# A Unified Inertial Encoder for Human and Humanoid Motion

unified-inertial-encoder.github.io

**Yutian Chen**[1], **Shibo Zhao**[1], **Rocky Duan**[1], **Karen Liu**[1,3], **Carmelo Sferrazza**[1], **Guanya Shi**[1,4], **Yuheng Qiu**[1] **and Koushil Sreenath**[1,2]

[1]Amazon FAR, [2]UC Berkeley, [3]Stanford University, [4]Carnegie Mellon University

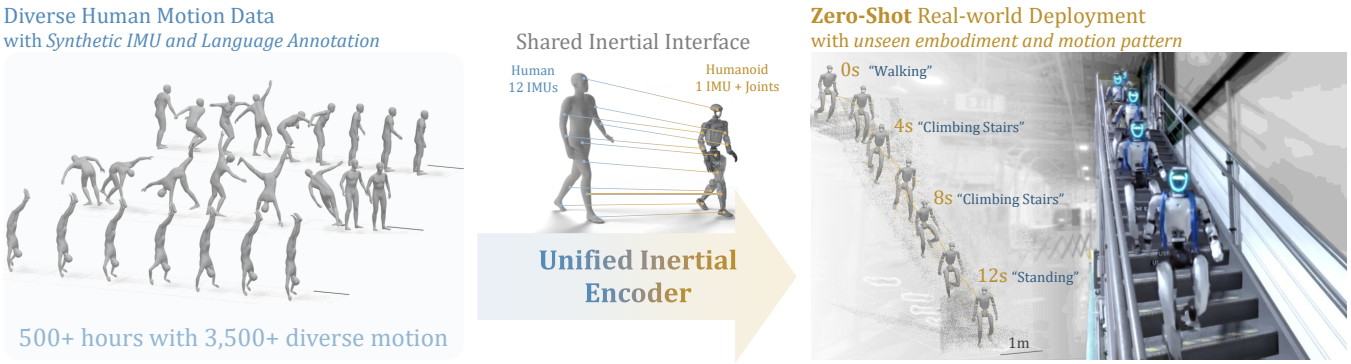



Fig. 1. We present a unified inertial encoder for human and humanoid motion. The encoder is trained on diverse human and IK-retargeted G1 motions (left), and uses a shared inertial interface to connect corresponding body sites across embodiments (center). Our encoder transfers to a physical G1 humanoid (right) for the inertial odometry and motion-description retrieval tasks without requiring physics simulation, real-robot trajectories, or fine-tuning.

***Abstract*— Learning humanoid state estimation and language-grounded motion understanding from robot recordings ties training coverage to the robot's motion repertoire. Human motion datasets offer broader coverage, yet transferring this knowledge requires overcoming differences in body geometry, sensing, and physical disturbances. To enable this transfer, we propose a unified inertial encoder whose shared inertial interface uses kinematic IMU synthesis to represent human and humanoid motion at corresponding body sites. Within this common input space, joint training on human and kinematically retargeted humanoid motion combines velocity estimation and motion–text alignment with continuous-time encoding and sensor and motion randomization. The resulting model transfers zero-shot to a physical Unitree G1 without physics simulation, real-robot training trajectories, or fine-tuning, achieving the lowest velocity error among evaluated methods across five datasets, including 0.099 m/s on real G1. Its motion-description retrieval also extends to retargeted G1 motion categories observed only in human training, demonstrating how human motion knowledge can support the generalization of humanoid inertial odometry and language-grounded motion understanding.**

## I. INTRODUCTION

Humanoids rely on state estimation for locomotion and balance [1], [2], and can use language-grounded motion understanding to interpret their behavior for monitoring and human–robot interaction [3]. Learning these capabilities from robot recordings, however, ties training coverage to behaviors that existing controllers can execute [4]. For state estimation, this creates a circular dependence: reliable control requires accurate estimates, yet estimators trained on robot recordings can inherit the controller's limited motion repertoire. Human datasets can broaden this coverage with diverse whole-body movements [5], [6] and natural-language descriptions [5], without requiring the humanoid to perform each behavior.

Realizing this opportunity remains challenging, even though inertial measurement units (IMUs) provide a common sensing modality. Existing learned inertial odometry typically specializes to a platform or dataset [7], [8], [9], [10], while reusable IMU representations primarily study human tasks [11], [12]. Extending these representations to humanoids requires reconciling differences in body geometry, sensor placement, and coordinate conventions [6], [13], along with varying sampling rates and measurement timing. Although retargeting can supply humanoid kinematics, the synthesized signals do not capture all contact effects, structural vibration, impacts, and sensor imperfections present in physical measurements [14]. A shared representation must therefore accommodate differences across embodiments while remaining robust to the gap between synthesized signals and physical IMU measurements.

Our central insight is to establish this correspondence through a shared inertial interface at anatomically corresponding body sites. For human motion, these sites receive signals from wearable IMUs or virtual sensors derived from kinematics; for humanoids, we synthesize the corresponding signals by combining pelvis IMU and joint encoder measurements with URDF kinematics. This kinematic IMU synthesis incorporates embodiment geometry into the input representation, which we then align to common coordinate conventions. This input space lets a unified inertial encoder learn from human and humanoid motion.

We jointly train this encoder on human and kinematically retargeted humanoid motion for velocity estimation and motion–text alignment, so that one model supports both inertial odometry and language-grounded motion understanding. Continuous-time encoding accommodates sampling differences across these sources, including different rates, missing samples, and irregular timing. Sensor and motion randomization further prepares the encoder for motion variation and disturbances absent from kinematic training signals.

Together, these choices enable zero-shot transfer to a physical Unitree G1 without physics simulation, real-robot training trajectories, or hardware-specific fine-tuning.

The benefit of learning from human motion is evident in real-G1 trajectory error with ground-truth attitude, which falls from 5.901 m with humanoid-only training to 3.153 m with joint training. The same checkpoint achieves the lowest absolute velocity error among evaluated methods across five datasets, including 0.099 m/s on real G1. The transfer extends to language-grounded motion understanding, with motion-description retrieval reaching 53.8% recall at 1 versus 17.3% for IMU2CLIP on retargeted G1 categories observed only in human training. Beyond this retargeted evaluation, demonstrations on a physical G1 illustrate motion-description retrieval on hardware.

In summary, our contributions are:

- We show human motion improves humanoid inertial odometry and enables zero-shot transfer to a physical G1 without real-robot training or fine-tuning.
- We present a shared inertial interface incorporating embodiment geometry via kinematic IMU synthesis, with continuous-time encoding across sampling rates.
- We learn a language-aligned IMU feature space supporting motion-description retrieval without task-specific classifiers, including humanoid motion categories observed only in human training.

## II. RELATED WORK

### A. Learned Inertial Odometry

Learned inertial odometry constrains drift from IMU integration by drawing on motion priors learned from data. IONet [7] and RoNIN [8] use these priors to predict windowed displacement and pedestrian velocity, respectively. Learned outputs also support model-based estimation: TLIO [9] fuses displacement and uncertainty in a filter, while AirIMU [15] combines IMU corrections and uncertainty with integration. For highly dynamic UAV motion, AirIO [10] uses body-frame measurements to improve velocity observability. Across these formulations, the learned priors and corrections reflect the sensor characteristics and behaviors encountered during training, so estimation accuracy can degrade when deployment motions fall outside that distribution [10], [16]. Their effectiveness in new settings thus depends in part on the breadth of training coverage.

Efforts to improve this generalization address both geometric variation and motion coverage: RIO [17] and EqNIO [18] use equivariance for geometric robustness, while TartanIMU [16] combines cross-platform training with fine-tuning and online adaptation. For humanoids, broadening coverage also raises the question of where diverse training motion can come from. Human datasets offer behaviors beyond those available from an existing robot controller [5], [6], motivating our use of human and kinematically retargeted humanoid motion to train one encoder. We examine whether this training improves humanoid inertial odometry and supports zero-shot transfer to physical G1 hardware without real-robot training or fine-tuning.

### B. General-Purpose and Language-Aligned IMU Models

General-purpose IMU models seek reusable inertial features for activity recognition and motion retrieval, with LIMU-BERT [19] learning contextual features from unlabeled sequences and PRIMUS [12] combining self-supervised and multimodal objectives. Although these encoders support transfer, their downstream evaluations rely on classifiers trained with task-specific labels. Connecting the features to natural-language descriptions provides a complementary route to motion understanding, allowing descriptions to serve as retrieval queries without training a classifier for each downstream label set.

This language alignment underlies IMU2CLIP [11], which projects human IMU signals into a joint video–text embedding space, and ImageBind [20], which includes IMU among six modalities in a shared representation. SensorLM [21] develops sensor–language representations from large-scale wearable data, while AURA-MFM [22] aligns IMU with third-person video, motion capture, and text for retrieval and activity recognition. Their recognition and retrieval evaluations, however, leave open whether these representations transfer motion semantics to humanoids or support velocity estimation. We address these questions through joint velocity estimation and motion–text alignment, evaluating both inertial odometry and motion-description retrieval, including humanoid categories excluded from humanoid training but represented in human training.

### C. Human-to-Humanoid Motion Transfer

Using diverse human motion for humanoid learning requires reconciling differences in morphology and physical capability. Retargeting addresses this mismatch through approaches ranging from shared latent spaces for human and humanoid kinematic structures [23] to the morphology-invariant and physics-aware constraints of Human2Humanoid [24]. Systems such as H2O [13], OmniH2O [25], and HumanPlus [4] use retargeted motion to train control policies in physics simulation for deployment on real humanoids. While these systems transfer reference motions or control policies to produce behavior, our aim is to transfer human motion knowledge at the sensing level. Our inertial encoder, trained on human and kinematically retargeted motion, uses a shared inertial interface for state estimation and language-grounded motion understanding on a physical G1, without simulation or real-robot trajectories.

## III. METHOD

### A. Problem Formulation

Let $\mathcal{S}$ denote a set of anatomically corresponding inertial sites shared by humans and humanoids. For a motion window, site $i \in \mathcal{S}$ provides a timestamped sequence

$$\mathcal{X}_i = \{(t_{i,j}, \mathbf{a}_{i,j}, \boldsymbol{\omega}_{i,j})\}_{j=1}^{T_i}, \tag{1}$$

where $\mathbf{a}_{i,j}, \boldsymbol{\omega}_{i,j} \in \mathbb{R}^3$ are measured specific force (including the gravity contribution) and angular velocity in the local sensor frame. The sequence length $T_i$ varies with the sensor

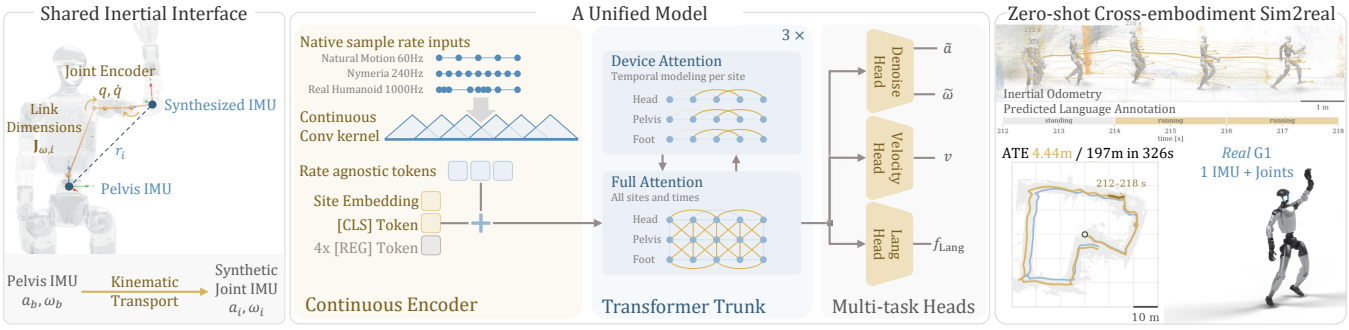



Fig. 2. Overview of the proposed method. Left: kinematic transport combines pelvis IMU measurements, joint encoders, and link kinematics to synthesize inertial signals at anatomically corresponding sites. Center: continuous convolution converts native-rate IMU streams into fixed-duration tokens. A shared transformer alternates within-site and full attention, with heads predicting body-frame velocity, gravity-free acceleration and clean angular velocity, and a motion embedding aligned with frozen text features. Right: example outputs on a physical G1 illustrate inertial odometry and motion-description retrieval.

rate and efficient inference is supported via the variable length attention [26]. Given $\mathcal{X} = \{\mathcal{X}_i\}_{i\in\mathcal{S}}$, our objective is to learn a shared inertial representation that captures both the metric and semantic properties of motion. For velocity estimation, the model predicts each site's body-frame linear velocity $\hat{\mathbf{v}}_{i,j}$ at its input timestamps, with pelvis velocity serving as root velocity. For semantic motion understanding, it maps the same motion window to a feature $\mathbf{f}_{\text{Lang}}$ aligned with a frozen text embedding space, enabling retrieval of natural-language motion descriptions. To support both tasks, auxiliary denoising supervision encourages the shared representation to preserve the underlying motion despite sensor disturbances by reconstructing gravity-free kinematic acceleration $\tilde{\mathbf{a}}_{i,j}$ and clean angular velocity $\tilde{\boldsymbol{\omega}}_{i,j}$. We learn this representation jointly from human and kinematically retargeted humanoid motion, with the goal of transferring both estimation and understanding to a physical humanoid.

### B. Unified Multi-IMU Representation

We first map human and humanoid motion signals into a common inertial input space. We instantiate $\mathcal{S}$ with 12 anatomical sites: the pelvis, head, and bilateral upper arms, forearms, upper legs, lower legs, and feet, discarding streams outside this set. Human signals come from wearable IMUs or virtual sensors derived from kinematics; humanoid signals are synthesized from the pelvis IMU and joint states. We align these streams to a common coordinate convention and encode them at their native timestamps, as detailed below.

*1) Humanoid IMU synthesis:* On G1, we transport the pelvis IMU measurements to the selected links using joint positions and velocities. Let $\mathbf{R}_{bi}$ map link-$i$ vectors into pelvis coordinates, and let $\mathbf{r}_i$, $\dot{\mathbf{r}}_i$, and $\ddot{\mathbf{r}}_i$ denote the link origin's position and its derivatives relative to the pelvis. Placing the base IMU at the pelvis origin gives

$$\begin{aligned}\boldsymbol{\omega}_i &= \mathbf{R}_{bi}^{\mathsf{T}}(\boldsymbol{\omega}_b + \mathbf{J}_{\omega,i}\dot{\mathbf{q}}),\\ \mathbf{a}_i &= \mathbf{R}_{bi}^{\mathsf{T}}\big(\mathbf{a}_b + \dot{\boldsymbol{\omega}}_b \times \mathbf{r}_i + \boldsymbol{\omega}_b \times (\boldsymbol{\omega}_b \times \mathbf{r}_i)\\ &\qquad + 2\boldsymbol{\omega}_b \times \dot{\mathbf{r}}_i + \ddot{\mathbf{r}}_i\big).\end{aligned} \tag{2}$$

Here $\boldsymbol{\omega}_b$ and $\mathbf{a}_b$ are the base IMU's angular velocity and acceleration in its sensor frame, $\mathbf{q}$ denotes joint angles, and $\mathbf{J}_{\omega,i}$ is the link's angular Jacobian in pelvis coordinates. Forward kinematics and Jacobians provide the relative link motion; numerical differentiation of the base angular velocity and joint velocities supplies angular and joint accelerations. The transport includes tangential, centripetal, Coriolis, and relative-link acceleration terms, and requires no world pose or root velocity.

For retargeted G1 motion, the base signals are derived from root kinematics; on hardware, they are read from the pelvis IMU. The synthesized accelerations inherit the gravity content of these base signals: gravity-free kinematic inputs yield gravity-free link accelerations, whereas measured specific force retains the gravity contribution at each link.

*2) Coordinate alignment:* Different datasets use different sensor-axis conventions, so we rotate all signals into the human Xsens convention [27]. For each body site, we first estimate the sensor's average orientation relative to an anatomical frame constructed from joint positions. Comparing this orientation with the Xsens reference gives a fixed rotation $\mathbf{M}_i$ from the reference axes to the dataset's native axes. We then apply this to sensor orientations and signals:

$$\mathbf{R}_{wi}^{\text{can}} = \mathbf{R}_{wi}\mathbf{M}_i, \quad \mathbf{x}_i^{\text{can}} = \mathbf{M}_i^{\mathsf{T}}\mathbf{x}_i, \quad \mathbf{x}_i \in \{\mathbf{a}_i, \boldsymbol{\omega}_i, \mathbf{v}_i\}. \tag{3}$$

Here $\mathbf{R}_{wi}$ is the native sensor-to-world rotation, and can marks quantities after alignment. The same correction is applied to acceleration, angular velocity, and velocity targets across recordings; world positions remain unchanged. Retargeted and real G1 motion share the same correction.

*3) Continuous-time encoding and decoding:* We encode each site's inertial stream into fixed-duration tokens through a learned hierarchical continuous convolution [28]. The convolution aggregates acceleration and gyroscope measurements using weights determined by their sampling times, accommodating different sampling rates and missing samples. This avoids resampling to a fixed time grid, which can introduce interpolation artifacts and distort the signals.

At each pooling stage, let $\mathcal{J}_w$ index features $\mathbf{h}_j$ at timestamps $t_j$ within an interval starting at $\tau_w$ with duration $\Delta$. Using normalized time $u_{w,j} = (t_j - \tau_w)/\Delta$, we compute

$$\begin{aligned}\bar{\mathbf{h}}_{w,k} &= \frac{\sum_{j\in\mathcal{J}_w}\phi_k(u_{w,j})\mathbf{h}_j}{\max(\epsilon, \sum_{j\in\mathcal{J}_w}\phi_k(u_{w,j}))},\\ \mathbf{y}_w &= \mathbf{b} + \sum_{k=1}^{K}\mathbf{W}_k\bar{\mathbf{h}}_{w,k}.\end{aligned} \tag{4}$$

TABLE I

TRAINING AUGMENTATION FOR MOTION DIVERSITY, SAMPLING VARIATION, AND HARDWARE DISTURBANCES.

| **Parameter** | **Setting** |
|---|---|
| *Motion and sampling* | |
| Gravity magnitude | $\mathcal{U}(9.76, 9.83)$ m/s$^2$ |
| Body scale | $\mathcal{U}(0.75, 1.0)$ / $\mathcal{U}(1.0, 1.33)$ |
| Motion speed | $\mathcal{U}(0.7, 1.3)$ / $\mathcal{U}(0.9, 1.6)$ |
| Human sampling rate | Target $\mathcal{U}(120, 240)$ Hz |
| Timestamp jitter | Spacing CV target $\mathcal{U}(0, 0.15)$ |
| *Sensor noise* | |
| Acc. noise density | $10^{-2}$ (m/s$^2$)/$\sqrt{\text{Hz}}$ |
| Gyro. noise density | $3 \times 10^{-4}$ (rad/s)/$\sqrt{\text{Hz}}$ |
| *Mechanical disturbances* | |
| Vibration | Base RMS 0.3–3.0 m/s$^2$ |
| Impacts | Base peak 3–30 m/s$^2$ |

* $\mathcal{U}$ denotes uniform distribution; paired ranges are human / G1. CV is the coefficient of variation. RMS and peak are values before modulation.

TABLE II

DATASETS USED IN THIS WORK. REAL G1 RECORDINGS ARE RESERVED FOR EVALUATION. EMB. DENOTES EMBODIMENT; FREQ. IS IN HZ; LANG. AND TRAIN MARK LANGUAGE SUPERVISION AND TRAINING USE.

| Dataset | Emb. | Motion | Real IMU | Freq. | Hours | Lang. | Train |
|---|---|---|---|---|---|---|---|
| Nymeria | Human | Real | ✓ | 240 | 300+ | ✓ | ✓ |
| Natural Motion | Human | Real | ✓ | 240 | 40.6 | ✗ | ✓ |
| BONES-Seed | Human | Real | ✓ | 120 | 144 | ✓ | ✓ |
| BONES-Seed | G1 | Retarget | ✓ | 120 | 144 | ✓ | ✓ |
| Real-Trajectory | G1 | Real | ✗ | 200 | 0.7 | ✗ | ✗ |

Here $\phi_k(u) = 1 - |u - c_k|/\delta$ are linear B-spline weights with uniformly spaced centers $c_k$ and spacing $\delta$, $\mathbf{W}_k$, $\mathbf{b}$ are learned parameters, and $\epsilon > 0$ is for numeric stability.

To capture both local motion and whole-body coordination, the transformer alternates attention within sites and across sites. Learned site embeddings are added once to the encoder's output to distinguish anatomical locations. Time-based rotary encoding [29] is applied in attention to capture elapsed time between tokens. The resulting features support velocity estimation and denoising, with a `[CLS]` token summarizing the motion for language alignment.

### C. Multi-Task Joint Training

We train a single model jointly on human motion and kinematically retargeted G1 motion. The shared trunk is optimized through four complementary objectives:

$$\mathcal{L} = \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{traj}}\mathcal{L}_{\text{traj}} + \lambda_{\text{lang}}\mathcal{L}_{\text{lang}} + \lambda_{\text{den}}\mathcal{L}_{\text{den}}. \quad (5)$$

*1) Velocity and trajectory consistency:* The velocity loss supervises each site's predicted velocity $\hat{\mathbf{v}}$ against ground truth $\mathbf{v}$: $\mathcal{L}_{\text{vel}} = \mathbb{E}\|\hat{\mathbf{v}} - \mathbf{v}\|_2$. To constrain accumulated drift, we also compare predicted and ground-truth displacements over short intervals. We integrate denoised angular velocity $\tilde{\boldsymbol{\omega}}$ [30] using fourth-order Runge–Kutta, initialized with the ground-truth attitude, to obtain the estimated sensor-to-world rotation $\hat{\mathbf{R}}$. We stop gradients through this rotation and integrate $\hat{\mathbf{R}}\hat{\mathbf{v}}$ using the trapezoidal rule to obtain world-frame position $\hat{\mathbf{p}}$. Hats denote estimates; $\mathbf{R}$ and $\mathbf{p}$ denote ground-truth rotation and position. For an interval from sample $s$ to $e$, let $\Delta\mathbf{p}_{se}^{\text{pred}} = \hat{\mathbf{R}}_s^{\mathsf{T}}(\hat{\mathbf{p}}_e - \hat{\mathbf{p}}_s)$ and $\Delta\mathbf{p}_{se}^{\text{gt}} = \mathbf{R}_s^{\mathsf{T}}(\mathbf{p}_e - \mathbf{p}_s)$ denote predicted and ground-truth displacements in their respective starting sensor frames. The trajectory loss is

$$\mathcal{L}_{\text{traj}} = \frac{1}{|\mathcal{I}|}\sum_{(s,e)\in\mathcal{I}} \left\| \frac{\Delta\mathbf{p}_{se}^{\text{pred}} - \Delta\mathbf{p}_{se}^{\text{gt}}}{t_e - t_s} \right\|_2^2. \quad (6)$$

Here $t_s, t_e$ are the endpoint timestamps, and $\mathcal{I}$ collects intervals of available durations from $\{0.25, 1, 2\}$ s without overlap. Here $|\mathcal{I}|$ is the number of intervals. Dividing by elapsed time makes errors comparable across durations.

*2) Language alignment and denoising:* For both Nymeria and BONES-Seed, a window receives language supervision only if it is fully covered by a single language annotation. A frozen CLIP ViT-L/14@336px text encoder [31] maps the motion description to a 768-dimensional target $\mathbf{z}$, and we align the window feature $\mathbf{f}_{\text{Lang}}$ using cosine distance:

$$\mathcal{L}_{\text{lang}} = 1 - \cos(\mathbf{f}_{\text{Lang}}, \mathbf{z}). \quad (7)$$

Windows without a valid language target still contribute to the motion and denoising objectives.

The auxiliary denoising loss supervises reconstruction of gravity-free acceleration and angular velocity derived from the underlying human or retargeted humanoid motion kinematics. These kinematic signals provide clean targets, while gravity and simulated sensor disturbances are added to construct the model inputs. The loss sums the mean Euclidean errors between the predicted signals and targets.

*3) Adaptive loss weights:* We fix $\lambda_{\text{den}} = \lambda_{\text{lang}} = 1$ and adapt the velocity and trajectory weights to balance their gradient norms against denoising at a shared feature layer. Each weight is an exponential moving average (EMA) of the denoising-to-task gradient norm ratio, clipped to $[0.02, 50]$.

### D. Sensor and Motion Randomization

We randomize gravity, sensor noise, bias drift, vibration, and impacts to improve robustness to disturbances absent from kinematic motion. The ranges are chosen heuristically from common sensor specifications and broadened to cover plausible operating conditions (Table I). We also vary body scale, motion speed, and sampling, applying scale and timing changes consistently to inputs and targets while keeping denoising targets free of gravity and sensor disturbances. Validation omits scale and temporal randomization while retaining noise injection, and real-G1 evaluation uses recorded signals without augmentation.

## IV. EXPERIMENTS

We evaluate one jointly trained model on inertial odometry (Sec. IV-B) and motion caption retrieval (Sec. IV-C), followed by physical G1 experiments and runtime analysis (Sec. IV-D). We then ablate the training data, input representation, and architecture (Sec. IV-E) and analyze compute allocation across model size and compute (Sec. IV-F).

TABLE III

INERTIAL ODOMETRY ACROSS HUMAN AND HUMANOID DATASETS. AVE MEASURES VELOCITY ERROR (M/S); ATE MEASURES TRAJECTORY ERROR (M), USING GROUND-TRUTH (GT) OR GYRO-INTEGRATED (PRED) ATTITUDE. LOWER IS BETTER; BOLD MARKS COLUMN MINIMA.

| | Human | | | | | | | | | Humanoid (G1) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Nymeria | | | Natural Motion | | | BONES-Seed (Human) | | | BONES-Seed (G1) | | | Real-Trajectory | | |
| Method | AVE | $\text{ATE}_{\text{GT}}$ | $\text{ATE}_{\text{Pred}}$ | AVE | $\text{ATE}_{\text{GT}}$ | $\text{ATE}_{\text{Pred}}$ | AVE | $\text{ATE}_{\text{GT}}$ | $\text{ATE}_{\text{Pred}}$ | AVE | $\text{ATE}_{\text{GT}}$ | $\text{ATE}_{\text{Pred}}$ | AVE | $\text{ATE}_{\text{GT}}$ | $\text{ATE}_{\text{Pred}}$ |
| *Pelvis IMU only* | | | | | | | | | | | | | | | |
| TLIO† [9] | 0.122 | 8.149 | **17.76** | 0.189 | 40.72 | 81.31 | 0.245 | 0.717 | 0.719 | 0.173 | 0.761 | 0.762 | 0.371 | 17.86 | 17.86 |
| AirIMU† [15] | 12.62 | – | – | 109.3 | – | – | 0.212 | 9.472 | 505.5 | 0.281 | 11.83 | 306.6 | 6.231 | 731.1 | – |
| AirIO† [10] | 0.069 | 7.481 | 113.9 | 0.280 | 158.1 | 157.9 | 0.140 | 0.530 | 1.918 | 0.118 | 0.720 | 1.690 | 0.350 | 16.82 | 20.48 |
| TartanIMU [16] | 0.301 | 37.74 | 35.53 | 0.332 | 155.8 | 156.3 | 0.328* | 2.970* | 2.964* | 0.219* | 2.176* | 2.176* | 0.391 | 18.94 | 18.96 |
| *Pelvis IMU + joint positions and velocities* | | | | | | | | | | | | | | | |
| AirIO w/ joints† | 0.060 | 6.912 | 39.64 | 0.080 | 14.11 | 209.2 | 0.205 | 0.907 | 1.652 | 0.123 | 0.620 | 1.565 | 0.190 | 5.976 | 22.25 |
| Ours | **0.026** | **2.468** | 18.20 | **0.049** | **7.456** | **50.54** | **0.031** | **0.113** | **0.148** | **0.024** | **0.101** | **0.127** | **0.099** | **3.153** | **11.50** |

† Trained separately on each dataset. * Only sequences longer than 10 seconds are evaluated.

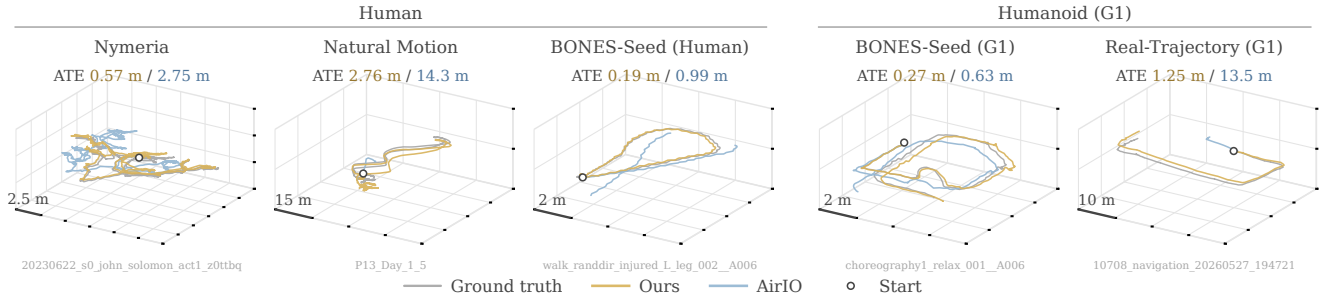



Fig. 3. Estimated trajectories sampled from each dataset. Predicted pelvis velocities from our model (gold) and AirIO (blue) are integrated using ground-truth attitude; gray shows the reference trajectory. Each panel reports per-recording $\text{ATE}_{\text{GT}}$ in meters (Ours/AirIO).

## A. Experimental Setup

*1) Datasets:* Training combines Nymeria [5], Natural Motion [6], and BONES-Seed [32] human motion with kinematically retargeted BONES-Seed (G1) motion; Table II reports total corpus hours before splitting. Human datasets use an 85%/15% training/validation split by trajectory, before window extraction. For retargeted G1, natural-language descriptions separate simple motions (e.g., walking and standing) from complex motions (e.g., dancing): simple motions follow the same split, while all complex trajectories form the unseen evaluation subset. Held-out G1 trajectories may have human counterparts in training, testing human-to-humanoid transfer (Sec. IV-C). All results in this section use held-out validation trajectories or real-G1 recordings, which are reserved entirely for zero-shot evaluation.

*2) Baselines:* For odometry, we provide the pelvis IMU as input to each single-IMU baseline and compare root-velocity estimates and their integrated trajectories across all methods. We train TLIO [9], AirIMU [15], and AirIO [10] per dataset and evaluate the released TartanIMU [16] human head. We also augment AirIO with relative joint positions and velocities to control for the additional information in our synthesized IMUs. For motion caption retrieval , we train IMU2CLIP [11] per dataset using the pelvis IMU and evaluate pretrained ImageBind [20] using the head IMU.

*3) Implementation Details:* We train for 20 epochs on 4096ms windows with bfloat16 precision and batch size of 1024. Muon [33] is used to update matrix parameters at learning rate of $2 \times 10^{-3}$, while AdamW [34] updates the rest at $5 \times 10^{-3}$. All *Ours* results in this section uses the exponential moving average (EMA) [35] with weight of 0.95.

## B. Inertial Odometry

**Task** We evaluate root-velocity estimation and the resulting integrated trajectories on human and humanoid motion, with representative trajectories in Fig. 3.

**Metric** We evaluate root-motion estimation using absolute velocity error (AVE, m/s) and absolute trajectory error (ATE, m), pooling pelvis samples across recordings. AVE measures the mean Euclidean error of body-frame velocity predictions:

$$\text{AVE} = \frac{1}{T}\sum_{t=1}^{T} \|\hat{\mathbf{v}}_t - \mathbf{v}_t\|_2 \,. \tag{8}$$

To evaluate accumulated trajectory error, we rotate predicted velocities into the world frame using attitude $\mathbf{R}_t$ and integrate from each recording's ground-truth initial position:

$$\begin{aligned} \hat{\mathbf{p}}_t &= \mathbf{p}_0 + \int_0^t \mathbf{R}_\tau \hat{\mathbf{v}}_\tau \, d\tau, \\ \text{ATE} &= \sqrt{\frac{1}{T}\sum_{t=1}^{T} \|\hat{\mathbf{p}}_t - \mathbf{p}_t\|_2^2}. \end{aligned} \tag{9}$$

We report $\text{ATE}_{\text{GT}}$ using ground-truth attitude to isolate velocity-estimation error, and $\text{ATE}_{\text{Pred}}$ using gyro-integrated attitude to additionally capture orientation drift. Gyroscope integration starts from the ground-truth initial attitude; neither setting uses filter updates or trajectory alignment.

**Analysis** Table III shows that our model achieves the lowest AVE and $\text{ATE}_{\text{GT}}$ on all five datasets, including

TABLE IV

MOTION CAPTION RETRIEVAL: MOTION-TO-TEXT RECALL AT 1 (R@1, ↑) / TEXT-TO-MOTION MEAN AVERAGE PRECISION (MAP %, ↑). UNSEEN G1 CATEGORIES OCCUR ONLY ON HUMAN EMBODIMENT IN TRAINING DATASET, SHOWING THE CROSS-EMBODIMENT KNOWLEDGE TRANSFER.

| | Human | | Humanoid (G1) | |
|---|---|---|---|---|
| Method | Nymeria | BONES-Seed | Seen cat. | Unseen cat. |
| Collapse | 4.4 / 0.1 | 19.7 / 0.2 | 21.9 / 0.8 | 9.8 / 0.1 |
| ImageBind [20] | 8.9 / 0.1 | 7.1 / 0.4 | 11.4 / 1.6 | 4.6 / 0.1 |
| IMU2CLIP† [11] | **16.2** / 0.4 | 51.8 / 5.2 | 36.1 / 11.2 | 17.3 / 0.5 |
| Ours | 14.9 / **0.6** | **60.7** / **15.9** | **63.4** / **32.9** | **53.8** / **11.3** |

†Trained separately on each dataset.

0.024 m/s AVE on retargeted G1 and 0.099 m/s on real G1. It also leads $\mathrm{ATE}_{\mathrm{Pred}}$ on four datasets, with TLIO slightly better on Nymeria. The higher errors under integrated attitude indicate that orientation drift remains a limitation.

### C. IMU–Language Retrieval

**Task** We evaluate motion–language alignment in both directions: retrieving descriptions from IMU windows and retrieving motion windows from text queries. Fig. 4 illustrates retrieval over a motion sequence.

**Metric** We rank retrieval candidates by cosine similarity within each dataset. For IMU-to-text retrieval, we report recall at 1 (R@1) against 19 random distractors averaged over eight draws. For text-to-motion retrieval, we report mean average precision (mAP) over all annotated motion windows, averaged across distinct text queries. To assess whether retrieval reflects motion-specific information, we include a *Collapse* baseline that ignores the IMU input and assigns every window the same embedding: the normalized mean of the evaluation text targets. This baseline measures performance attainable from a constant representation, providing a reference for interpreting retrieval scores when the embedding does not distinguish between motions.

**Analysis** Table IV shows that our model leads both metrics on BONES-Seed human motion and both G1 splits. On categories excluded from G1 training, it reaches 53.8% R@1 and 11.3% mAP versus 17.3% and 0.5% for IMU2CLIP, supporting semantic transfer from human motion. On Nymeria, IMU2CLIP achieves higher R@1 (16.2% versus 14.9%), while our model achieves higher mAP.

### D. Real-world Experiment and Runtime Analysis

We evaluate motion tracking and language retrieval on a physical G1 without real-robot training or fine-tuning. Fig. 5 shows the tracked stair descent, with retrieved descriptions changing from "Walking" to "Standing" as the robot stops.

Despite its greater capacity and joint support for state estimation and motion understanding, our encoder remains suitable for real-time robot use. With TensorRT on an NVIDIA L40S, it processes a 4-s window of 12 IMUs sampled at 200 Hz in 1.78 ms (Fig. 5). This latency is well below the 5-ms input sampling interval, allowing inference to keep pace with incoming measurements.

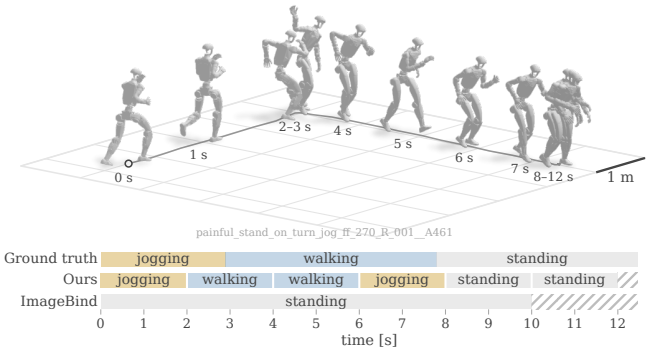



Fig. 4. Motion-description retrieval on a retargeted G1 sequence. Sampled poses (top) accompany annotated categories and descriptions retrieved by our model and ImageBind (bottom). Our model uses 2s and ImageBind uses 10s windows; hatching marks intervals without a complete input window.

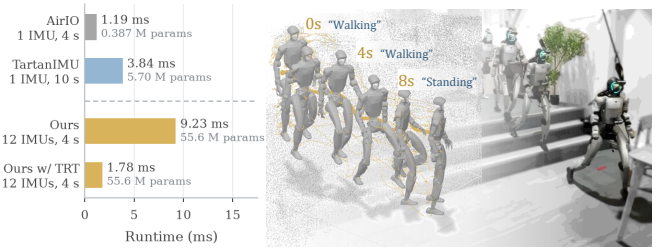



Fig. 5. Inference runtime and deployment on a physical G1. Left: latency on an NVIDIA L40S with 200-Hz inputs; labels show IMU count, window duration, and model size. TRT denotes TensorRT. Right: stair descent shown by robot poses over a 3D point cloud, with retrieved descriptions changing from "Walking" to "Standing" as the robot stops.

### E. Ablation Studies

We examine how human motion data, the shared multi-IMU representation, encoder design, and training-data diversity contribute to motion estimation across embodiments. Table V shows the results on human and humanoid datasets.

*1) Human-to-Humanoid Transfer:* Diverse human motion may provide useful motion priors beyond those learned from humanoid data alone. We compare our jointly trained model with a humanoid specialist trained exclusively on retargeted G1 motion. Joint training improves both trajectory metrics on retargeted and real G1, reducing $\mathrm{ATE}_{\mathrm{GT}}$ from 0.162 to 0.101 m and from 5.901 to 3.153 m, respectively (Table V). These results suggest that broader human motion experience can benefit humanoid odometry even relative to embodiment-specific training. Fig. 6 further illustrates the shared representation, with corresponding human and G1 motion categories overlapping in the t-SNE plot.

*2) Multi-IMU Representation:* Our shared inertial interface incorporates embodiment geometry through kinematic IMU synthesis, aligning human and humanoid motion at corresponding body sites. We compare against pelvis-only input to assess the value of whole-body motion cues, and against pelvis IMU plus joint positions and velocities to test their representation in a common inertial space. Table V shows that multi-IMU input reduces real-G1 $\mathrm{ATE}_{\mathrm{GT}}$ from 8.114 to 3.153 m relative to pelvis-only input and improves both trajectory metrics over direct joints on retargeted and real G1. Although direct joints perform better on Nymeria, the humanoid results support incorporating embodiment ge-

TABLE V

MODEL AND TRAINING RECIPE ABLATIONS (ATE IN M, LOWER IS BETTER), GROUPED BY THE FOUR QUESTIONS IN SEC. IV-E.

| Variant | Nymeria | | BONES-Seed (G1) | | Real-Trajectory | |
|---|---|---|---|---|---|---|
| | $ATE_{GT}$ | $ATE_{Pred}$ | $ATE_{GT}$ | $ATE_{Pred}$ | $ATE_{GT}$ | $ATE_{Pred}$ |
| *1. Training embodiment* | | | | | | |
| Humanoid-only | 8.041 | 17.50 | 0.162 | 0.180 | 5.901 | 13.15 |
| *2. Input representation* | | | | | | |
| Single pelvis IMU | 3.592 | 16.66 | 0.330 | 0.335 | 8.114 | 12.81 |
| Pelvis IMU + joints | **2.283** | 16.12 | 0.136 | 0.164 | 3.513 | 18.36 |
| *3. Model design* | | | | | | |
| w/o Alternating Attn. | 2.466 | **14.31** | 0.110 | 0.128 | 3.530 | 11.71 |
| w/o Continuous Enc. | 31.42 | 31.51 | 1.189 | 1.189 | 17.37 | 17.67 |
| *4. Data diversity* | | | | | | |
| 1/64 of data | 4.882 | 18.40 | 0.154 | 0.166 | 4.369 | 12.44 |
| 1/16 of data | 3.500 | 16.93 | 0.111 | 0.131 | 3.792 | 11.95 |
| 1/4 of data | 2.847 | 14.64 | **0.099** | **0.121** | 3.716 | 11.98 |
| Ours | 2.468 | 18.20 | 0.101 | 0.127 | **3.153** | **11.50** |

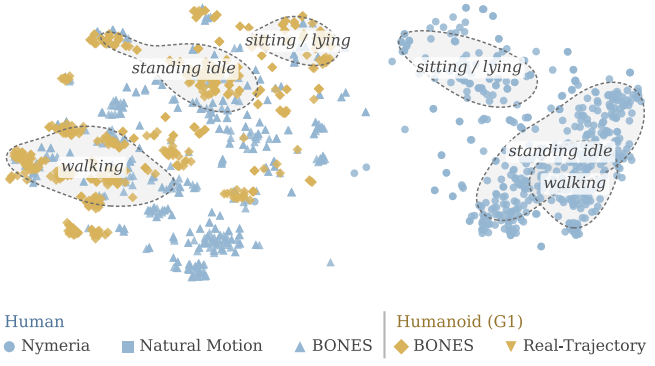



Fig. 6. Human and humanoid motion in the learned feature space. A t-SNE projection of [CLS] embeddings uses blue for humans, gold for G1, and marker shapes for datasets. Dashed regions enclose 60% of windows in each motion category, showing overlap between human and humanoid data.

ometry into a common inertial representation.

*3) Model Design:* We examine whether explicitly modeling each site's temporal structure and physical sampling time benefits the shared encoder (Table V). Alternating within-site and full attention modestly improves humanoid odometry over full attention alone, suggesting that dedicated per-site modeling complements cross-body interactions, although gains on Nymeria are mixed. Continuous-time tokenization yields larger gains across all three datasets over fixed-sample patches, whose duration varies with sampling rate, supporting a consistent temporal representation across embodiments.

*4) Data Diversity:* We test whether broader motion coverage improves generalization by varying unique training sequences at fixed model size and approximately matched optimization budgets (Table V). More diverse data consistently reduces $ATE_{GT}$ on Nymeria and real G1, supporting broader motion exposure over repeated training on a smaller corpus. Retargeted G1 performs best with quarter-data training, suggesting that additional diversity can benefit hardware transfer without improving retargeted-motion accuracy.

### F. Compute Allocation

Efficient scaling requires balancing model capacity with training exposure under a limited compute budget. To characterize this tradeoff, we fit a Chinchilla-style scaling law [36]

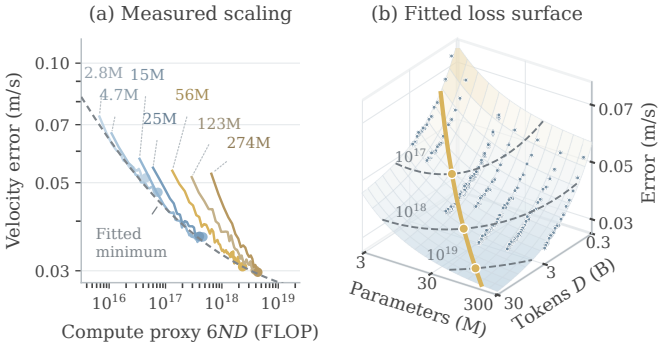



Fig. 7. Model capacity and training exposure under a fixed compute budget. (a) Measured velocity error and fitted compute-optimal trend (dashed). (b) Fitted error surface with observations (blue), constant-$6ND$ contours (gray), and predicted optima (gold). $N$ counts parameters and $D$ counts processed tokens, including repeats; the surface includes extrapolation.

across models of varying width and depth:

$$L(N, D) = E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}}, \tag{10}$$

where $L$ is shared-site validation velocity error, $N$ is the parameter count, and $D$ is the number of processed tokens, including repeated windows. The fit estimates an error floor $E$, coefficients $A, B$, and scaling exponents $\alpha, \beta$.

Under the compute proxy $C = 6ND$, the fit yields $N_{\text{opt}} \propto C^{0.52}$ and $D_{\text{opt}} \propto C^{0.48}$ (Fig. 7). These nearly equal exponents suggest that increasing model capacity and training exposure together uses additional compute effectively. The fitted contours illustrate this tradeoff and guide allocation across compute budgets. These estimates include extrapolation and optimize validation error, so they need not yield the best hardware transfer.

## V. CONCLUSIONS

We presented a unified inertial model that zero-shot transfers human motion knowledge to a physical humanoid for motion estimation and language understanding, without physics simulation, real-robot training trajectories, or fine-tuning. A shared inertial interface incorporates embodiment geometry through kinematic IMU synthesis, while continuous-time encoding accommodates different sampling rates. Together, these designs enable a single checkpoint to achieve the lowest velocity error among the evaluated methods across five datasets. The learned language-aligned representation also supports motion interpretation through text retrieval without task-specific classifiers, including humanoid motion categories absent from G1 training.

Hardware evaluation currently covers a single humanoid platform, and language supervision is partly synthetic with a limited vocabulary. Despite accurate velocity estimates, orientation drift remains a major source of trajectory error. Future work will extend evaluation across humanoid platforms and integrate learned velocities with filtering for closed-loop state estimation and control.

## REFERENCES

[1] S. Kuindersma, R. Deits, M. Fallon, A. Valenzuela, H. Dai, F. Permenter, T. Koolen, P. Marion, and R. Tedrake, "Optimization-based locomotion planning, estimation, and control design for the Atlas

humanoid robot," *Autonomous Robots*, vol. 40, no. 3, pp. 429–455, 2016.

[2] N. Rotella, M. Bloesch, L. Righetti, and S. Schaal, "State estimation for a humanoid robot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2014, pp. 952–958.

[3] B. Hayes and J. A. Shah, "Improving robot controller transparency through autonomous policy explanation," in *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2017, pp. 303–312.

[4] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, "HumanPlus: Humanoid shadowing and imitation from humans," in *Proceedings of The 8th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 270. PMLR, 2025, pp. 2828–2844. [Online]. Available: https://proceedings.mlr.press/v270/fu25a.html

[5] L. Ma, Y. Ye, F. Hong, V. Guzov, Y. Jiang, R. Postyeni, L. Pesqueira, A. Gamino, V. Baiyya, H. J. Kim, K. Bailey, D. S. Fosas, C. K. Liu, Z. Liu, J. Engel, R. De Nardi, and R. Newcombe, "Nymeria: A massive collection of multimodal egocentric daily motion in the wild," in *European Conference on Computer Vision (ECCV)*, 2024.

[6] J. H. Geissinger and A. T. Asbeck, "Motion inference using sparse inertial sensors, self-supervised learning, and a new dataset of unscripted human motion," *Sensors*, vol. 20, no. 21, p. 6330, 2020.

[7] C. Chen, X. Lu, A. Markham, and N. Trigoni, "IONet: Learning to cure the curse of drift in inertial odometry," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018, pp. 6468–6476.

[8] S. Herath, H. Yan, and Y. Furukawa, "RoNIN: Robust neural inertial navigation in the wild: Benchmark, evaluations, & new methods," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 3146–3152.

[9] W. Liu, D. Caruso, E. Ilg, J. Dong, A. I. Mourikis, K. Daniilidis, V. Kumar, and J. Engel, "TLIO: Tight learned inertial odometry," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5653–5660, 2020.

[10] Y. Qiu, C. Xu, Y. Chen, S. Zhao, J. Geng, and S. Scherer, "AirIO: Learning inertial odometry with enhanced IMU feature observability," *IEEE Robotics and Automation Letters*, vol. 10, no. 9, pp. 9368–9375, 2025. [Online]. Available: https://doi.org/10.1109/LRA.2025.3581130

[11] S. Moon, A. Madotto, Z. Lin, A. Saraf, A. Bearman, and B. Damavandi, "IMU2CLIP: Language-grounded motion sensor translation with multimodal contrastive learning," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 13 246–13 253.

[12] A. M. Das, C. I. Tang, F. Kawsar, and M. Malekzadeh, "PRIMUS: Pretraining IMU encoders with multimodal self-supervision," in *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5. [Online]. Available: https://doi.org/10.1109/ICASSP49660.2025.10888874

[13] T. He, Z. Luo, W. Xiao, C. Zhang, K. Kitani, C. Liu, and G. Shi, "Learning human-to-humanoid real-time whole-body teleoperation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024. [Online]. Available: https://doi.org/10.1109/IROS58592.2024.10801984

[14] M. Fallon, "Accurate and robust localization for walking robots fusing kinematics, inertial, vision and LIDAR," *Interface Focus*, vol. 8, no. 4, p. 20180015, 2018.

[15] Y. Qiu, C. Wang, C. Xu, Y. Chen, X. Zhou, Y. Xia, and S. Scherer, "AirIMU: Learning uncertainty propagation for inertial odometry," *arXiv preprint arXiv:2310.04874*, 2023.

[16] S. Zhao, S. Zhou, R. Blanchard, Y. Qiu, W. Wang, and S. Scherer, "Tartan IMU: A light foundation model for inertial positioning in robotics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 22 520–22 529.

[17] X. Cao, C. Zhou, D. Zeng, and Y. Wang, "RIO: Rotation-equivariance supervised learning of robust inertial odometry," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[18] R. K. Jayanth, Y. Xu, Z. Wang, E. Chatzipantazis, K. Daniilidis, and D. Gehrig, "EqNIO: Subequivariant neural inertial odometry," in *International Conference on Learning Representations (ICLR)*, 2025.

[19] H. Xu, P. Zhou, R. Tan, M. Li, and G. Shen, "LIMU-BERT: Unleashing the potential of unlabeled data for IMU sensing applications," in *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems (SenSys)*, 2021.

[20] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "ImageBind: One embedding space to bind them all," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 15 180–15 190.

[21] Y. Zhang, K. Ayush, S. Qiao, A. A. Heydari, G. Narayanswamy, M. A. Xu, A. A. Metwally, S. Xu, J. Garrison, X. Xu, T. Althoff, Y. Liu, P. Kohli, J. Zhan, M. Malhotra, S. Patel, C. Mascolo, X. Liu, D. McDuff, and Y. Yang, "SensorLM: Learning the language of wearable sensors," in *Advances in Neural Information Processing Systems*, vol. 38, 2025.

[22] K. Matsuishi, K. Ukita, and T. Okita, "Multimodal foundation model for cross-modal retrieval and activity recognition tasks," *International Journal of Activity and Behavior Computing*, vol. 2025, no. 3, pp. 1–25, 2025.

[23] S. Choi, M. Pan, and J. Kim, "Nonparametric motion retargeting for humanoid robots on shared latent space," in *Robotics: Science and Systems*, 2020.

[24] T. Huang, F. Yuan, J. Gu, S. Fang, X. Zhang, Y. Wang, W. Gao, and S. Zhang, "Human2Humanoid: Physics-aware cross-morphology motion retargeting for humanoid robots," *arXiv preprint arXiv:2606.03476*, 2026.

[25] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. M. Kitani, C. Liu, and G. Shi, "OmniH2O: Universal and dexterous human-to-humanoid whole-body teleoperation and learning," in *Proceedings of The 8th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 270. PMLR, 2025, pp. 1516–1540. [Online]. Available: https://proceedings.mlr.press/v270/he25b.html

[26] T. Dao, "FlashAttention-2: Faster attention with better parallelism and work partitioning," in *International Conference on Learning Representations (ICLR)*, 2024.

[27] *MVN User Manual*, Xsens, sections on coordinate systems and segment axes. Accessed: 2026-09-10. [Online]. Available: https://www.xsens.com/hubfs/Downloads/Manuals/MVN_User_Manual.pdf

[28] D. W. Romero, A. Kuzina, E. J. Bekkers, J. M. Tomczak, and M. Hoogendoorn, "CKConv: Continuous kernel convolution for sequential data," in *International Conference on Learning Representations (ICLR)*, 2022. [Online]. Available: https://openreview.net/forum?id=8FhxBtXSl0

[29] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "RoFormer: Enhanced transformer with Rotary Position Embedding," *Neurocomputing*, vol. 568, p. 127063, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231223011864

[30] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-manifold preintegration for real-time visual–inertial odometry," *IEEE Transactions on Robotics*, vol. 33, no. 1, pp. 1–21, 2017.

[31] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 139, 2021, pp. 8748–8763.

[32] Bones Studio, "BONES-SEED: Skeletal everyday embodied dataset," 2026, accessed: 2026-08-16. [Online]. Available: https://huggingface.co/datasets/bones-studio/seed

[33] K. Jordan, Y. Jin, V. Boza, J. You, F. Cesista, L. Newhouse, and J. Bernstein, "Muon: An optimizer for hidden layers in neural networks," 2024. [Online]. Available: https://kellerjordan.github.io/posts/muon/

[34] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[35] D. Morales-Brotons, T. Vogels, and H. Hendrikx, "Exponential moving average of weights in deep learning: Dynamics and benefits," *Transactions on Machine Learning Research*, 2024. [Online]. Available: https://openreview.net/forum?id=2M9CUnYnBA

[36] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, O. Vinyals, J. Rae, and L. Sifre, "An empirical analysis of compute-optimal large language model training," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 30 016–30 030.