Nymeria: A Massive Collection of Multimodal Egocentric Daily Motion in the Wild

Lingni Ma[1], Yuting Ye[1], Fangzhou Hong[2†], Vladimir Guzov[3†], Yifeng Jiang[4†], Rowan Postyeni[1], Luis Pesqueira[1], Alexander Gamino[1], Vijay Baiyya[1], Hyo Jin Kim[1], Kevin Bailey[1], David S. Fosas[1], C. Karen Liu[4], Ziwei Liu[2], Jakob Engel[1], Renzo De Nardi[1], and Richard Newcombe[1]

[1] Meta Reality Labs Research
[2] Nanyang Technological University, Singapore
[3] University of Tübingen and Max Planck Institute for Informatics, Germany
[4] Stanford University, USA
https://www.projectaria.com/datasets/nymeria

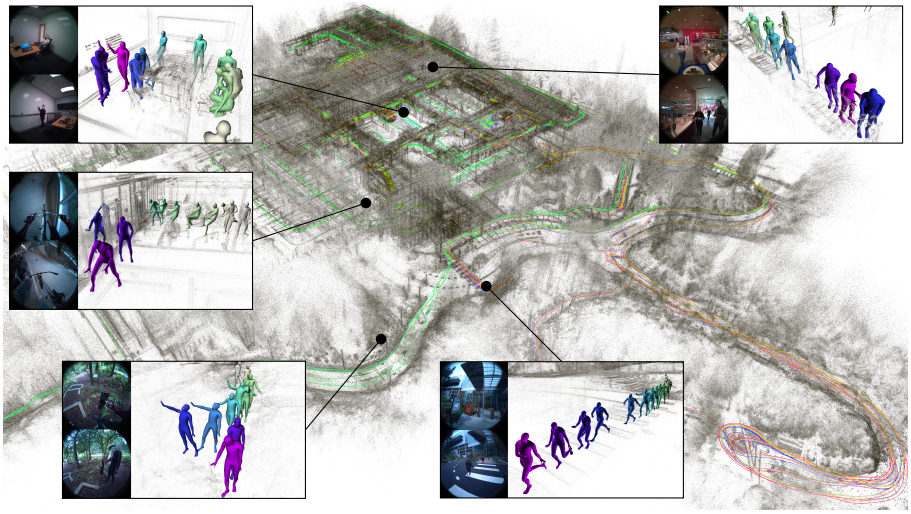


Fig. 1: A glimpse of Nymeria dataset. The figure shows example indoor and outdoor activities captured on a campus, where the point clouds and trajectories are the SLAM output by tracking all egocentric devices *i.e.* the glasses and wristbands. Each sub-figure is a motion clip from a different participant, where the top left gives the latest egocentric view, the right is the 3D localized full-body motion synchronized with the headset and the bottom left provides an auxiliary third-person view.

Abstract. We introduce Nymeria - a large-scale, diverse, richly annotated human motion dataset collected in the wild with multiple multi-

† Work done during internships at Meta Reality Labs Research.

modal egocentric devices. The dataset comes with a) full-body ground-truth motion; b) multiple multimodal egocentric data from Project Aria devices with videos, eye tracking, IMUs and etc; and c) an third-person perspective by an additional "observer". All devices are precisely synchronized and localized in one metric 3D world. We derive hierarchical protocol to add in-context language descriptions of human motion, from fine-grain motion narrations, to simplified atomic actions and high-level activity summarization. To the best of our knowledge, Nymeria dataset is the world's largest human motion in the wild; first of its kind to provide synchronized and localized multi-device multimodal egocentric data; and the world's largest motion-language dataset. It provides 300 hours of daily activities from 264 participants across 50 locations, total travelling distance over 399Km . The language descriptions contain 310.5K sentences in 8.64M words from a vocabulary size of 6545. To demonstrate the potential of the dataset we evaluate several SOTA algorithms for egocentric body tracking, motion synthesis, and action recognition.

Keywords: human motion · egocentric · multimodal · dataset

1 Introduction

The advent of AI is leading to a surge of smart glasses [1–3, 5, 6, 9, 11, 30] and other wearables. These devices not only provide seamless access to LLM-based AI assistants, but also are multimodal data-capture vehicles that provide immediate and long-term personalized context, allowing AI assistants to evolve into the next generation of human-centric contextualized AI, and unlock a new era of *contextualized computing* combined with AR/VR technology.

In this paradigm, the wearer's own body motion and action provides important context. The problem is challenging to solve given insufficient self-observations from wearable devices. In reality, a critical limiting factor to advance research is data. Currently, people either use real data with limited scale, diversity and modality [35, 40, 47, 57, 69, 79, 106, 113] or simulations that lack realism and completeness [12, 13, 46, 48, 60, 99]. There are three key technical challenges to create large motion datasets.

- *Obtaining long-term ground-truth motion in the wild.* There are two main motion capture (mocap) approaches. Vision-based solutions such as the ones relying on optical markers [57, 69, 79] or cameras [35, 56, 63, 101, 113] are adversely affected by line-of-sight visibility, and require a complex multi-camera-setup to cover limited range of motions inside a constrained volume. Inertial-based solutions [7, 10] suffer from dead-reckoning and thus are inferior in global positioning accuracy [70, 89].
- *Multi-device alignment.* Combining multiple capture devices or ground-truth systems requires accurate temporal and spatial alignment, which can be challenging with off-the-shelf hardware that cannot be modified or lack support for universal synchronization protocols. The existing datasets work around this problem using visual cues [35, 56, 113] or audio [40]. Such approaches

offer limited accuracy and reliability. In order to counter clock drift for long recordings, they can be intrusive and interrupt the natural activity. Consequently, existing datasets mostly record short motions (*cf*. Tab. 2).
– *Data processing and annotations.* These are critical for a dataset to develop its full potential. In addition to the body motion, device localization and scene representation, we believe natural language descriptions are crucial for future research directions. Existing work provides simple descriptions or action labels without scene context [26,36,83], and is of significantly smaller scale compared to the text corpus for training LLMs [16,78,100].

To fill the gap and accelerate the research, we introduce Nymeria - the world largest human motion dataset with 300 hours in-the-wild daily activities from 264 participants performing 20 scenarios from 50 indoor and outdoor locations. With avarege 15-min per recording, the data captures natural activities with spontaneous unscripted actions and authentic interactions. Nymeria is first-of-its-kind dataset recorded with multiple multimodal egocentric devices. Participants worn XSens mocap suit [7], Project Aria glasses [30] and Aria-alike wristbands to record egocentric motion, RGB, grayscale, eye tracking (ET) videos, inertial measurement units (IMUs), magnetometer, barometer and etc. Devices are synchronized with a non-intrusive hardware solution with sub-millisecond accuracy and localized into a single metric 3D leveraging Project Aria Machine Perception Service (MPS) [8]. We also developed novel algorithms to retarget XSens skeleton motion into a full parametric human model and correct the global drift with optimization. To connect human motion with natural languages, we developed a coarse-to-fine narration schema to describe in-context human motion at different granularity. With 310.5K sentences and 8.64M words from 6545 vocabulary size, Nymeria stands out as the world's largest motion-language dataset.

2 Related Works

Motion datasets – scale, multimodal, in-the-wild and perspectives. Datasets are crucial ingredients in developing algorithms, particularly machine learning approaches. AMASS [69] is a pioneering effort in large motion dataset, which unifies multiple marker-based datasets into SMPL [65]. While AMASS provides diverse human motion, it lacks scene context. Recent works [47, 57, 79] extend the solution to include objects. Monocular [18, 33, 51, 63, 68, 81, 88, 90, 94, 108]

Seq	Qty	Pts	Sce	Loc	Pose	Img	IMU	Gaze	Traj	Sent	Word	Voc
1200	300h	264	20	50	260M	201M	11.7B	10.8M	399Km	310.5K	8.64M	6545

Table 1: Highlight statistics of Nymeria dataset. We capture 1200 sequences of 300-hour daily activity from 264 people performing 20 scenarios at 50 locations with 399Km traveling distance, 260M body poses, 201.2M images, 11.7B IMUs, 10.8M gazes. The motion narrations contain 310.5K sentences in 8.64M words from 6545 vocabulary.

and multi-view cameras [35, 49, 50, 56, 113] are common mocap alternatives. For monocular camera, Motion-X [63] stands out as a comprehensive large collection of whole-body motion with facial expressions and hand gestures. For multi-view setting, EgoExo4D [35] stands out as a large dataset of skilled activities. Vision-based algorithms require good line-of-sight. Consequently, they are better suited to record motion with clear body observations bounded by a volume. To capture data in the wild, mocap suit is a popular candidate [22, 40, 53, 54, 57, 59, 70, 101, 106]. To address dead-reckoning for inertial-based tracking, previous works fuse IMUs with vision [70, 101, 110], optimize motion with 3D scenes [40, 59] and limit the range of motion locally [106]. Simulation as a more scalable solution, offers valuable supplement to real data. Existing works leverage gaming engine for character animation [19], render marker-based mocap with virtual characters and scenes [12, 15, 47], simulate motion with VR [13] or by generative algorithms [60] etc. Simulations often struggle to present noise characteristics of real data, result in domain gaps. While many solutions are developed for third-person views, egocentric motion datasets remain sparse, leaving a gap to the recent advance of egocentric perception [6, 30, 34]. Existing works focus on hands with object-interaction [14, 24, 25, 58, 91], lack ground-truth [34] or parametric body motion [35], or is limited in scale, diversity, and modality [12, 40, 56, 59, 79, 103, 104, 113]. Nymeria is designed to fill the gaps with significant delta to existing datasets (*cf*. Tab. 2).

dataset	q/h	p/M	μ/m	pp	tt/K	voc	pm	hd	3p	wd	gp	od	gz	sr	mp	hh
AMASS [69]	42	0.9	0.22	346			✓									
HPS [40]	4.5	0.5	8.2	7			✓	✓			✓	✓		✓	✓	✓
EgoBody [113]	2	0.4	1	36			✓	✓	✓				✓	✓	✓	✓
HML3D [36]	28.6	2.9	0.12	-	45.0	5371	✓									
EgoHuman [56]	3.5	0.4	0.5	7			✓	✓	✓			✓			✓	✓
MotionX [63]	144	15.6	0.11	-	81.1	-	✓		✓			✓				
DivaTrack [106]	16.5	3.6	0.13	22												
EgoExo4D [35]	88.8	9.6	<u>2.6</u>	<u>740</u>	<u>432</u>	<u>4405</u>		✓	✓			✓	✓	✓		
ParaHome [57]	7.33	56	4.4	30			✓		✓		✓			✓		
LaHuman [22]	3	-	0.51	-	12.3	-	✓		✓		✓	✓		✓	✓	✓
Nymeria(ours)	300	260	15	264	310.5	6545	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 2: Human motion datasets by releasing date. Columns 2 to 5 show activity in hour (q/h), pose frames in millions (p/M), mean sequence duration in minute (μ/m) and number of participants (pp). We then compare language narrations w.r.t. number of descriptions (tt/K) and vocabulary size (voc). The remaining columns mark following features: parametric model for motion representation (pm), egocentric head-mounted device (hd), third-person perspectives (3p), wristbands (wd), global positioning (gp), outdoor scene (od), gaze (gz), 3D scene representations (sr), multi-people scenarios (mp) and human-human interactions (hh). Note EgoExo4D [35] reports 1422h by summing per camera recording time, where the total activity is 180h, with 88.8h annotated with MSCOCO keypoints. The underline numbers are reported for the full dataset.

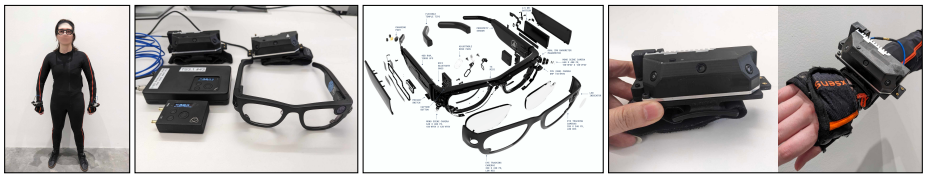
Motion with natural language. Adding language descriptions to motion leads to unique perspective in motion understanding, especially given the powerful capability of large language models (LLMs). Early datasets [17, 32, 38, 41, 43, 93] offer sparse annotations as action categories or semantic labels [84]. KIT [83] is the first attempt in using complete sentences to describe locomotion. HumanML3D [36] enriches HumanAct12 [38] and AMASS [69] with multiple descriptions per motion clip. Motion-X [63] leverages narration algorithm to obtain large-scale fine-grained descriptions at sequence and frame levels. Compared to available text corpus, motion-language data is rather sparse with brief text on brief motion [22]. Ego4D [34] and EgoExo4D [35] offer large-scale descriptions of atomic actions, however, ground-truth motion is missing or without parametric representations. In our work, we provide in-context coarse-to-fine narration by annotators, with the amount of data a magnitude larger than prior works. Since Closely related to tracking is motion synthesis

Egocentric body tracking, synthesis and action recognition. Tracking one's own body motion with wearable devices is challenging. Early approaches leverage dense body-worn IMUs [42,70,71]. Recent methods reduce IMUs to improve practicality [48,107,111,112], assume IMU-enabled AR/VR devices [45,46], combine IMU with cameras [40,103,104,110,118], mobile phones [73] and wristbands [59]. While sparse sensors are more practical, our work centers on constructing an offline dataset to serve "ground truth". Consequently, we use dense IMUs to ensure high accuracy and correct global drift via optimization. Since full-body motion is ill-posed given insufficient observation from egocentric headset, motion synthesis is often used to produce plausible motion. To condition the generation, research explore sparse motion measures [20, 105, 106], headset motion [61] and eye gaze [117] and text [37,82,96]. The success of diffusion models lead to active develop in text-driven motion synthesis [21, 23, 52, 92, 97, 116]. Similarly LLMs inspire novel motion understanding algorithms [31,44,66,115] that tightly couple motion with natural language. Our work constructs a dataset with rich hierarchical narrations to inspire further breakthroughs in the field.

3 Building Nymeria Dataset

3.1 Data collection setup

Hardware. Each participant wears a mocap suit, a pair of glasses, two wristbands, and a synchronization device (*cf*. Fig. 2. XSens MVN Link [7] is adopted for mocap, which is a tight-fit body suit wired with 17 inertial trackers and a magnetometer. MVN Link supports on-device recording, making it ideal to collect in-the-wild data. We use Project Aria glasses [30] as a lightweight headset to record multimodal data. The sensor suite includes 1 RGB camera, 2 grayscale peripheral cameras, 2 ET cameras, 2 IMUs, 1 barometer, 1 magnetometer, 7 microphones, 1 thermometer, GNSS, WiFi and BT. We repackaged the electronics and sensors of Project Aria into a new wristband device called *miniAria*, in order



(a) (b) (c) (d)

Fig. 2: Capture setup. (a) A full-dressed participant. (b) The set of hardwares including Project Aria glasses, two miniAria wristbands and synchronization device. (c) The sensor suite of Project Aria and (d) the miniAria wristband.

to closely resemble current AR/VR headsets and provide data to better constrain body tracking algorithms. The wristband matches Project Aria's sensing ability, with exclusion of microphones, barometer, magnetometer and ET cameras. The supplementary provides detailed sensor configuration and recording profiles.

Synchronization. Project Aria can record an externally provided time signal to aid synchronization. We further enable MVN Link to accept the same signal. A synchronization device is developed to supply the timestamps for all devices, which can optionally receive time from a wireless server located in radio range (∼100m). This facilitates synchronizing multiple devices with sub-millisecond accuracy. The alignment between XSens and Aria is within 1 motion frame *i.e.* 4.2 ms. To capture multiple people simultaneously, we replicate the described setup per participant and leverage a common time server.

Recording protocols. Data collection is managed by 2-3 onsite operators. In addition to participants, a trained observer wearing Project Aria is present to record participants from third-person perspective. All people interact naturally as per activity requires, contributing to rich dynamics as opposed to staged motion. To complete each recording, participants first perform a brief mocap calibration, then gaze calibration, and finally 15-20min activity. We collect 4-8 recordings per person, where a bulk of data is captured at family houses.

Scenarios. We define 20 scenarios (*cf*. examples in Fig. 1 and Fig. 3). For indoor activities, scenarios include cooking, working, entertaining, searching objects etc. For outdoor activities, scenarios include hiking, biking, dining, sports etc. To encourage natural interactions and authentic motions, participants are instructed with high-level guidelines *e.g.* "grab food in the cafeteria and eat on the patio". Operators also prompt in-context actions to increase dynamics.

Privacy considerations. We follow Project Aria research guideline for responsible innovation. Prior to data collection, consents were obtained from participants and home owners for recording and data usage. The SOTA de-identification algorithm EgoBlur [86] is used to blur faces and license plates for all videos.

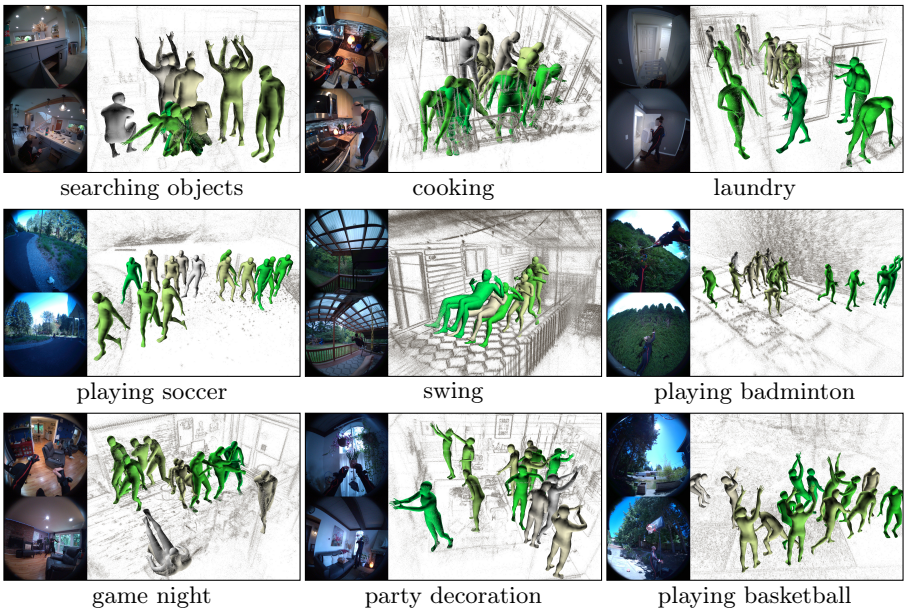


Fig. 3: Diverse scenarios by diverse people. We show different participants performing different indoor/outdoor activities at different locations. In each subfigure, we show an egocentric view on the top left, a third-person view on the bottom left, and motion rendering on the right.

3.2 Data processing

To process data, we first use XSens software to obtain the skeleton motion, and Project Aria MPS [8] for device localization, scene representation and gaze estimation. Then motion is retarget to a parametric human model [4] and registered into the coordinates of Aria devices via optimization.

Full-body mocap and retargetting. We record motion at 240Hz, following the recommended procedures: 1) carefully measuring body dimensions of participants; 2) performing calibration prior to every recording; and 3) processing XSens with the highest quality with single- or multi-floor specification.

XSens represents the body motion as the global transformation and 3D local joint angles of a template skeleton. The skeleton consists of 23 segments, where each segment matches the measured body dimensions of the subject. In addition, a set of $K = 79$ anatomical landmarks are defined on the skeleton model [75]. Their global positions, $\{\mathbf{p}_i\}_K$, can be computed by evaluating the forward kinematics at each frame. We utilize these landmarks to retarget the body motion onto an anatomically-inspired human model for improved realism and visual validation. Our human model is parameterized by $\{\boldsymbol{\theta}, \boldsymbol{\phi}\}$, where the pose parameters $\boldsymbol{\theta}$ define the global transformation and local joint angles,

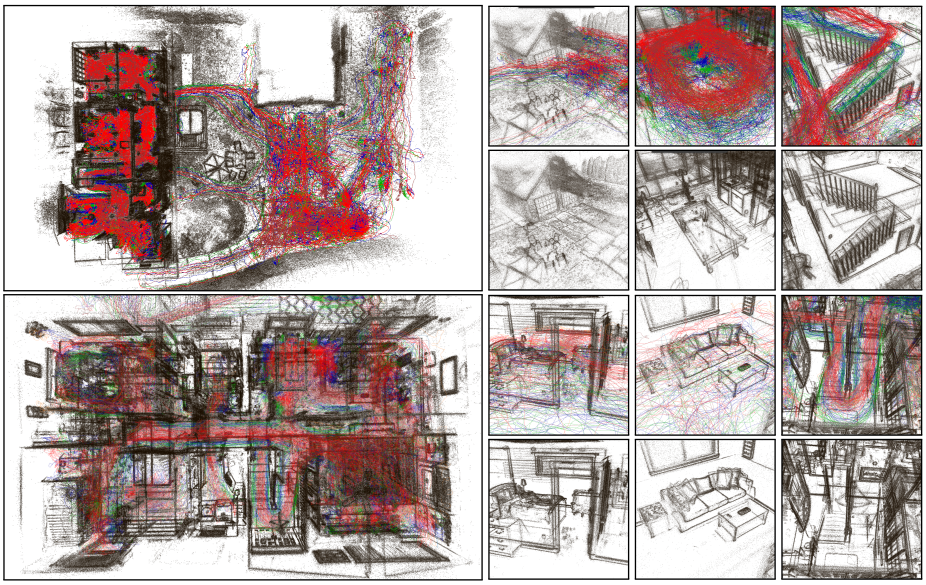


Fig. 4: Global aligned trajectories and point clouds by locations. We show examples of split-level residential house with gardens, where each contain ≈5 hours of recording. The left shows the top-down views of accumulated trajectories where red, green and blue indicate the head, the left and right wrist. On the right we sample closed-up views where the clusters emerge from human 3D motion distribution.

and the shape parameters $\boldsymbol{\phi}$ represent a global body scale and individual bone length. Given a motion of N frames, we solve the following inverse kinematics optimization problem:

$$\underset{\boldsymbol{\phi},\boldsymbol{\theta}_{0\cdots T-1},\mathbf{v}^{0\cdots K-1}}{\arg\min} \sum_{t=0}^{N-1}\sum_{i=0}^{K-1} \left\| T^i(\boldsymbol{\phi},\boldsymbol{\theta}_t)\mathbf{v}^i - \mathbf{p}_t^i \right\|^2 , \tag{1}$$

where $\mathbf{v}^i$ is the local offset of the ith landmark defined on our model, and T^i is the global transformation of its parent joint. We initialize $\mathbf{v}^i$ by manually placing them on the model. The supplementary provides more details about our human model and motion retargetting.

6DoF localization and mapping with global alignment. Data recorded at the same location are globally aligned into a single metric 3D world via Project Aria MPS [8], which employs state-of-the-art visual inertial odometry (VIO), SLAM and mapping algorithms [29, 74, 76]. In a nutshell, first SLAM is run for each individual recording independently. Subsequently, the resulting maps are loop-closed and jointly optimized via visual-inertial bundle adjustment. The output are highly accurate 1KHz trajectories (*cf*. Fig. 4 and supplementary) – allowing for example to visualize head- and wrist-motion-clusters respectively.

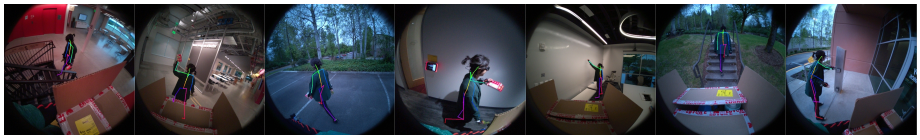


Fig. 5: End-to-end quality assessment. We uniformly sample a 20min recording over 1.5Km moving distance and project skeleton in observer's camera. The rendering and image aligns well, due to precise tracking and synchronization.

Given the device trajectories, we align the body poses into this same reference coordinates by correlating the dead-reckoned trajectory from XSens. Since the latter accumulates significant drift, there exists no static global transformation to align them. Instead, we assume a constant transformation T_{HD} between the user's head segment from XSens H and the Aria device D (Aria is firmly held in place with straps and participants are asked to avoid adjusting the glasses during the recording). We then cut the trajectory into a large number of short 4.2 ms segments, and solve the following optimization problem

$$\arg\min_{T_{HD}} \sum_{t} \left\| \log\left(\left({T_{OH}^{t}}^{-1} T_{OH}^{t+1} \right) \cdot \left(T_{HD} {T_{WD}^{t}}^{-1} T_{WD}^{t+1} {T_{HD}}^{-1} \right)^{-1} \right) \right\|^2, \qquad (2)$$

where O is the drifting odometry frame of XSens and W is the world coordinates of the MPS output. The is a HandEye calibration problem with closed-form solution [95]. Note the formulation effectively aligns a large number of local motion clips by comparing the local velocity. In practice, Aria is not completely rigid during recording, resulting in a main source of inaccuracy. Precision can be improved with a rolling window optimization. Figure 5 provides a qualitative end-to-end assessment of our multi-device location, XSens motion registration, and time synchronization.

3.3 In-context motion-language description

To build the connection between body motion, natural language and activity recognition, we ask human annotators to write textual descriptions of in-context human motion by viewing playback videos of the dataset. The annotators segment the video into clips to write descriptions by answering predefined questions. To give annotators a holistic understanding of the motion, the playback video contains synchronized views of the egocentric video, third-person video, and human motion rendered with 3D scene.

We define three annotation tasks to describe motion coarse to fine, and scale up human efforts in a meaningful way. The finest level is *motion narration* for detailed body posture, *e.g.* motion direction, velocity, interactions and attention. Next, we annotate for *atomic action*. Compared to motion narration, annotators are encouraged to use verbs whenever possible, *e.g.* using "dancing" instead of "swing both arms while rotating the body to the right with legs slightly apart".

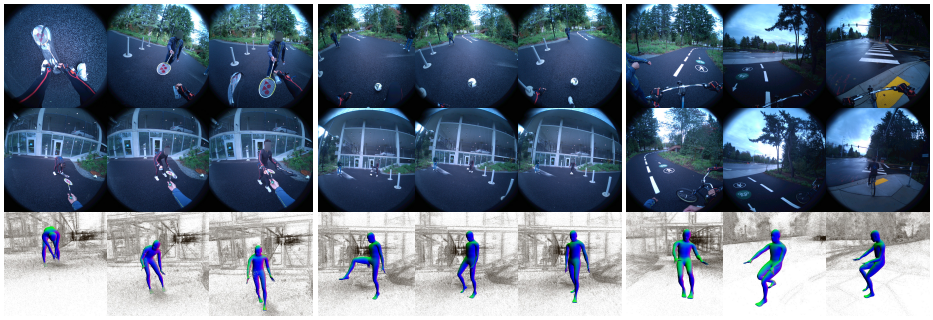


Fig. 6: Hierarchical narration example. (left) motion narration of 5s clip – "C straightens the body to receive a badminton racket from his peer. C extends the right arms towards the bat, with the left arm slightly bent. C takes a step forward with the left foot. C focuses on the racket." (middle) Atomic action of 5s clip – "C lifts the right leg to kick a soccer ball in front of a building." (right) Activity summarization of 30s clip – "C bikes on the road with a peer."

motion narration (39h, 117.2K, 2.72M, 3739) | atomic action (207h, 170.6K, 5.47M, 5129) | summarization (196h, 22.6K, 0.45M, 3168) | combined (230h, 310.5K, 8.64M, 6545)

Fig. 7: Distribution of language descriptions. The word cloud visualizes annotations of each task in separation and with all data combined. The tuple (N, X, Y, Z) means N hours of data is described by X sentences, Y words, and Z vocabulary size.

The two tasks are done for <5s clips. The last task *activity summarization* is to give one-sentence summary over 30s activity. Figure 6 provide examples of each annotation task. The figure shows the benefit of providing annotators three synchronized views. While the egocentric view captures closed-up hand-object interactions, the third-person and motion rendering help annotators grasp a holistic understanding of the actions.

3.4 Statistics

Data statistics. We collected 300 daily activities from 264 participants, which amounts to 1200 sequences with average 15min duration. The accumulated trajectory from all participants is 399Km for headset and 1053Km for both wristbands. Figure 8 shows the participant demographics w.r.t. the self-reported ethnicity, age, height and weight. The statistics is split by gender, where 48.5% participants self-identified as female, and 51.4% as male. The dataset captures 47 houses, where 31 are multi-floor. In total, there are 201 rooms and 45 gardens. We also capture three locations from an open campus, including 1 cafeteria with an outdoor patio, 1 office building, and a public parking connected to multiple

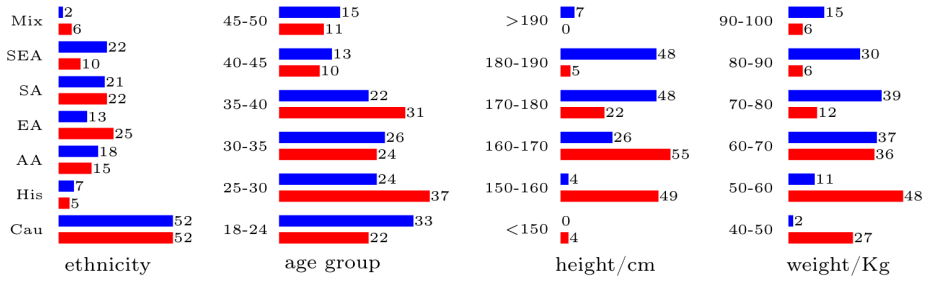


Fig. 8: Demographics by female and male. Meaning of symbols: Caucasian (Cau), Hispanic (His), African American (AA), East/South/Southeast Asian (EA/SA/SEA).

biking/hiking trails. Approximately 15% recordings contain outdoor activities. Among all scenarios, the highest occurrences are cooking, searching objects and improvised actions. We provide detailed breakdown in supplementary.

Annotation statistics. For language descriptions, we annotated 38.6 hours motion narration, 207 hours atomic action and 196 hours activity summarization. The average video segment is 5 seconds for narration and atomic action and 30 seconds for summarization. In total, the dataset provides 310.5K sentences 8.64M words from a vocabulary of 6545 distinctive words. Note the average word per sentence is 27.8, which is significant longer than existing motion-language narrations. Figure 7 visualizes the language distribution.

4 Benchmark Tasks and Baselines

4.1 Research opportunities

With an enormous amount of contextualized human motion, Nymeria dataste provides unprecedented research opportunities. Following, we highlight a few research domains. This is by no means complete, but to inspire novel directions and ideas by scratching the surface of full potential.

Motion tasks. Nymeria is created to assist human motion understanding, with an emphasis on, but not limited to, egocentric perception. The data supports various topics, *e.g.* full-body tracking, motion synthesis, motion forecasting, path planning, action recognition, human behavior analysis and etc. With a multimodal multi-device dataset, we encourage exploring unique algorithms with novel settings, *e.g.* gaze-conditioned motion prediction, action recognition from headset and wristbands, interaction generation from language etc.

Multimodal spatial reasoning and video understanding. In addition to body motion, Nymeria also provides extensive egocentric videos with precise localization and synchronization. Nymeria therefore is a great asset for algorithms requiring

	MPJPE (cm)			Hand	MPJVE
	Mean	Lower	Upper	PE(cm)	(cm/s)
AMASS	4.20	8.06	1.88	2.34	28.23
Real	7.97	16.74	3.13	6.25	16.71
Synthetic	7.31	15.97	2.51	3.47	16.63

Table 3: AvatarPoser [46] train/test with real vs. synthetic poses from Nymeria. The first row reports AvatarPoser train/test on AMASS [69] for reference.

	MPJPE (cm)			FID
	Mean	Lower	Upper	
BoDiffusion(A)	3.63	7.07	1.53	-
BoDiffusion	7.98	15.27	5.28	2.32
EgoEgo	13.22	19.03	10.00	5.14

Table 4: Motion synthesis with EgoEgo [61] and BoDiffusion [20] on Nymeria. BoDiffusion(A) reports results train/test on AMASS [69].

camera poses as priori, *e.g.* scene reconstruction [55, 72]. In this regard, the rich dynamics raise novel real-world challenges. Nymeria also facilitates video understanding with numerous in-context narration. Image retrieval and relocalization also benefit from Nymeria , sincell we capture multiple videos per location and align them in global coordinates.

Simulation. Simulation is a promising technique to gather massive data, however, synthetic data is typically seed in real world [15, 47]. By design, Nymeria is naturally useful to drive in-context character animation. Since human motion is a function of the environment, we believe Nymeria also aid simulating 3D scene [109]. Sensor simulation can benefit from our data as well, especially for IMUs, magnetormeter, and barometer in combination with motion priors.

4.2 Motion tasks baselines

We use three case studies to showcase motion algorithms on Nymeria. The goal is to validate the data and provide future research with common baselines. Due to page limit, details of model training is present in supplementary.

Motion tracking and synthesis from sparse inputs. In this case study, we take the 1-point/3-point body tracking problem in modern AR/VR, where people wear a headset and optionally controllers in VR, and a pair of glasses and optionally wristbands in AR. The task is to recover wearer's full-body motion using only headset (1-point) [61, 67] or additional wrist devices (3-point) [20, 28, 45, 46, 106]. The problem is considered mixture of tracking and synthesis, due to insufficient lower-body observation. Existing works simulate sensor inputs from motion datasets, which lack the noise characteristic of real data. Nymeria is the first large dataset with multimodal real data to support model training and evaluation. In particular, we provide raw IMU and device poses, as opposed to filtered [89] acceleration and velocity released in previous datasets [40, 101, 106]. In the first experiment, we adapt the 3-point regression method AvatarPoser [46] to train on Nymeria with two variants – one model trained with "real" device poses from SLAM and the other trained with "synthetic" input mocked-up from body motion same as in the original work. The evaluation uses the same metric

as AvatarPoser, including mean per joint position error (MPJPE), hand position error and mean per joint velocity error (MPJVE). Results are shown in Tab. 3. As expected, "real" data lead to worse performance due to additional error-sources. However the gap is small, which indicates the quality of device tracking. Compare to the original AvatarPoser trained on AMASS [69], Nymeria yields reasonable but higher MPJPE, mainly from worse lower-body tracking. We hypothesize our data contain harder motion, *e.g.* hiking uneven terrain, taking stairs, playing sports and etc. The second experiment evaluates two diffusion models for motion synthesis, *i.e.* 3-point Bodiffusion [20] and 1-point EgoEgo [61]. In addition to MPJPE, we report the Frechet Inception Distance (FID) which measures the distribution distance between generated motion and real motion, following the same procedure in [39] (*cf.* Tab. 4). Results trained from our data are comparable with the original works trained on subset of AMASS, where 3-point diffusion yields better performance as expected.

Representing human motion manifold. Learning the embedding space to model human motion manifold has many benefits, *e.g.* dimension reduction, learning motion priors, motion denoising by projection, motion synthesis by sampling and interpolation etc. Previous attempts mainly rely on AMASS [69] to learn the representation [64, 81, 87, 98], where the data heavily focused on isolated locomotion or professional motion. By capturing rich daily activities of common people interacting with real world, our data distribution can better represent the motion manifold concerning everyday human activities. To take a stab in this direction, we to train Vector-Quantized Variational Autoencoder (VQ-VAE) [27, 77] for Nymeria motion data, following the previous work [44]. VQ-VAE can be used as a "motion tokenizer" to generate motion with auto-regression [66], and to de-noise motion by projecting the input onto the manifold [87]. An ablation is perform to study the impact of product quantization, codebook size and latent dimension. For evaluation, we adopt the same metrics in [44] and include the VQ-VAE trained on AMASS for comparison. As shown in Tab. 5, Nymeria motion data can be well tokenized to achieve similar performance as VQ-VAE trained with AMASS data. By leveraging product quantization, increasing the codebook size and decreasing the latent dimension, the reconstruction quality is further improved. The resulting motion tokenizer can therefore be used with LLMs akin to language tokenizer to foster motion understanding [44, 115] as detailed in the next case study.

PQ	CB	Dim	MPJPE	PA-MPJPE	ACC
*	512	-	55.80	40.10	7.50
1	2048	512	51.60	37.55	1.09
2	2048	512	39.63	29.77	0.71
2	4096	512	39.20	29.66	0.82
2	16384	64	**34.49**	**26.83**	**0.67**

Table 5: Ablation of motion VQ-VAE trained on Nymeria (metric unit: mm). We compare product quantization (PQ), codebook (CB) size and latent dimensions (Dim). The first row show results of AMASS as reported in [44]. PA stands for Procrustes-aligned and ACC for joint position acceleration.

	Bert	Bleu@1	Bleu@4	CIDEr	RougeL
TM2T	11.08	40.11	8.99	20.85	30.70
MotionGPT	**14.09**	**42.22**	**10.31**	**37.27**	**32.33**

Table 6: Evaluation of motion-to-text. Models are trained with small subset of Nymeria.

Motion and language. While parametric human motion is useful for machine algorithms, natural language description is a better interface with human. An valuable feature of Nymeria dataset is the high-quality hierarchical narrations. Compared with existing data [36, 63], our narrations are in longer natural sentences with context descriptions of objects and environments. It can be used to learn models for text-driven motion generation and motion-to-text descriptions. More importantly, the contextual descriptions are not only paired with the human motion, but also with videos, point clouds, and other sensory data and annotations. The corroboration of both 2D and 3D environment information with language offers exciting opportunities in grounding language and motion research in the physical world. Leveraging VQ-VAE experiment, we train MotionGPT [44] and TM2T [37] for the motion-to-text task with a subset of 30h motion narrations from Nymeria (*cf*. Tab. 6) to be directly comparable with previous results, using the same metrics of BERT [114], BLEU [80], CIDEr [102] and ROUGE-L [62]. TM2T performs worse than MotionGPT since it lacks strong language prior with the T5 [85] backbone. Compare to the original works trained with HumanML3D [36] and KIT [83], our results are worse as expected. Given similar hours of data, our narrations are much more complex and diverse. By using the full narration data, we expect the performance to be significantly better.

5 Conclusions and Discussions

We propose Nymeria dataset to accelerate research in egocentric motion understanding. The dataset is the world's largest collection of human motion in the wild with 300 hours daily activity, 260M body poses of 264 participants across 50 locations. We provide accurate 6DoF tracking, 3D scene points and gaze, with all modalities synchronized and aligned into one metric 3D world. Collectively, the dataset captured 399Km of travel by the participants for a total of 201.2M egocentric images, 11.7B IMU samples and 10.8M gaze point. The Nymeria dataset also stands out as the world largest motion-language dataset with 310.5K sentences in 8.64M words with 38.6 hours of fine-grained motion narration, 207 hours atomic actions and 196 hours activity summarization.

Limitations. The mocap suit and wristbands lead to unnatural appearance in videos, and restrict certain range of motion. XSens quality is known to be affected by motion calibration and body measurements. Our dataset only covers a portion of daily activities, leaving out common public scenarios.

Social impact. Understanding egocentric full-body motion is crucial towards contextual AI, however it heavily involves personal data. We make best effort to respect privacy via consent, de-identification, minimum data retention and permissive research license.

Acknowledgements

We gratefully acknowledge the following colleagues for their valuable discussions and technical support. Genesis Mendoza, Jacob Alibadi, Ivan Soeria-Atmadja, Elena Shchetinina, and Atishi Bali worked on data collection. Yusuf Mansour supported gaze estimation on Project Aria. Ahmed Elabbasy, Guru Somasundaram, Omkar Pakhi, and Nikhil Raina supported EgoBlur as the solution to anonymize video and explored bounding box annotation. Evgeniy Oleinik, Maien Hamed, and Mark Schwesinger supported onboarding Nymeria dataset into Project Aria dataset explorer and data release. Melissa Hebra helped with coordinating narration annotation. Edward Miller served as research program manager. Pierre Moulon provided valuable guidance to open source code repository. Tassos Mourikis, Maurizio Monge, David Caruso, Duncan Frost, and Harry Lanaras provided technical support for SLAM. Daniel DeTone, Dan Barnes, Raul Mur Artal, Thomas Whelan, and Austin Kukay provided valuable discussions on annotating semantic bounding box. Julian Nubert adopted the dataset for early dogfooding. Pedro Cancel Rivera, Gustavo Solaira, Yang Lou, and Yuyang Zou provided support from Project Aria program. Svetoslav Kolev provided frequent feedback. Arjang Talattof supported MPS. Gerard Pons-Moll served as senior advisor. Carl Ren and Mingfei Yan served as senior managers.

Contribution Statements. Lingni Ma led the project, developed the pipeline to construct the dataset, coordinated data collection/processing, and led the baseline evaluations. Yuting Ye developed the solution for human motion retargetting, advised data collection and evaluations. Fangzhou Hong, Vladimir Guzov and Yifeng Jiang, validated the dataset and implemented baselines. Rowan Postyeni supported daily operations, processed XSens motion data and performed quality assessment for XSens motion and narration. Luis Pesqueria served the program manager for data collection and narration. Alexander Gamino was responsible for multi-device tracking. Vijay Baiyya was responsible to Project Aria MPS scaled processing. Hyo Jin Kim supported narration annotations. Kevin Bailey and David Soriano Fosas lead hardware development of miniAria wristband. C. Karen Liu and Ziwei Liu served as technical advisor to data collection, annotation and baseline evaluations. Jakob Engel, Renzo De Nardi and Richard Newcombe were the senior technical and scientific advisors.

References

1. Apple Vision Pro, `https://www.apple.com/apple-vision-pro/`
2. HTC VIVE, `vive.com`
3. Magic Leap 2, `https://www.magicleap.com/magic-leap-2`
4. Meta momentum library, `https://github.com/facebookincubator/momentum/`
5. Meta Quest, `https://www.meta.com/quest/`
6. Microsoft HoloLens, `https://learn.microsoft.com/en-us/hololens/`
7. Movella XSens MVN Link motion capture, `https://www.movella.com/products/motion-capture/xsens-mvn-link`

8. Project Aria Machine Perception Services, https://facebookresearch.github.io/projectaria_tools/docs/ARK/mps
9. Ray-Ban Meta smart glasses, https://www.meta.com/smart-glasses/
10. Rokoko, https://www.rokoko.com/
11. Vuzix smart glasses, https://www.vuzix.com/pages/smart-glasses
12. Akada, H., Wang, J., Shimada, S., Takahashi, M., Theobalt, C., Golyanik, V.: UnrealEgo: A new dataset for robust egocentric 3d human motion capture. In: European Conference on Computer Vision (ECCV) (2022)
13. Araujo, J.P., Li, J., Vetrivel, K., Agarwal, R., Gopinath, D., Wu, J., Clegg, A., Liu, C.K.: Circle: Capture in rich contextual environments. CVPR (2023)
14. Banerjee, P., Shkodrani, S., Moulon, P., Hampali, S., Zhang, F., Fountain, J., Miller, E., Basol, S., Newcombe, R., Wang, R., Engel, J.J., Hodan, T.: Introducing hot3d: An egocentric dataset for 3d hand and object tracking (2024)
15. Black, M.J., Patel, P., Tesch, J., Yang, J.: BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In: Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (2023)
16. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems. vol. 33, pp. 1877–1901 (2020)
17. Cai, Z., Ren, D., Zeng, A., Lin, Z., Yu, T., Wang, W., Fan, X., Gao, Y., Yu, Y., Pan, L., et al.: Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In: European Conference on Computer Vision (2022)
18. Cai, Z., Yin, W., Zeng, A., Wei, C., Sun, Q., Wang, Y., Pang, H.E., Mei, H., Zhang, M., Zhang, L., Loy, C.C., Yang, L., Liu, Z.: Smpler-x: Scaling up expressive human pose and shape estimation (2023)
19. Cai, Z., Zhang, M., Ren, J., Wei, C., Ren, D., Lin, Z., Zhao, H., Yang, L., Loy, C.C., Liu, Z.: Playing for 3d human recovery. arXiv preprint arXiv:2110.07588 (2021)
20. Castillo, A., Escobar, M., Jeanneret, G., Pumarola, A., Arbeláez, P., Thabet, A., Sanakoyeu, A.: BoDiffusion: Diffusing sparse observations for full-body human motion synthesis. ICCV (2023)
21. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: CVPR (2023)
22. Cong, P., Wang, Z., Dou, Z., Ren, Y., Yin, W., Cheng, K., Sun, Y., Long, X., Zhu, X., Ma, Y.: Laserhuman: Language-guided scene-aware human motion generation in free environment (2024)
23. Dabral, R., Mughal, M.H., Golyanik, V., Theobalt, C.: Mofusion: A framework for denoising-diffusion-based motion synthesis. In: Computer Vision and Pattern Recognition (CVPR) (2023)
24. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. In: ECCV (2018)
25. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: The epic-kitchens dataset: Collection, challenges and baselines. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **43**(11), 4125–4141 (2021)

26. Delmas, Ginger and Weinzaepfel, Philippe and Lucas, Thomas and Moreno-Noguer, Francesc and Rogez, Grégory: PoseScript: 3D Human Poses from Natural Language. In: ECCV (2022)
27. Dhariwal, P., Jun, H., Payne, C., Kim, J.W., Radford, A., Sutskever, I.: Jukebox: A generative model for music. arXiv preprint arXiv:2005.00341 (2020)
28. Du, Y., Kips, R., Pumarola, A., Starke, S., Thabet, A., Sanakoyeu, A.: Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 481–490 (2023)
29. Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry (2016)
30. Engel, J., Somasundaram, K., Goesele, M., Sun, A., Gamino, A., Turner, A., Talattof, A., Yuan, A., Souti, B., Meredith, B., Peng, C., Sweeney, C., Wilson, C., Barnes, D., DeTone, D., Caruso, D., Valleroy, D., Ginjupalli, D., Frost, D., Miller, E., Mueggler, E., Oleinik, E., Zhang, F., Somasundaram, G., Solaira, G., Lanaras, H., Howard-Jenkins, H., Tang, H., Kim, H.J., Rivera, J., Luo, J., Dong, J., Straub, J., Bailey, K., Eckenhoff, K., Ma, L., Pesqueira, L., Schwesinger, M., Monge, M., Yang, N., Charron, N., Raina, N., Parkhi, O., Borschowa, P., Moulon, P., Gupta, P., Mur-Artal, R., Pennington, R., Kulkarni, S., Miglani, S., Gondi, S., Solanki, S., Diener, S., Cheng, S., Green, S., Saarinen, S., Patra, S., Mourikis, T., Whelan, T., Singh, T., Balntas, V., Baiyya, V., Dreewes, W., Pan, X., Lou, Y., Zhao, Y., Mansour, Y., Zou, Y., Lv, Z., Wang, Z., Yan, M., Ren, C., Nardi, R.D., Newcombe, R.: Project Aria: A new tool for egocentric multi-modal AI research (2023)
31. Feng, Y., Lin, J., Dwivedi, S.K., Sun, Y., Patel, P., Black, M.J.: Chatpose: Chatting about 3d human pose. In: CVPR (2024)
32. Ghorbani, S., Mahdaviani, K., Thaler, A., Kording, K., Cook, D.J., Blohm, G., Troje, N.F.: Movi: A large multi-purpose human motion and video dataset. Plos one (2021)
33. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa*, A., Malik*, J.: Humans in 4D: Reconstructing and tracking humans with transformers. In: International Conference on Computer Vision (ICCV) (2023)
34. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., Martin, M., Nagarajan, T., Radosavovic, I., Ramakrishnan, S.K., Ryan, F., Sharma, J., Wray, M., Xu, M., Xu, E.Z., Zhao, C., Bansal, S., Batra, D., Cartillier, V., Crane, S., Do, T., Doulaty, M., Erapalli, A., Feichtenhofer, C., Fragomeni, A., Fu, Q., Gebreselasie, A., González, C., Hillis, J., Huang, X., Huang, Y., Jia, W., Khoo, W., Kolář, J., Kottur, S., Kumar, A., Landini, F., Li, C., Li, Y., Li, Z., Mangalam, K., Modhugu, R., Munro, J., Murrell, T., Nishiyasu, T., Price, W., Ruiz, P., Ramazanova, M., Sari, L., Somasundaram, K., Southerland, A., Sugano, Y., Tao, R., Vo, M., Wang, Y., Wu, X., Yagi, T., Zhao, Z., Zhu, Y., Arbeláez, P., Crandall, D., Damen, D., Farinella, G.M., Fuegen, C., Ghanem, B., Ithapu, V.K., Jawahar, C.V., Joo, H., Kitani, K., Li, H., Newcombe, R., Oliva, A., Park, H.S., Rehg, J.M., Sato, Y., Shi, J., Shou, M.Z., Torralba, A., Torresani, L., Yan, M., Malik, J.: Ego4D: Around the world in 3,000 hours of egocentric video. In: CVPR. pp. 18995–19012 (June 2022)
35. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., Byrne, E., Chavis, Z., Chen, J., Cheng, F., Chu, F.J., Crane, S., Dasgupta, A., Dong, J., Escobar, M., Forigua, C., Gebreselasie, A., Haresh, S., Huang, J., Islam, M.M., Jain, S., Khirodkar, R., Kukreja, D., Liang, K.J., Liu, J.W., Majumder, S., Mao, Y., Martin, M.,

Mavroudi, E., Nagarajan, T., Ragusa, F., Ramakrishnan, S.K., Seminara, L., Somayazulu, A., Song, Y., Su, S., Xue, Z., Zhang, E., Zhang, J., Castillo, A., Chen, C., Fu, X., Furuta, R., Gonzalez, C., Gupta, P., Hu, J., Huang, Y., Huang, Y., Khoo, W., Kumar, A., Kuo, R., Lakhavani, S., Liu, M., Luo, M., Luo, Z., Meredith, B., Miller, A., Oguntola, O., Pan, X., Peng, P., Pramanick, S., Ramazanova, M., Ryan, F., Shan, W., Somasundaram, K., Song, C., Southerland, A., Tateno, M., Wang, H., Wang, Y., Yagi, T., Yan, M., Yang, X., Yu, Z., Zha, S.C., Zhao, C., Zhao, Z., Zhu, Z., Zhuo, J., Arbelaez, P., Bertasius, G., Crandall, D., Damen, D., Engel, J., Farinella, G.M., Furnari, A., Ghanem, B., Hoffman, J., Jawahar, C.V., Newcombe, R., Park, H.S., Rehg, J.M., Sato, Y., Savva, M., Shi, J., Shou, M.Z., Wray, M.: Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In: CVPR (2024)

36. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5152–5161 (June 2022)
37. Guo, C., Zuo, X., Wang, S., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: European Conference on Computer Vision. pp. 580–597. Springer (2022)
38. Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2021–2029 (2020)
39. Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2021–2029 (2020)
40. Guzov, V., Mir, A., Sattler, T., Pons-Moll, G.: Human poseitioning system (HPS): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
41. Harvey, F.G., Yurick, M., Nowrouzezahrai, D., Pal, C.: Robust motion inbetweening. ACM Transactions on Graphics (TOG) **39**(4), 60–1 (2020)
42. Huang, Y., Kaufmann, M., Aksan, E., Black, M.J., Hilliges, O., Pons-Moll, G.: Deep inertial poser: learning to reconstruct human pose from sparse inertial measurements in real time. ACM Transactions on Graphics (TOG) **37**(6), 1–15 (2018)
43. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE transactions on pattern analysis and machine intelligence **36**(7) (2013)
44. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. In: Advances in Neural Information Processing Systems (2024)
45. Jiang, J., Streli, P., Meier, M., Fender, A., Holz, C.: Egoposer: Robust real-time ego-body pose estimation in large scenes. arXiv preprint arXiv:2308.06493 (2023)
46. Jiang, J., Streli, P., Qiu, H., Fender, A., Laich, L., Snape, P., Holz, C.: Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In: European Conference on Computer Vision. pp. 443–460. Springer (2022)
47. Jiang, N., Zhang, Z., Li, H., Ma, X., Wang, Z., Chen, Y., Liu, T., Zhu, Y., Huang, S.: Scaling up dynamic human-scene interaction modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
48. Jiang, Y., Ye, Y., Gopinath, D., Won, J., Winkler, A.W., Liu, C.K.: Transformer inertial poser: Real-time human motion reconstruction from sparse imus with

simultaneous terrain generation. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022)
49. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: The IEEE International Conference on Computer Vision (ICCV) (2015)
50. Joo, H., Simon, T., Li, X., Liu, H., Tan, L., Gui, L., Banerjee, S., Godisart, T.S., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social interaction capture. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
51. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5614–5623 (2019)
52. Karunratanakul, K., Preechakul, K., Suwajanakorn, S., Tang, S.: Guided motion diffusion for controllable human motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2151–2162 (2023)
53. Kaufmann, M., Song, J., Guo, C., Shen, K., Jiang, T., Tang, C., Zárate, J.J., Hilliges, O.: EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild. In: International Conference on Computer Vision (ICCV) (2023)
54. Kaufmann, M., Zhao, Y., Tang, C., Tao, L., Twigg, C., Song, J., Wang, R., Hilliges, O.: Em-pose: 3d human pose estimation from sparse electromagnetic trackers. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2021)
55. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023), `https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/`
56. Khirodkar, R., Bansal, A., Ma, L., Newcombe, R., Vo, M., Kitani, K.: EgoHumans: An egocentric 3d multi-human benchmark. In: ICCV (2023)
57. Kim, J., Kim, J., Na, J., Joo, H.: Parahome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions (2024)
58. Kwon, T., Tekin, B., Stühmer, J., Bogo, F., Pollefeys, M.: H2o: Two hands manipulating objects for first person interaction recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10138–10148 (October 2021)
59. Lee, J., Joo, H.: Mocap everyone everywhere: Lightweight motion capture with smartwatches and a head-mounted camera. arXiv preprint arXiv:2401.00847 (2024)
60. Li, G., Zhao, K., Zhang, S., Lyu, X., Dusmanu, M., Zhang, Y., Pollefeys, M., Tang, S.: Egogen: An egocentric synthetic data generator (2024)
61. Li, J., Liu, K., Wu, J.: Ego-body pose estimation via ego-head pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17142–17151 (2023)
62. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
63. Lin, J., Zeng, A., Lu, S., Cai, Y., Zhang, R., Wang, H., Zhang, L.: Motion-x: A large-scale 3d expressive whole-body human motion dataset. Advances in Neural Information Processing Systems (2023)
64. Ling, H.Y., Zinno, F., Cheng, G., van de Panne, M.: Character controllers using motion vaes. ACM Trans. Graph. **39**(4) (2020)

65. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015)
66. Lucas, T., Baradel, F., Weinzaepfel, P., Rogez, G.: Posegpt: Quantization-based 3d human motion generation and forecasting. In: European Conference on Computer Vision. pp. 417–435. Springer (2022)
67. Luo, Z., Hachiuma, R., Yuan, Y., Kitani, K.: Dynamics-regulated kinematic policy for egocentric pose estimation. In: Neural Information Processing Systems (2021)
68. Luvizon, D., Habermann, M., Golyanik, V., Kortylewski, A., Theobalt, C.: Scene-Aware 3D Multi-Human Motion Capture from a Single Camera. Computer Graphics Forum **42**(2), 371–383 (2023)
69. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: International Conference on Computer Vision. pp. 5442–5451 (Oct 2019)
70. von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
71. Marcard, T., Rosenhahn, B., Black, M., Pons-Moll, G.: Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. Computer Graphics Forum 36(2), Proceedings of the 38th Annual Conference of the European Association for Computer Graphics (Eurographics), 2017 **36** (02 2017)
72. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
73. Mollyn, V., Arakawa, R., Goel, M., Harrison, C., Ahuja, K.: 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. CHI '23, Association for Computing Machinery, New York, NY, USA (2023)
74. Mourikis, A.I., Roumeliotis, S.I.: A multi-state constraint kalman filter for vision-aided inertial navigation. In: Proceedings 2007 IEEE international conference on robotics and automation. pp. 3565–3572. IEEE (2007)
75. Movella: MVN User Manual, `https://www.movella.com/hubfs/MVN_User_Manual.pdf`
76. Mur-Artal, Raúl, M.J.M.M., Tardós, J.D.: ORB-SLAM: a versatile and accurate monocular SLAM system. IEEE Transactions on Robotics **31**(5), 1147–1163 (2015)
77. van den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (2017)
78. OpenAI, :, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H.W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus,

L., Felix, N., Fishman, S.P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S.S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N.S., Khan, T., Kilpatrick, L., Kim, J.W., Kim, C., Kim, Y., Kirchner, H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C.M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S.M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H.P., Michael, Pokorny, Pokrass, M., Pong, V., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F.P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J.F.C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J.J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., Zoph, B.: Gpt-4 technical report (2023)

79. Pan, X., Charron, N., Yang, Y., Peters, S., Whelan, T., Kong, C., Parkhi, O., Newcombe, R., Ren, C.Y.: Aria digital twin: A new benchmark dataset for egocentric 3d machine perception (2023)
80. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
81. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 10975–10985 (2019)
82. Petrovich, M., Black, M.J., Varol, G.: Temos: Generating diverse human motions from textual descriptions. In: ECCV. pp. 480–497. Springer Nature Switzerland, Cham (2022)
83. Plappert, M., Mandery, C., Asfour, T.: The KIT motion-language dataset. Big Data **4**(4), 236–252 (dec 2016)
84. Punnakkal, A.R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., Black, M.J.: BABEL: Bodies, action and behavior with english labels. In: Proceed-

ings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 722–731 (Jun 2021)
85. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research **21**(1), 5485–5551 (2020)
86. Raina, N., Somasundaram, G., Zheng, K., Saarinen, S., Messiner, J., Schwesinger, M., Pesqueira, L., Prasad, I., Miller, E., Gupta, P., Yan, M., Newcombe, R.A., Ren, C.Y., Parkhi, O.M.: Egoblur: Responsible innovation in aria. ArXiv **abs/2308.13093** (2023)
87. Rempe, D., Birdal, T., Hertzmann, A., Yang, J., Sridhar, S., Guibas, L.J.: Humor: 3d human motion model for robust pose estimation. In: International Conference on Computer Vision (ICCV) (2021)
88. Riza Alp Gueler, Natalia Neverova, I.K.: Densepose: Dense human pose estimation in the wild. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
89. Roetenberg, D., Luinge, H., Slycke, P.: Xsens mvn: Full 6dof human motion tracking using miniature inertial sensors. Xsens Motion Technol. BV Tech. Rep. **3** (01 2009)
90. Rong, Y., Shiratori, T., Joo, H.: Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1749–1759 (2021)
91. Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., Yao, A.: Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In: CVPR (2022)
92. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. ICLR (2023)
93. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1010–1019 (2016)
94. Shimada, S., Golyanik, V., Xu, W., Theobalt, C.: Physcap: Physically plausible monocular 3d motion capture in real time. ACM Transactions on Graphics (ToG) **39**(6), 1–16 (2020)
95. Sorkine-Hornung, O., Rabinovich, M.: Least-squares rigid motion using svd. Computing **1**(1), 1–5 (2017)
96. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: Motionclip: Exposing human motion generation to clip space. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII. pp. 358–374. Springer (2022)
97. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: ICLR (2023)
98. Tiwari, G., Antic, D., Lenssen, J.E., Sarafianos, N., Tung, T., Pons-Moll, G.: Pose-ndf: Modeling human pose manifolds with neural distance fields. In: European Conference on Computer Vision (ECCV) (October 2022)
99. Tome, D., Alldieck, T., Peluse, P., Pons-Moll, G., Agapito, L., Badino, H., de la Torre, F.: Selfpose: 3d egocentric pose estimation from a headset mounted camera. IEEE Transactions on Pattern Analysis and Machine Intelligence (Oct 2020)
100. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models (2023)

101. Trumble, M., Gilbert, A., Malleson, C., Hilton, A., Collomosse, J.: Total capture: 3d human pose estimation fusing video and inertial sensors. In: Proceedings of 28th British Machine Vision Conference. pp. 1–13 (2017)
102. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
103. Wang, J., Liu, L., Xu, W., Sarkar, K., Luvizon, D., Theobalt, C.: Estimating egocentric 3d human pose in the wild with external weak supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13157–13166 (June 2022)
104. Wang, J., Luvizon, D., Xu, W., Liu, L., Sarkar, K., Theobalt, C.: Scene-aware egocentric 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13031–13040 (2023)
105. Wouwe, T., Lee, S., Falisse, A., Delp, S., Liu, C.: Diffusion inertial poser: Human motion reconstruction from arbitrary sparse imu configurations. In: CVPR (2024)
106. Yang, D., Kang, J., Ma, L., Greer, J., Ye, Y., Lee, S.H.: Divatrack: Diverse bodies and motions from acceleration-enhanced three-point trackers. EuroGraphics (2024)
107. Yang, D., Kim, D., Lee, S.H.: Lobstr: Real-time lower-body pose prediction from sparse upper-body tracking signals. In: Computer Graphics Forum. vol. 40, pp. 265–275. Wiley Online Library (2021)
108. Ye, V., Pavlakos, G., Malik, J., Kanazawa, A.: Decoupling human and camera motion from videos in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
109. Yi, H., Huang, C.H.P., Tripathi, S., Hering, L., Thies, J., Black, M.J.: MIME: Human-aware 3D scene generation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12965–12976 (June 2023)
110. Yi, X., Zhou, Y., Habermann, M., Golyanik, V., Pan, S., Theobalt, C., Xu, F.: Egolocate: Real-time motion capture, localization, and mapping with sparse body-mounted sensors. ACM Transactions on Graphics (TOG) **42**(4) (2023)
111. Yi, X., Zhou, Y., Habermann, M., Shimada, S., Golyanik, V., Theobalt, C., Xu, F.: 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. pp. 13167–13178 (2022)
112. Yi, X., Zhou, Y., Xu, F.: Transpose: real-time 3d human translation and pose estimation with six inertial sensors. ACM Transactions on Graphics (TOG) **40**(4), 1–13 (2021)
113. Zhang, S., Ma, Q., Zhang, Y., Qian, Z., Kwon, T., Pollefeys, M., Bogo, F., Tang, S.: EgoBody: Human body shape and motion of interacting people from head-mounted devices. In: European conference on computer vision (ECCV) (Oct 2022)
114. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)
115. Zhang, Y., Huang, D., Liu, B., Tang, S., Lu, Y., Chen, L., Bai, L., Chu, Q., Yu, N., Ouyang, W.: Motiongpt: Finetuned llms are general-purpose motion generators. arXiv preprint arXiv:2306.10900 (2023)
116. Zhang, Z., Liu, R., Aberman, K., Hanocka, R.: Tedi: Temporally-entangled diffusion for long-term motion synthesis (2023)
117. Zheng, Y., Yang, Y., Mo, K., Li, J., Yu, T., Liu, Y., Liu, K., Guibas, L.J.: Gimo: Gaze-informed human motion prediction in context. ECCV (2022)

118. Zheng, Z., Yu, T., Li, H., Guo, K., Dai, Q., Fang, L., Liu, Y.: Hybridfusion: Real-time performance capture using a single depth sensor and sparse imus. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)