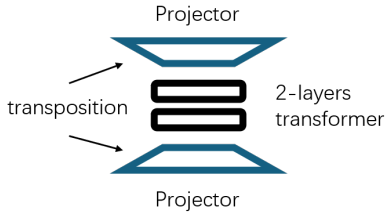

Figure 4: The structure of distillation design. The input teacher features $\mathbf{z}^{(18)}$ are projected down to 64 dimensions, processed by the 2-layer student f_S, and projected back to 1024 dimensions to match $\mathbf{z}^{(21)}$. The two projectors share weights via the transpose constraint (Eq. 5) and are discarded after training.

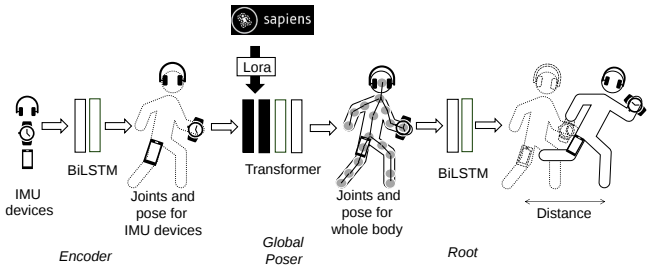

Figure 5: The pipeline of the pose estimation system.

we adopt a three-stage pipeline: (1) a compact BiLSTM first estimates local joint positions and orientations at the instrumented sites; (2) a masked-completion Transformer, whose encoder is distilled from Sapiens, infers the full body positions and rotations; (3) a BiLSTM model to predict the global translation base on the whole-body pose. The complete pipeline is shown in Fig. 5.

4.2.1 Stage 1: Local Joint Estimation at IMU Sites. Each IMU stream provides a 12-dimensional input vector per timestep, comprising a 3-D linear acceleration $\mathbf{a} \in \mathbb{R}^3$ and a flattened 3×3 rotation matrix $\mathbf{R} \in \mathbb{R}^9$ representing orientation:

$$\mathbf{x}_t = [\,\mathbf{a}_t \parallel \text{vec}(\mathbf{R}_t)\,] \in \mathbb{R}^{12}. \quad (8)$$

The five streams are concatenated into a 60-dimensional input and processed by a BiLSTM (hidden size 256, 2 layers) [16], which produces, for each of the five ROI sites, an estimated 3-D joint position $\hat{\mathbf{p}} \in \mathbb{R}^3$ and a 6-D continuous rotation representation $\hat{\mathbf{r}} \in \mathbb{R}^6$ [59]. Decoupling local inference at instrumented sites from whole-body completion provides cleaner intermediate representations for the stage 2.

4.2.2 Stage 2: Whole-Body Completion. Recovering globally consistent motion from 1~3 IMUs is substantially harder than local estimation: IMUs provide only local accelerations and

orientations at sparse body locations with no absolute position reference and significant integration drift [35, 50]. We address this by incorporating human-centric priors distilled from Sapiens. We treat whole-body inference as a masked-autoencoding problem, a common design in vision field [13]: given the local estimates from Stage 1 at the instrumented sites, the network completes the whole body joints.

Global Poser input. The Global Transformer Poser receives a per-timestep feature vector that concatenates four streams:

$$\mathbf{h}_t = \left[\, \hat{\mathbf{p}}_t \parallel \hat{\mathbf{r}}_t \parallel \mathbf{x}_t \parallel \Delta\hat{\mathbf{p}}_t \,\right] \in \mathbb{R}^{5\times(3+6+12+3)}, \tag{9}$$

where $\hat{\mathbf{p}}_t$ and $\hat{\mathbf{r}}_t$ are the Stage-1 position and rotation estimates for the five ROI sites, $\mathbf{x}_t$ is the raw IMU input (Eq. (8)), and $\Delta\hat{\mathbf{p}}_t = \mathrm{d}\hat{\mathbf{p}}_t/\mathrm{d}t$ is the temporal velocity of the predicted ROI positions.

Distilled Transformer encoder with LoRA. The encoder is the compact 2-layer, 64-dimensional student Transformer f_S obtained via the distillation procedure described in Section 4.1. To preserve the human-centric priors encoded in f_S while adapting to the IMU input space, we freeze the student parameters and attach LoRA adapters [18] to its linear projections, training only the low-rank updates $\Delta\mathbf{W} = \mathbf{BA}$ (rank $r \ll 64$). Another 2-layer, 64-dimensional Transformer decoder predicts the whole-body pose with outputs of a 24×9 matrix per frame, representing the flattened 3×3 rotation matrix for each SMPL joint. At inference the 24×9 output is projected to the nearest valid rotation via SVD [27].

4.2.3 Stage 3: Global Translation and Foot Contact. A 2-layer, 256-dimensional BiLSTM head receives the concatenation of the predicted 24×9 pose and the raw 60-D IMU input, and jointly regresses root velocity $\hat{\mathbf{v}}_{\text{root}} \in \mathbb{R}^3$ and left/right foot-contact logits $\hat{c} \in \mathbb{R}^2$. Global translation is obtained by cumulative integration of $\hat{\mathbf{v}}_{\text{root}}$. Following MobilePoser [50], contact signals suppress drift by zeroing velocity when a foot is detected as planted, stabilising trajectories in mobile settings. The full training objective and the staged optimisation schedule are described in Section 4.3.

4.3 Training Procedure

Staged Optimisation. Because each stage conditions on the predictions of the previous one at inference, naively training the full pipeline end to end from scratch is unstable: early modules produce noisy outputs that cause downstream training to diverge. We therefore adopt a two-phase schedule. In *Phase I* we pretrain the three stages—the BiLSTM encoder (Stage 1), the Global Transformer Poser (Stage 2), and the root head (Stage 3)—in isolation, with each module receiving its corresponding ground-truth input and optimised with only its own subset of the losses defined below (i.e. $\mathcal{L}_{\text{j}} + \mathcal{L}_{\text{p}}$ for Stage 1, $\mathcal{L}_{\text{GP}}$ for Stage 2, and $\mathcal{L}_{\text{vel}} + \mathcal{L}_{\text{ct}}$ for Stage 3). In *Phase II* we jointly train the entire pipeline end to end with the full objective (10), where each module receives only the predictions of its upstream module rather than ground truth, forcing the network to adapt to its own prediction errors. To bridge the gap between the clean ground-truth inputs used in *Phase I* and the imperfect predictions encountered in *Phase II* and at inference, we inject zero-mean Gaussian noise into the *Phase I* inputs: $\sigma_{\text{roi}} = 0.04$ for the ROI joint positions fed to the Global Poser, and $\sigma_{\text{pose}} = 0.02$ for the 24×9 pose matrices fed to the root head during *Phase I* training.

Loss Function. The overall training objective is a weighted sum of six terms, each supervising a different aspect of the pipeline:

$$\mathcal{L} = \lambda_{\text{GP}}\,\mathcal{L}_{\text{GP}} + \lambda_{\text{sm}}\,\mathcal{L}_{\text{sm}} + \lambda_{\text{j}}\,\mathcal{L}_{\text{j}} + \lambda_{\text{p}}\,\mathcal{L}_{\text{p}} + \lambda_{\text{vel}}\,\mathcal{L}_{\text{vel}} + \lambda_{\text{ct}}\,\mathcal{L}_{\text{ct}}. \tag{10}$$

Stage-1 supervision (ROI encoder). $\mathcal{L}_{\text{j}}$ is the masked mean squared error on the five ROI joint positions predicted by the BiLSTM encoder, computed only at active (non-masked) sites. $\mathcal{L}_{\text{p}}$ is the geodesic distance on $SO(3)$ between predicted and ground-truth ROI rotations, likewise masked:

$$\mathcal{L}_{\text{p}} = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} \arccos\!\left(\frac{\mathrm{tr}(\hat{R}_s^{\top} R_s^{*}) - 1}{2}\right), \tag{11}$$

where $\mathcal{S}$ denotes the set of active ROI sites.

Stage-2 supervision (whole-body pose). $\mathcal{L}_{\text{GP}}$ is the mean squared error between the predicted 24×9 rotation matrices $\hat{R}$ and the ground truth R^*; $\mathcal{L}_{\text{sm}}$ penalises temporal jitter via the mean squared discrete acceleration of $\hat{R}$ along the time axis:

$$\mathcal{L}_{\text{GP}} = \mathrm{MSE}\big(\hat{R},\, R^*\big), \tag{12}$$

$$\mathcal{L}_{\text{sm}} = \frac{1}{T-2}\sum_{t=2}^{T} \left\| \hat{R}_t - 2\hat{R}_{t-1} + \hat{R}_{t-2} \right\|^2. \tag{13}$$

Stage-3 supervision (root head). $\mathcal{L}_{\text{vel}}$ is the MSE on root velocity, and $\mathcal{L}_{\text{ct}}$ is binary cross-entropy on foot-contact logits.

The loss weights are set to $(\lambda_{\text{GP}}, \lambda_{\text{sm}}, \lambda_{\text{j}}, \lambda_{\text{p}}, \lambda_{\text{vel}}, \lambda_{\text{ct}}) = (6.5,\ 450,\ 25,\ 0.68,\ 2,\ 0.05)$. Due to the differing scales of each loss term, these weights are calibrated so that the effective gradient is dominated by full-body pose reconstruction, with ROI encoder supervision as secondary and root translation, foot contact and temporal smoothness terms as lightweight regularisation.

5 EXPERIMENT

We train our model on a synthetic dataset rendered from SMPL-based motion sequences [33, 34] and then evaluate it on real-world IMU datasets, including DIP-IMU, TotalCapture, and the IMUPoser dataset [19, 35, 44].

5.1 Datasets

5.1.1 Synthetic Training Dataset. For synthetic pre-training, we follow MobilePoser and build our corpus from the AMASS

motion-capture archive [34], which retargets a large collection of mocap datasets to the SMPL/SMPL-H body models [33]. AMASS provides roughly 40 hours of high-quality 3D human motion with full-body joint rotations and global root translation, covering everyday activities, sports, and complex transitions across many subjects and capture setups. AMASS serves as the only source of synthetic pose trajectories used to pre-train the IMU-to-pose backbone.

5.1.2 Real-world Evaluation Dataset. For evaluation on real data, we use three public inertial datasets:

DIP-IMU[19] records 10 subjects with commercial-grade Xsens IMUs at 60 Hz and provides accurate full-body pose but no global translation, focusing on everyday movements such as stretches, squats, lunges, and punches. **TotalCapture[44]** provides synchronized Xsens IMU streams and optical mocap ground truth at 60 Hz, including both articulated pose and global translation for multi-person indoor activities. **IMUPoser [35]** contains 10 participants captured with commodity devices (iPhone 11 Pro, Apple Watch, and AirPods) at 25 Hz, together with ground-truth SMPL pose and root translation, and covers a broad range of daily, exercise, and freestyle motions. Together, these datasets span both commercial-grade and consumer-grade IMUs and allow us to test generalization across hardware and motion domains.

5.1.3 Data Pre-processing. All signals (virtual IMU readings, joint rotations/positions, root translation, and foot-contact labels) are resampled to a fixed frame rate (30Hz) and segmented into overlapping windows of fixed length (125 frames) for training. For each window, we randomly select between one and three of the five possible device locations as "active" and keep their IMU readings; all remaining device slots are set to zero. This masking strategy mirrors MobilePoser's design [50] and allows the model to handle different subsets of available consumer devices at inference time while training on a unified representation.

5.2 Evaluation

5.2.1 Quantitative Evaluation Metric. Following prior work [35, 50], we evaluate pose estimation quality with the metrics below, where lower values indicate better performance, and we also adopt MPJVE as the primary metric and DIP-IMU as the primary dataset to facilitate direct comparison.

- **Mean Per-Joint Rotation Error (MPJRE)**: Mean angular error between predicted and ground-truth joint rotations, averaged over all root-aligned joints and all frames, reported in degrees (°).
- **Mean Per-Joint Position Error (MPJPE)**: Mean Euclidean distance between predicted and ground-truth joint positions, averaged over all root-aligned joints and all frames, reported in centimeters (cm).
- **Mean Per-Vertex Position Error (MPJVE)**: Mean Euclidean distance between predicted and ground-truth SMPL body-mesh vertices, averaged over all root-aligned vertices and all frames, reported in centimeters (cm).
- **Mean Translation Error (MTE)**: Mean Euclidean distance between the predicted and ground-truth global root joint positions (i.e., body translation in the world coordinate frame), averaged over all frames and reported in centimeters (cm), without applying root alignment.

5.2.2 Biomechanical Validity Metric. While standard metrics such as MPJPE and MPJVE measure geometric accuracy, low reconstruction error does not necessarily imply anatomical plausibility: a model can achieve competitive quantitative scores while still producing poses that violate the physiological range-of-motion (ROM) limits of human joints [46]. As discussed in Section 2.3, such implausible predictions are not merely cosmetic failures—they can trigger downstream errors in safety-critical applications. We therefore introduce a complementary metric that directly evaluates whether predicted poses respect human anatomical constraints.

Joint Violation Rate (JVR). We check whether predicted rotations exceed the normative range-of-motion limits published by the American Academy of Orthopaedic Surgeons (AAOS) [11], a widely adopted clinical reference that tabulates maximum flexion, extension, abduction, and adduction angles for each major joint. We decompose each local rotation $R \in SO(3)$ into swing and twist components via quaternion projection [7]:

$$R = R_{\text{swing}} \cdot R_{\text{twist}}, \tag{14}$$

where the twist is taken about the bone's longitudinal axis (Y in the SMPL local frame). The swing quaternion is further resolved into two independent angular components: flexion/extension (θ_{flex}, swing about X) and abduction/adduction (θ_{abd}, swing about Z).

For each joint we define per-axis limits from the AAOS table: $\theta^{\max}_{\text{flex}}$ is set to the larger of the published flexion and extension values, and $\theta^{\max}_{\text{abd}}$ is set analogously from the abduction and adduction values. A 5% tolerance is applied to accommodate individual variation and sensor noise. Ball-and-socket joints (shoulder, hip, ankle, wrist) are checked on both axes independently, while hinge joints (elbow, knee) are checked on flexion/extension only. A frame is flagged as violated if any joint exceeds its limit on any applicable axis (with the 5% tolerance), and JVR is the percentage of violated frames over the sequence; lower is better.

5.3 Comparison with Existing Works

We compared our model, ViPoser, with other existing models, IMUposer and Mobileposer [35, 50], in the literature.

Table 1: Quantitative comparison on 3 datasets. Best results are shown in bold.

Dataset	Model	MPJVE↓ (cm)	MPJRE↓ (deg)	MPJPE↓ (cm)	MTE↓ (cm)
DIP-IMU	IMUPoser	16.40	29.92	12.78	–
	MobilePoser	15.32	29.43	11.89	–
	ViPoser (Ours)	**13.94**	**26.67**	**10.80**	–
IMUPoser	IMUPoser	14.70	25.44	11.41	–
	MobilePoser	14.18	24.79	11.03	16.91
	ViPoser (Ours)	**12.85**	**22.44**	**10.07**	**16.87**
Total-Capture	IMUPoser	16.44	26.91	13.51	–
	MobilePoser	17.88	28.88	14.77	**25.77**
	ViPoser (Ours)	**13.90**	**23.18**	**11.48**	26.69

IMUposer[35]: IMUPoser is a real-time full-body pose estimation system that uses only the IMUs embedded in everyday consumer devices such as smartphones, smartwatches, and earbuds. The method defines a set of possible on-body device locations and trains a bidirectional LSTM network to directly regress 3D full-body pose from any valid subset of available devices, without requiring specialized motion-capture hardware. This represents a strong IMU-only baseline for sparse, user-controlled instrumentation.

Mobileposer[50]: MobilePoser extends this line of IMUposer by estimating both full-body pose and 3D global translation from commodity mobile IMUs. It employs a multi-stage neural network for kinematic pose estimation followed by a physics-based motion optimizer, enabling real-time, on-device inference from any subset of phones, watches, and earbuds while achieving state-of-the-art accuracy under sparse IMU configurations. This makes MobilePoser a strong, real-time baseline closely aligned with our problem setting.

5.4 Quantitative Overall Evaluation

We first compare the overall performance of our models and 2 baseline models. As shown in Table 1, our method consistently outperforms both IMUPoser and MobilePoser in all pose estimation metrics (MPJVE, MPJRE, MPJPE) across the three evaluation datasets. The improvement is most pronounced on TotalCapture, where MPJVE is reduced from 17.88 cm (MobilePoser) to 13.90 cm. We attribute these gains to the distilled human-structure priors, which provide biomechanical regularization that is especially beneficial when IMU signals are sparse and ambiguous.

For global translation estimation, DIP-IMU provides no trajectory annotations and IMUPoser lacks a translation module, so MobilePoser is the only available baseline on the remaining two datasets. Our method achieves comparable translation errors to MobilePoser on the IMUPoser dataset

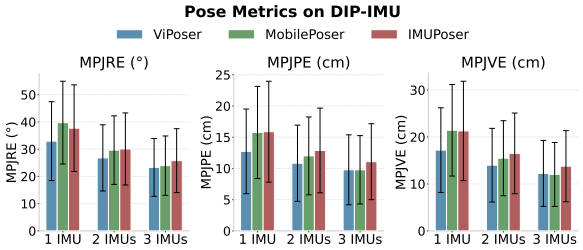

Figure 6: Pose estimation results for 1–3 IMU configurations on DIP-IMU dataset.

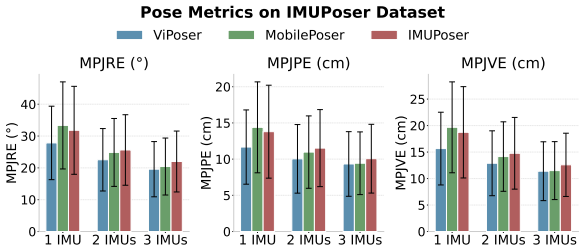

Figure 7: Pose estimation results for 1–3 IMU configurations on IMUPoser dataset.

(16.87 cm vs. 16.91 cm) and slightly higher error on TotalCapture (26.69 cm vs. 25.77 cm). This is expected, as our knowledge distillation targets joint-level pose quality rather than global translation, which depends primarily on foot-ground contact estimation and velocity integration, where ViPoser uses a single shared BiLSTM to jointly predict root velocity and foot contact to minimise inference latency while MobilePoser employs two separate modules for these tasks.

5.5 Robustness Across IMU Configurations

In daily use, a person may carry between one and three IMU devices at varying body locations. To evaluate robustness under these realistic conditions, we test all models on the DIP-IMU and IMUPoser datasets across every valid 1–3 device combination, without any additional fine-tuning beyond the base training [19, 35]. Results are shown in Fig. 6 and Fig. 7.

ViPoser consistently outperforms both baselines across all device counts and both datasets. Notably, the performance gap widens as active IMUs decrease: with 3 IMUs, all methods perform comparably (e.g., MPJVE ≈ 12-14 cm on DIP-IMU), whereas with a single IMU the advantage of ViPoser becomes more pronounced. This trend is precisely what we would expect: fewer sensors means fewer kinematic constraints from the input signal, making the model more reliant on human-structure priors internalized from the vision teacher. The fact that our advantage grows in the most under-constrained setting validates the core motivation of our approach.

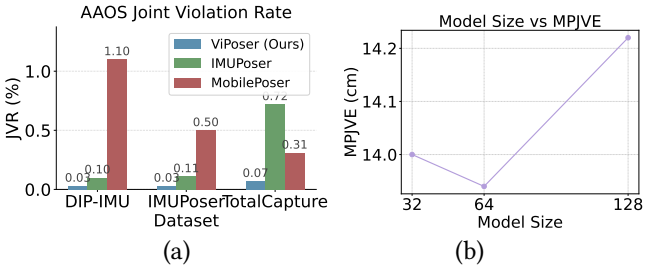

Figure 8: (a) AAOS Joint Violation Rate (%) across three datasets; ViPoser consistently produces the most anatomically plausible motion. (b) Performance of different student model sizes on DIP-IMU.

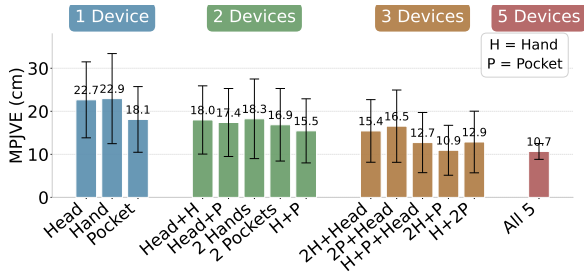

Figure 9: MPJVE under different device placements.

5.6 Biomechanical Validity Evaluation

We evaluate the biomechanical validity using the Joint Violation Rate (JVR) defined in Section 5.2.2, which measures the percentage of frames containing at least one anatomically impossible joint angle according to the AAOS clinical reference [11]. Fig. 8 (a) reports JVR across three benchmark datasets. ViPoser achieves the lowest violation rate on all datasets, with JVR at 0.03% on both DIP-IMU and IMUPoser, and 0.07% on TotalCapture—an order of magnitude lower than the baselines in most settings. IMUPoser gets relatively low JVRs on DIP-IMU and IMUPoser (0.10%, 0.11%) but degrades substantially on TotalCapture (0.72%). MobilePoser exhibits the highest violation rate on DIP-IMU (1.10%), indicating that its predicted rotations fall outside the anatomical range of motion more frequently.

5.7 Ablation Studies

We conduct ablation studies to validate the key design choices of our system, covering device placement, student model capacity, and the distillation layer selection strategy.

5.7.1 Device Placement. We investigate how different IMU placement configurations affect pose estimation accuracy with ViPoser, shown in Fig. 9.

Among single-device configurations, the hip-pocket IMU contributes most to accuracy (MPJVE 18.1 cm), while head and hand placements perform similarly but worse (22.7 cm and 22.9 cm respectively). This is consistent with the intuition that hip motion provides the most informative signal for lower-body pose, which accounts for the majority of body joints. With two devices, the hand–pocket combination yields the best result (15.5 cm), outperforming two-hand (18.3 cm) / two-pocket (16.9 cm) setups, suggesting that coverage of both upper and lower limbs is more beneficial than redundant coverage of a single region. With three devices, the best configuration achieves 10.9 cm MPJVE, a 40% reduction from the best single-device result, confirming that each additional device provides meaningful complementary information. Besides, the MPJVE reaches 10.7 cm while the model gets IMU inputs from all 5 positions.

5.7.2 Distillation Model Size. A larger student model may capture richer knowledge but incurs higher computational cost, which conflicts with our on-device deployment constraints. We compare students of hidden dimension 32, 64, and 128 while keeping all other settings fixed. As shown in Fig. 8(b), the 64-channel student achieves the best MPJVE (13.94 cm), outperforming both the 32-channel variant (14.00 cm, insufficient capacity) and the 128-channel variant (14.22 cm, prone to overfitting given the limited distillation signal). We adopt the 64-channel student in all subsequent experiments.

5.7.3 Distillation Layer Selection. Given a fixed student capacity, distilling a wider range of teacher layers forces more diverse knowledge into the same bottleneck, inevitably coarsening the learned representations. We evaluate this from two complementary angles.

Impact of Distillation Depth. This experiment serves a dual purpose: beyond identifying the optimal layer depth, it provides a direct empirical test of a central claim of this paper—that transferring human-centric priors from a vision foundation model genuinely benefits IMU-based pose estimation. To this end, we include a no-knowledge random initialisation baseline (*FromRandom*)—the same three-stage pipeline without any knowledge transferred from Sapiens—alongside three distillation variants. As shown in Fig. 10(a), the *FromRandom* yields an MPJVE of 15.00 cm, which is clearly worse than *all* distillation variants regardless of which layers are used. This confirms that vision priors provide a consistent and meaningful benefit to HPE, validating the core motivation of our approach. Among the distilled variants, blocks 0–3 reduce error to 14.53 cm, while blocks 9–12 perform worse at 14.85 cm—even below the shallow-layer variant. This non-monotonic pattern is consistent with the probing findings in Section 3: mid-depth layers carry mixed representations that are neither purely structural nor cleanly image-specific, making them the least transferable target for an

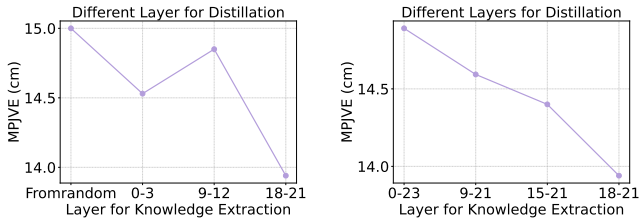

Figure 10: Impact of the distillation layer selection of Sapiens. (a) The impact of the distillation depth. (b) The impact of the distillation range.

IMU student. Blocks 18–21 achieve the best result (13.94 cm, 7.1% improvement), further corroborating that the deeper, human-structure-rich layers identified in Section 3 are the most informative source of cross-modal knowledge for our task.

Impact of Distillation Range. The relation between the range of the layer and the content inside are already discussed in Section 3, but we are still curious about how those features impact the performance of pose estimation. As shown in Fig. 10 (b), progressively restricting the distillation range to deeper blocks yields monotonically decreasing MPJVE: 14.89 cm (blocks 0–23), 14.59 cm (9–21), 14.40 cm (15–21), and 13.94 cm (18–21). Broader coverage does not translate to richer student knowledge; rather, shallower layers dilute the supervision signal with image-specific redundancy that is irrelevant—and potentially harmful—for the IMU modality.

5.7.4 Biomechanism Knowledge Transfer in Joints Violation. To disentangle the contribution of the distilled human priors in producing anatomically plausible human poses from the architectural design, we compare ViPoser against an identical model trained from random initialisation (denoted *FromRandom*). Table 2 reports the AAOS Joint Violation Rate on all three benchmarks.

FromRandom already achieves a considerably low JVR compared with the external baselines (cf. Fig. 8 (a)), indicating that our pipeline architecture inherently encourages anatomically plausible outputs. However, ViPoser with distilled priors consistently reduces JVR by a further 1.5×–7× compared to *FromRandom*, reaching near-zero violation rates across all datasets. This gap demonstrates that the human biomechanical knowledge transferred from the vision foundation model provides meaningful additional guidance for producing anatomically valid poses, beyond what the architecture alone can achieve.

Taken together, these ablations support a clear principle for cross-modal knowledge transfer from a large vision

Table 2: Ablation on knowledge distillation: AAOS JVR between *FromRandom* and ViPoser.

Method	DIP-IMU	IMUPoser	TotalCapture
FromRandom	0.20	0.19	0.10
ViPoser	**0.03**	**0.03**	**0.07**

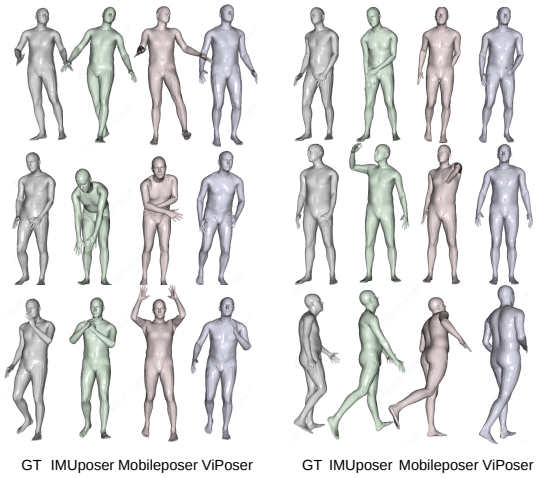

Figure 11: Qualitative comparison of pose estimation results. Left to right: Ground truth (gray), IMUPoser (green), MobilePoser (red), and our ViPoser (blue).

model to a modality-constrained IMU student: *precision matters more than coverage*. A moderately sized student (64-dim) avoids both underfitting and overfitting to the distillation signal. More importantly, targeting a narrow, semantically coherent band of teacher layers (18–21) outperforms both broader distillation ranges and arbitrarily chosen layer positions, with a 7.1% gain over the *FromRandom*, and also get lower violate rate. Distilling shallower or wider layer ranges dilutes the supervision signal with image-specific features that are redundant—or misleading—for an IMU input space, confirming the layer analysis in Section 3.

5.8 Qualitative Results

Fig. 11 visualizes inference results under a three-IMU setup (smartphone in the right pocket, smartwatch on the left wrist, earbuds on the head), comparing ground truth (gray), IMUPoser (green), MobilePoser (red), and our ViPoser (blue). As expected, all methods struggle to fully recover body parts that have no direct IMU coverage, and our method is no exception. However, as highlighted in Section 2.3, a critical failure mode of IMU-based methods is the generation of anatomically implausible poses—limbs intersecting the body or joints twisted beyond physiological limits. As shown

Table 3: Inference efficiency on DIP-IMU (1448 frames).

Method	Total (ms)	Per-frame (ms)	FPS
IMUPoser	192.43	0.133	7,524.8
MobilePoser	1206.16	0.833	1,200.5
ViPoser (Ours)	**156.69**	**0.108**	**9,241.2**

in Fig. 11, ViPoser exhibits such artifacts far less frequently than both baselines, demonstrating that the human-structure priors distilled from Sapiens effectively act as a biomechanical regularizer, constraining predictions to the manifold of plausible human motion even in the absence of direct observations at uninstrumented joints.

5.9 Inference Efficiency

We evaluate the inference computational overhead of all three methods on the DIP-IMU dataset (1448 frames, 48.27 s of motion) on a consumer-grade laptop equipped with an NVIDIA RTX 3070 GPU. Results are reported in Table 3.

Our method achieves a per-frame inference time of 0.108 ms, the lowest among all three methods, corresponding to a throughput of approximately 9,241 FPS on a consumer GPU. Compared to MobilePoser, which processes each frame in 0.833 ms, our approach is nearly 7.7× faster—a direct result of replacing the heavy vision backbone with a compact distilled student at runtime. Relative to IMUPoser (0.133 ms per frame), our method is also marginally faster despite incorporating additional human-structure priors, demonstrating that the knowledge distillation design successfully decouples the representational richness of the vision teacher from its computational cost. These results confirm that our system achieves the intended goal of lightweight yet knowledge-rich inference, and establish a practical foundation for future real-time deployment on edge devices.

6 RELATED WORK

Vision-Based Human Pose Estimation. Vision-based methods have long dominated HPE research. Early CNN-based approaches such as the Stacked Hourglass network [36] established the multi-scale heatmap detection. OpenPose [5] enabled real-time multi-person tracking via introducing Part Affinity Fields for limb association. HRNet [42] advanced single-person estimation by maintaining high-resolution representations through parallel multi-resolution subnetworks with repeated cross-scale fusion. More recently, ViTPose [51] demonstrated that a plain Vision Transformer backbone suffices to achieve state-of-the-art accuracy on COCO, and Sapiens [25] further scaled human-centric pretraining to over 300 M images, achieving strong transfer across pose, segmentation, depth, and surface normal tasks. Despite their accuracy, vision-based methods require camera coverage and are susceptible to occlusion and privacy concerns, limiting deployment in unconstrained mobile settings.

IMU-Based Human Pose Estimation. IMU-based methods offer camera-free, occlusion-robust alternatives for motion capture. SIP [45] pioneered full-body SMPL reconstruction from 6 IMUs via offline optimization. DIP [19] enabled real-time inference by training a bidirectional RNN on synthesized IMU data. TransPose [54] improved accuracy through a multi-stage architecture with intermediate joint-position estimation and added global translation tracking. PIP [53] introduced a physics-aware motion optimizer to enforce physical plausibility. TIP [23] adopted a Transformer decoder to improve temporal consistency, and DynaIP [56] proposed part-based modeling with pseudo-velocity regression trained on real inertial data, improving generalization over synthetic-only training. To move beyond six dedicated sensors, IMUPoser [35] explored pose estimation from any subset of consumer-device IMUs (phone, watch, earbuds), and MobilePoser [50] further added global translation estimation with real-time on-device inference. However, all these methods frame pose estimation as pure input–output mapping without explicitly encoding biomechanical priors, which can yield anatomically implausible outputs.

Other Sensor Modalities for Pose Estimation. Alternative modalities have also been explored. RF-Pose [57] used WiFi-frequency signals with cross-modal vision supervision to estimate 2D poses through walls. Person-in-WiFi 3D [52] extended this to multi-person 3D estimation from commercial WiFi devices, and HuPR [26] established a benchmark for mmWave radar-based pose estimation. NeuroPose [32] demonstrated hand pose tracking from wearable EMG sensors by exploiting muscle-activation signals.

7 CONCLUSION

We propose ViPoser, which addresses the fundamental limitations of sparse-IMU based pose estimation by leveraging the structural intelligence of large vision foundation models. By distilling biomechanical priors from Sapiens into a lightweight Transformer, the framework effectively bridges the gap between limited sensor data and anatomically accurate motion capture. Experimental results demonstrate that ViPoser not only yields a 10–30% reduction in error over current state-of-the-art methods but also ensures physical plausibility in real-time scenarios. ViPoser establishes a robust pathway for deploying high-fidelity, camera-free motion capture on edge devices, making it a viable solution for healthcare, athletics, and immersive technologies.

REFERENCES

[1] Karan Ahuja, Sven Mayer, Mayank Goel, and Chris Harrison. 2021. Pose-on-the-go: Approximating user pose with smartphone sensor fusion and inverse kinematics. In *Proceedings of the 2021 CHI Conference*

on Human Factors in Computing Systems. 1–12.

[2] Rayan Armani, Changlin Qian, Jiaxi Jiang, and Christian Holz. 2024. Ultra inertial poser: Scalable motion capture and tracking from sparse inertial sensors and ultra-wideband ranging. In *ACM SIGGRAPH 2024 Conference Papers*. 1–11.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[4] Yize Cai, Baoshen Guo, Flora Salim, and Zhiqing Hong. 2025. Towards generalizable human activity recognition: A survey. *arXiv preprint arXiv:2508.12213* (2025).

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence* 43, 1 (2019), 172–186.

[6] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. 2023. Scaling vision transformers to 22 billion parameters. In *International conference on machine learning*. PMLR, 7480–7512.

[7] PrzemysŁ Dobrowolski et al. 2015. Swing-twist decomposition in clifford algebra. *arXiv preprint arXiv:1506.05481* (2015).

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

[9] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence* 2, 11 (2020), 665–673.

[10] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[11] Walter B. Greene and James D. Heckman. 1994. *The Clinical Measurement of Joint Motion* (1 ed.). American Academy of Orthopaedic Surgeons, Rosemont, IL.

[12] Song Han, Huizi Mao, and William J Dally. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149* (2015).

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16000–16009.

[14] Carlos Hinojosa, Juan Carlos Niebles, and Henry Arguello. 2021. Learning privacy-preserving optics for human pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2573–2582.

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[16] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[17] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International conference on machine learning*. PMLR, 2790–2799.

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *Iclr* 1, 2 (2022), 3.

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

[20] Raul Igual, Carlos Medrano, and Inmaculada Plaza. 2013. Challenges, issues and trends in fall detection systems. *Biomedical engineering online* 12, 1 (2013), 66.

[21] IMARC Group. 2024. Virtual Reality Market Size, Share and Trends 2025–2033. https://www.imarcgroup.com/virtual-reality-market. Accessed 2026-02-28.

[22] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*. PMLR, 4904–4916.

[23] Yifeng Jiang, Yuting Ye, Deepak Gopinath, Jungdam Won, Alexander W Winkler, and C Karen Liu. 2022. Transformer inertial poser: Real-time human motion reconstruction from sparse imus with simultaneous terrain generation. In *SIGGRAPH Asia 2022 Conference Papers*. 1–9.

[24] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

[25] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. 2024. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*. Springer, 206–228.

[26] Shih-Po Lee, Niraj Prakash Kini, Wen-Hsiao Peng, Ching-Wen Ma, and Jenq-Neng Hwang. 2023. Hupr: A benchmark for human pose estimation using millimeter wave radar. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5715–5724.

[27] Jake Levinson, Carlos Esteves, Kefan Chen, Noah Snavely, Angjoo Kanazawa, Afshin Rostamizadeh, and Ameesh Makadia. 2020. An analysis of svd for deep rotation estimation. *Advances in Neural Information Processing Systems* 33 (2020), 22554–22565.

[28] Yadong Li, Shuning Wang, Yongjian Fu, Justin Chen, Xingyu Chen, Ju Ren, Xinyu Zhang, Akshay Gadre, and Ke Sun. 2025. UltraPoser: Pushing the Limits of IMU-based Full-Body Pose Estimation with Ultrasound Sensing on Consumer Wearables. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 1–15.

[29] Ji Lin, Wei-Ming Chen, Yujun Lin, Chuang Gan, Song Han, et al. 2020. Mcunet: Tiny deep learning on iot devices. *Advances in neural information processing systems* 33 (2020), 11711–11722.

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.

[32] Yilin Liu, Shijia Zhang, and Mahanth Gowda. 2021. NeuroPose: 3D hand pose tracking using EMG wearables. In *Proceedings of the Web Conference 2021*. 1471–1482.

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 851–866.

[34] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. 2019. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5442–5451.

[35] Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. 2023. Imuposer: Full-body pose estimation using imus in phones, watches, and earbuds. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–12.

[36] Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*. Springer, 483–499.

[37] Ana Filipa Rodrigues Nogueira, Hélder P Oliveira, and Luís F Teixeira. 2025. Markerless multi-view 3D human pose estimation: A survey. *Image and Vision Computing* 155 (2025), 105437.

[38] OpenAI. 2024. *Hello GPT-4o.* https://openai.com/index/hello-gpt-4o/

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[40] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. 2021. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems* 34 (2021), 12116–12128.

[41] Ali Raza, Azam Mehmood Qadri, Iqra Akhtar, Nagwan Abdel Samee, and Maali Alabdulhafith. 2023. LogRF: An approach to human pose estimation using skeleton landmarks for physiotherapy fitness exercise correction. *IEEE Access* 11 (2023), 107930–107939.

[42] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5693–5703.

[43] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58, 1 (1996), 267–288.

[44] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. 2017. Total capture: 3d human pose estimation fusing video and inertial sensors. In *Proceedings of 28th British Machine Vision Conference*. 1–13.

[45] Timo Von Marcard, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll. 2017. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer graphics forum*, Vol. 36. Wiley Online Library, 349–360.

[46] Bastian Wandt and Bodo Rosenhahn. 2019. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7782–7791.

[47] World Health Organization. 2021. Falls. https://www.who.int/news-room/fact-sheets/detail/falls. WHO fact sheet, accessed 2026-02-28.

[48] Huatao Xu, Pengfei Zhou, Rui Tan, and Mo Li. 2023. Practically adopting human activity recognition. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–15.

[49] Huatao Xu, Pengfei Zhou, Rui Tan, Mo Li, and Guobin Shen. 2021. Limu-bert: Unleashing the potential of unlabeled data for imu sensing applications. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 220–233.

[50] Vasco Xu, Chenfeng Gao, Henry Hoffmann, and Karan Ahuja. 2024. Mobileposer: Real-time full-body pose estimation and 3d human translation from imus in mobile consumer devices. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–11.

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

[52] Kangwei Yan, Fei Wang, Bo Qian, Han Ding, Jinsong Han, and Xing Wei. 2024. Person-in-wifi 3d: End-to-end multi-person 3d pose estimation with wi-fi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 969–978.

[53] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. 2022. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13167–13178.

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

[55] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? *Advances in neural information processing systems* 27 (2014).

[56] Yu Zhang, Songpengcheng Xia, Lei Chu, Jiarui Yang, Qi Wu, and Ling Pei. 2024. Dynamic inertial poser (dynaip): Part-based motion dynamics learning for enhanced human pose estimation with sparse inertial sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1889–1899.

[57] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. 2018. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7356–7365.

[58] Wenqi Zheng and Yutaka Arakawa. 2025. Few-shot Vision-based Human Activity Recognition with MLLM-based Visual Reinforcement Learning. *arXiv preprint arXiv:2508.10371* (2025).

[59] Yijun Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2019. On the Continuity of Rotation Representations in Neural Networks. In *CVPR*. https://openaccess.thecvf.com/content_CVPR_2019/papers/Zhou_On_the_Continuity_of_Rotation_Representations_in_Neural_Networks_CVPR_2019_paper.pdf