OMNIIMU-BENCH: BENCHMARKING WRIST IMU REPRESENTATIONS ACROSS KINEMATIC, SEMANTIC AND CLINICAL TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Wrist-worn inertial measurement units (IMUs) provide a common sensing modality for tasks ranging from activity and gesture recognition to pose estimation, inertial odometry and clinical motor assessment. Existing wearable benchmarks evaluate performance across multiple tasks, but leave open how supervision for one task affects transfer to others and whether these tasks benefit from a shared representation. We present OmniIMU-Bench, a benchmark of ten tasks from eight datasets for studying IMU multi-task performance and cross-task transfer. We compare six self-supervised methods pretrained on unlabeled recorded and simulated wrist motion, evaluating classification, pose and velocity estimation, and IMU–language retrieval. We find that our masked autoencoder, InrtlWrist, achieves the best or joint-best frozen-probe performance on nine of ten tasks, and fine-tuning improves on training from scratch on nine of ten tasks, including full-body pose, odometry and retrieval. Across transfer and joint-training experiments, supervision that benefits one target can reduce performance on others, with supervised sources transferring within kinematic estimation while, under frozen probing, none approaches the pretrained encoder on language retrieval, and joint-training gains that come at a cost to the paired task. These findings distinguish broad downstream usefulness from complementary task supervision. Code, splits, and a leaderboard will be released.

1 INTRODUCTION

Wrist-worn inertial measurement units (IMUs) provide a common sensing modality for studying human movement. Embedded in everyday smartwatches, accelerometers and gyroscopes capture motion relevant to activity recognition (Yuan et al., 2024), clinical motor assessment (Varghese et al., 2024) and human–computer interaction (Laput et al., 2016; Wen et al., 2016), with additional cues for articulated pose estimation and inertial odometry. Like the shadows in Plato's allegory of the cave (Plato, 1991, Book VII, 514a–517a), different downstream tasks reveal distinct aspects of the same underlying human motion, each offering only a partial view (Huh et al., 2024). These perspectives raise the question of whether these views share a representation.

Self-supervised learning has demonstrated substantial reuse of unlabeled wearable recordings. Yuan et al. (2024) show transfer from large-scale free-living accelerometry to activity recognition across datasets and populations. Inertia-1 examines objectives, architectures, scale and sensing configurations across activity recognition, freezing-of-gait detection and disease prediction (Xu et al., 2026). These studies establish transfer across classification tasks and sensing conditions, while leaving less understood how the same representations support articulated joint rotations, metric velocity and motion–language correspondence.

These targets also allow us to examine whether different forms of motion supervision are complementary. Task-transfer studies in vision provide a precedent for measuring such relationships (Zamir et al., 2018). For wrist IMUs, we ask both how broadly pretrained features transfer and whether supervision for one motion target improves or compromises their usefulness for others.

We present OmniIMU-Bench, a benchmark of ten tasks from eight datasets addressing these questions (Figure 1). We compare six self-supervised methods pretrained on unlabeled activity and daily-living

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

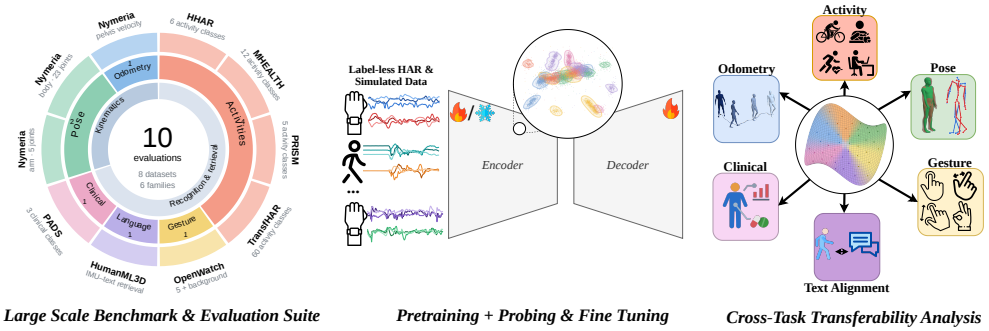

Figure 1: Overview of OmniIMU-Bench. Ten tasks from eight datasets evaluate representations learned from single-wrist accelerometer and gyroscope signals. Pretraining uses unlabeled recorded and simulated motion. Frozen probing and fine-tuning measure downstream performance, while source-task transfer and joint training examine the effects of supervision across tasks. Representation visualizations are schematic.

recordings together with simulated wrist motion. Evaluation spans classification, pose and velocity estimation, and IMU–language retrieval, using six-channel inputs from a single wrist. Beyond frozen probing and fine-tuning, we examine supervised source-task transfer and selected joint-training pairs. The results reveal a distinction between broadly useful pretraining and compatible task supervision. Full-body pose and odometry improve over training from scratch after fine-tuning, and each kinematic task's best frozen source is another kinematic task, whereas under frozen probing no supervised source approaches the pretrained encoder on language retrieval. Five of the six kinematic transfers survive, at about half their size, when the tasks are evaluated on disjoint participants, and the observed trade-offs show why individual-task performance alone is insufficient to characterize a shared wrist representation.

We make three contributions:

- **A benchmark and comprehensive evaluations.** OmniIMU-Bench brings together ten tasks across six families to evaluate wrist-IMU representations through categorical, kinematic and motion–language prediction targets. We compare six self-supervised methods under matched encoder capacity and common task protocols, alongside training from scratch and retrained task-specific systems.
- **A reusable pretrained wrist-IMU model.** We provide InrtlWrist, a masked autoencoder pretrained on unlabeled recorded and simulated wrist motion. Among the six controlled methods, it achieves the best or joint-best frozen-probe performance on nine of ten tasks. Fine-tuning improves on training from scratch on nine of the ten tasks, including full-body pose, odometry and IMU–language retrieval.
- **An analysis of cross-task transfer and complementarity.** We measure how source-task supervision and joint training affect other prediction targets. The experiments distinguish beneficial transfer between kinematic tasks from the gap that remains to the pretrained encoder on motion–language retrieval under frozen probing, and identify joint-training gains that come at the expense of the paired task. Additional analyses examine pretraining volume and motion coverage.

We will release the benchmark code, pretrained models, task definitions, splits and a public website with a leaderboard.

2 Related Work

Wearable sensor benchmarks. Wearable-IMU representations are mostly evaluated on activity recognition: pretraining objectives (Haresamudram et al., 2022; Cai et al., 2026), cross-dataset generalization (Hong et al., 2024) and foundation models on activity and health outcomes (Yuan et al., 2024; Narayanswamy et al., 2025; Abbaspourazad et al., 2024; Xu et al., 2026; Tarale et al.,

2026); SlotFM adds gesture, gait and sports (Park et al., 2025), NormWear physiological signals (Luo et al., 2026). Sparse-IMU pose (Huang et al., 2018; Yi et al., 2021; 2022; Jiang et al., 2022; Mollyn et al., 2023; Xu et al., 2024) and inertial odometry (Chen et al., 2018; Herath et al., 2020; Liu et al., 2020; Qiu et al., 2025) are studied apart, though label-free physical decoders now link them to self-supervised learning (Leng et al., 2026). Task-specific systems for activity (Ordóñez & Roggen, 2016), gestures (Wen et al., 2016; Laput et al., 2016; Bonazzi et al., 2026), Parkinson's disease (Bot et al., 2016; Varghese et al., 2024) and IMU–language alignment (Zhang et al., 2024; Moon et al., 2023; Zhang et al., 2025b) are our reference points (Appendix B). None measures how supervision on one task affects the others, which OmniIMU-Bench adds.

Self-supervised learning for wearable sensors. Pretext tasks for wearable IMUs include transformation prediction (Saeed et al., 2019; Yuan et al., 2024), self-training (Tang et al., 2021), masked reconstruction (Haresamudram et al., 2020; Xu et al., 2021; He et al., 2022), masked modeling with contrast (Dong et al., 2023), factorized multimodal contrast (Liu et al., 2023), relative contrast (Xu et al., 2025) and video and text supervision (Das et al., 2025); language models can also synthesize virtual IMU (Leng et al., 2024). Each came with its own backbone, data and protocol, so we compare five with masked autoencoding under one corpus and matched capacity (Section 4.2). Transformers for multivariate series encode channels independently (Nie et al., 2023; Goswami et al., 2024), as whole-channel tokens (Liu et al., 2024) or in separate time and variate stages (Zhang & Yan, 2023); the first loses the accelerometer–gyroscope coupling and the second the patch-level resolution pose and odometry need, hence the alternating frame-wise and global attention of InrtlWrist, adapted from VGGT (Wang et al., 2025) (Section 4.1).

Task transfer, probing and pretraining data. Task relationships have been measured by transferring supervision (Zamir et al., 2018; Vu et al., 2020; Achille et al., 2019), by choosing tasks to train together (Caruana, 1997; Standley et al., 2020) and by comparing representations directly (Kornblith et al., 2019; Huh et al., 2024); we take the transfer route for wrist IMUs (Section 6). Because probe accuracy reflects the probe as well as the representation (Alain & Bengio, 2016; Hewitt & Liang, 2019), we fix light decoders and add untrained-encoder and raw-input controls (Section 4.3). Scaling laws and data pruning (Kaplan et al., 2020; Sorscher et al., 2022) motivated wearable studies of data volume and participant count (Narayanswamy et al., 2025; Xu et al., 2026); we vary motion coverage at matched volume (Section 7).

3 Datasets and Tasks

3.1 Pretraining Data

Pretraining uses 197,094 unlabeled accelerometer–gyroscope windows (10.24 s, 50% overlap) from approximately 280 hours of recorded and simulated wrist motion. We evaluate transfer to ten tasks from eight datasets (Table 1). The five pretraining sources are an activity-recognition pool (Bhattacharya et al., 2022; Kwapisz et al., 2011; Weiss et al., 2019; Morris et al., 2014; Reiss & Stricker, 2012; Mollyn et al., 2022; Shoaib et al., 2016), OCDetect free-living recordings (Burchard et al., 2026), EatingSpeed sessions (Wang et al., 2024a;b), and wrist motion simulated from Motion-X++ (Lin et al., 2023; Zhang et al., 2025a) and BONES-SEED (Bones Studio, 2025). Evaluation datasets are excluded from pretraining. Sampling and provenance checks, including unresolved PRISM–SAMoSA participant overlap, appear in Appendices A and C.

Table 1: Pretraining sources and evaluation datasets. Parentheses give class counts.

Regime / task	Dataset	Hours
Pretraining		
HAR/Everyday	UT-Watch	45.3
	WISDM	44.7
	RecoFit	43.0
	PAMAP2	7.1
	SAMoSA	6.7
	Shoaib	6.5
	OCDetect	2,600
	EatingSpeed	525.9
Simulated	Motion-X++	40.4
	BONES-SEED	144.2
Evaluation		
Arm Pose	Nymeria	300
Body Pose	Nymeria	300
Odometry	Nymeria	300
Gesture	OpenWatch	1.9
Retrieval	HumanML3D	28.6
Clinical (3)	PADS	35.9
Activity (6)	HHAR	4.8
Activity(12)	MHEALTH	1.9
Activity(5)	PRISM	16.3
Activity(60)	TransfHAR	2.0

3.2 Tasks

Pose estimation. Nymeria's real right-wrist IMU and motion capture (Ma et al., 2024) support rotation pre-

diction for five right-arm and 23 SMPL joints every 0.32 s within 10.24 s windows. We report arm joint position error (cm) and root-relative body MPJPE (mm). All three Nymeria tasks share a recording-session split; participants recur across splits.

Inertial odometry. Nymeria pelvis velocities predicted in the wrist frame are rotated into world coordinates using ground-truth initial wrist orientation propagated by gyroscope integration. Integrated trajectories are scored by absolute trajectory error (ATE).

Microgesture recognition. OpenWatch (Bonazzi et al., 2026) votes predictions from 1 s windows into clip labels, scored by macro-F1. Subject splits use canonical identities after duplicate and mirrored participant records are merged.

IMU–language retrieval. Simulated right-wrist HumanML3D clips (Guo et al., 2022) are aligned with CLIP text embeddings (Radford et al., 2021). We report mean Recall@1 across both directions in galleries of 100 on the official test split. Shared source-motion segments across splits are documented in Appendix C.

Clinical classification. PADS (Varghese et al., 2024) assigns 469 participants to healthy control, Parkinson's disease or differential diagnosis. Inputs use one wrist; participant predictions average window posteriors across both wrists. We report participant macro-F1 on a subject-disjoint split.

Activity recognition. All four activity tasks classify 2.56 s windows using macro-F1. HHAR, MHEALTH and PRISM use subject-disjoint splits (Stisen et al., 2015; Banos et al., 2014; Arakawa et al., 2022); PRISM uses its five global classes. TransfHAR has ten participants, but most activities were performed by only one (Bradshaw et al., 2026). Each recording is therefore split chronologically into 80% training and 20% evaluation, dropping one boundary window.

4 METHODS

4.1 ARCHITECTURE

Given a wrist-worn window $\mathbf{X} \in \mathbb{R}^{T\times 6}$, we divide each accelerometer and gyroscope channel into 0.32 s patches, forming a $6 \times P$ token grid, where P is the number of temporal patches. A 12-layer transformer f_θ with width $D = 384$ and approximately 21 million parameters produces patch representations $\mathbf{H} \in \mathbb{R}^{6P\times D}$.

Following VGGT (Wang et al., 2025), layers alternate between frame-wise attention across the six channels at each patch position and global attention across all channels and positions, including a learnable summary token.

4.2 TRAINING

We pretrain InrtlWrist by masked reconstruction. The encoder processes visible patches, and a reconstruction decoder predicts the masked measurements. The reconstruction decoder is discarded after pretraining.

To support different input durations, each batch is cropped to 1.28–10.24 s in multiples of 0.32 s. Positional embeddings are restricted to the crop, avoiding resizing or interpolation of the position code, following resolution-agnostic vision transformers (Dehghani et al., 2023).

We compare against LIMU-BERT (masked reconstruction), SimMTM (reconstruction and contrastive losses), FOCAL (contrastive, orthogonality and temporal losses), RelCon (relative contrastive loss) and SSL-Wearables (transformation-prediction cross-entropy) using a shared pretraining corpus and matched encoder capacity. These objectives use plain attention over the flattened token grid. A plain-attention InrtlWrist variant controls for this architectural difference. Preprocessing and pretraining controls are detailed in Appendix A.

4.3 EVALUATION PROTOCOLS

For each task, a decoder g_ϕ predicts $\hat{\mathbf{y}} = g_\phi(f_\theta(\mathbf{X}))$. Let $\mathcal{L}_{\text{task}}$ denote the supervised task loss, θ_{SSL} pretrained encoder weights and θ_{rand} a random initialization. We evaluate frozen probing, fine-tuning and training from scratch under the following optimization settings. Frozen probing fixes the encoder

at θ_{SSL}. Fine-tuning and training from scratch optimize both encoder and decoder, initializing θ from θ_{SSL} and θ_{rand}, respectively.

$$\underbrace{\min_{\phi} \mathcal{L}_{\text{task}}(\theta_{\text{SSL}}, \phi)}_{\text{Frozen probing}} \qquad \underbrace{\min_{\theta,\phi} \mathcal{L}_{\text{task}}(\theta, \phi)}_{\text{Fine-tuning}} \qquad \underbrace{\min_{\theta,\phi} \mathcal{L}_{\text{task}}(\theta, \phi)}_{\text{From scratch}}. \tag{1}$$

Within each task, controlled comparisons share the decoder architecture, optimizer, training budget, split and model-selection rule. Classification probes use linear heads. Kinematic probes use width-192 attention decoders with two cross-channel and two temporal layers for arm pose, one cross-channel layer for full-body pose, and one temporal layer for odometry. Decoder-capacity sensitivity is examined in Section 7.

Appendix B details task definitions, splits and metrics. Each encoder's pretraining corpus is documented; evaluations on datasets included in that corpus are flagged and excluded from matched comparisons. Released external models and retrained task-specific systems are reported separately.

5 Results

Among the six matched pretraining methods, InrtlWrist achieves the best or joint-best primary score on nine of ten tasks under frozen probing and eight after fine-tuning; fine-tuning improves on training from scratch on nine of ten tasks, gesture being the exception (54.2 against 60.8) (Figure 2). Its frozen features outperform training from scratch on mean activity recognition and IMU–language retrieval, but trail it on full-body pose and odometry; fine-tuning surpasses the scratch baseline on both kinematic tasks (Tables 2–4).

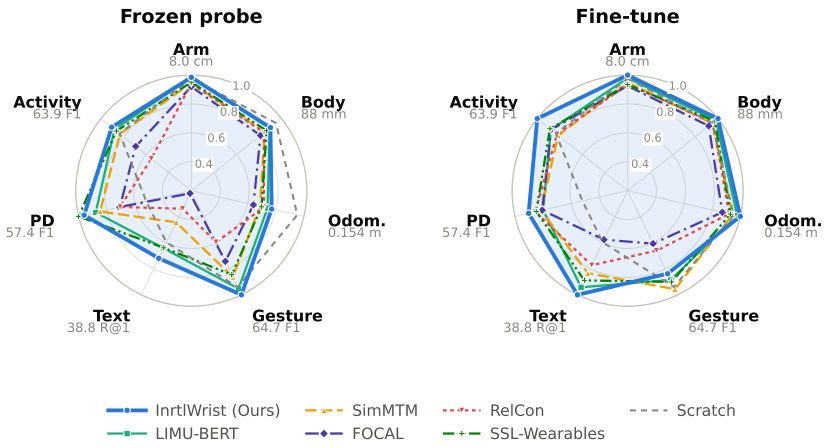

Figure 2: **Downstream performance across task families.** Frozen probing (left) and fine-tuning (right), with training from scratch shown in both panels. Activity averages macro-F1 across four datasets. Each axis is normalized to the best score across both panels, including scratch, using score/best for F1 and Recall@1 and best/error for error metrics. Axis labels give these reference values in the original units. Larger radii indicate better performance; the displayed range is 0.20–1.00.

5.1 Activity, Gesture, and Clinical Recognition

Activity recognition. InrtlWrist leads the six matched methods on all four activity datasets under frozen probing, with 57.8 mean macro-F1 against 56.2 for LIMU-BERT and 53.3 for training from scratch (Table 2). Fine-tuning raises its mean to 63.9, with the largest gains on TransfHAR (72.7 to 86.1) and PRISM (27.9 to 36.8). HHAR instead favors frozen probing, declining from 62.1 to 60.3, while FOCAL achieves the highest fine-tuned score at 61.3. The random frozen encoder reaches 26.2 mean macro-F1, indicating that pretraining improves recognition under the same readout. TransfHAR evaluates later portions of the same recordings; the other three datasets hold out participants.

Table 2: **Activity, gesture, and clinical recognition.** Macro-F1 and balanced accuracy (BA), reported as percentages with higher values better. Scores are evaluated per window for HAR, per clip for OpenWatch and per participant for PADS. Mean averages the four HAR datasets. Bold marks the best of six matched methods within each protocol, including ties. Gray rows show variants and external references excluded from ranking; ‡ denotes released NormWear. Task-specific references are retrained DeepConvLSTM for HAR, Mix-Token for OpenWatch and the PADS authors' BOSS pipeline with CatBoost.

	Activity recognition					Gesture	PADS	
Encoder / system	HHAR	MHEALTH	PRISM	TransfHAR	Mean	OpenWatch	F1	BA
Frozen probe								
InrtlWrist (Ours)	**62.1**	**68.6**	**27.9**	**72.7**	**57.8**	**64.7**	55.2	60.0
LIMU-BERT	61.2	67.6	25.1	71.1	56.2	61.7	50.5	58.4
SimMTM	57.5	61.4	26.7	66.8	53.1	55.9	48.7	59.3
FOCAL	44.9	63.0	21.9	47.0	44.2	48.1	40.7	53.9
RelCon	33.3	38.4	18.3	51.6	35.4	38.5	40.5	51.5
SSL-Wearables	59.2	66.3	27.6	67.4	55.1	54.7	**57.4**	**61.7**
InrtlWrist, plain attn.	60.2	68.1	26.5	73.0	56.9	63.9	48.1	59.7
InrtlWrist, accel. only	57.8	68.5	28.0	69.5	55.9	62.2	48.2	61.6
NormWear‡	43.7	56.1	19.2	33.7	38.2	43.8	56.6	59.7
Random frozen encoder	32.4	46.6	16.1	9.7	26.2	36.4	51.7	55.9
Raw input, 0 layers	5.3	1.4	15.6	0.1	5.6	11.4	12.6	33.3
Fine-tuning								
InrtlWrist (Ours)	60.3	**72.4**	**36.8**	**86.1**	**63.9**	54.2	**51.9**	**63.9**
LIMU-BERT	56.3	62.9	29.5	75.8	56.1	57.3	46.0	55.0
SimMTM	47.0	56.8	30.8	73.7	52.1	**61.9**	46.8	59.3
FOCAL	**61.3**	65.9	28.8	68.3	56.1	39.3	46.3	57.7
RelCon	42.9	60.3	28.9	79.0	52.8	42.7	49.1	58.1
SSL-Wearables	53.5	64.8	32.8	75.9	56.7	58.3	48.7	59.0
InrtlWrist, plain attn.	52.6	68.6	36.2	78.6	59.0	70.3	61.0	68.3
InrtlWrist, accel. only	56.8	68.3	29.8	77.2	58.0	60.6	59.9	67.7
NormWear‡	37.1	45.6	18.3	10.3	27.8	21.0	43.1	53.3
Same backbone (scratch)	48.8	63.2	27.8	73.5	53.3	60.8	29.9	45.5
Task-specific system	44.8	55.6	35.6	71.6	51.9	50.0	48.1	58.5

Gesture recognition. On OpenWatch, frozen InrtlWrist reaches 64.7 clip macro-F1, compared with 61.7 for LIMU-BERT, 60.8 for scratch and 36.4 for the random encoder. Fine-tuning reduces its score to 54.2, while SimMTM leads the matched methods at 61.9. Frozen InrtlWrist therefore remains ahead of every fine-tuned matched method on this task.

Clinical recognition. On PADS, frozen InrtlWrist reaches 55.2 participant macro-F1, below SSL-Wearables at 57.4. Although it exceeds training from scratch by 25.3 points, its advantage over the frozen random encoder is only 3.5 points, providing more limited evidence of a pretraining benefit than on activity and gesture recognition. Fine-tuning reduces macro-F1 to 51.9 while increasing balanced accuracy from 60.0 to 63.9. The retrained BOSS reference reaches 48.1 macro-F1 under the same participant aggregation (Varghese et al., 2024). All scores use participant-level evaluation; aggregation and class weighting are detailed in Appendix B.

5.2 Kinematic Estimation

Frozen InrtlWrist outperforms the random encoder on all three kinematic targets, but remains behind training from scratch on full-body pose and odometry (Table 3). Fine-tuning reduces full-body MPJPE from 97.9 to 88.1 mm and odometry ATE from 0.200 to 0.154 m, surpassing scratch at 92.7 mm and 0.162 m. It achieves the lowest errors among the six matched methods on all five fine-tuned metrics, with LIMU-BERT the closest alternative.

Arm pose changes less with fine-tuning. Frozen InrtlWrist and LIMU-BERT tie at 8.14 cm at the reported precision, compared with 8.21 cm for scratch. Fine-tuning reduces InrtlWrist's position error to 8.01 cm, while angular error changes from 9.89° to 9.86°, matching frozen LIMU-BERT. The frozen primary-metric ranking therefore does not extend to angular error.

Table 3: **Kinematic estimation on Nymeria.** Lower values are better for all metrics. Bold marks the best of six matched methods per column, including ties. Spanning cells report each reference under its stated training condition. The gray row reports retrained MobilePoser for pose and Air-IO for odometry, using their own inputs and training recipes. Velocity MSE is reported in units of 10^{-2}.

	Arm pose				Full-body pose		Odometry			
	Position (cm)		Angle (°)		MPJPE (mm)		ATE (m)		Vel. MSE (10^{-2})	
Encoder / system	Probing	Fine-tuning	Probing	Fine-tuning	Probing	Fine-tuning	Probing	Fine-tuning	Probing	Fine-tuning
InrtlWrist (Ours)	**8.14**	**8.01**	9.89	**9.86**	**97.9**	**88.1**	**0.200**	**0.154**	**2.732**	**1.824**
LIMU-BERT	**8.14**	8.18	**9.86**	9.97	98.5	89.7	0.208	0.158	2.856	1.885
SimMTM	8.47	8.44	10.27	10.29	103.6	92.9	0.224	0.163	3.064	1.946
FOCAL	8.71	8.66	10.70	10.69	108.4	95.9	0.240	0.177	3.334	2.218
RelCon	8.50	8.55	10.41	10.53	104.3	92.1	0.215	0.167	2.893	2.037
SSL-Wearables	8.42	8.57	10.14	10.58	101.7	90.6	0.220	0.166	3.078	2.030
Random encoder (frozen)	8.46		10.35		105.0		0.233		3.217	
Same backbone (scratch)	8.21		9.89		92.7		0.162		1.940	
Task-specific system	9.74		24.28		91.9		0.180		3.148	

The retrained MobilePoser and Air-IO references reach 91.9 mm body MPJPE and 0.180 m ATE, respectively, compared with 88.1 mm and 0.154 m for fine-tuned InrtlWrist. Their inputs and training recipes differ from the matched comparison. MobilePoser also receives wrist orientation, and the arm-pose evaluation supplies ancestor joints that it otherwise predicts. These results use session-held-out Nymeria recordings with recurring participants; odometry additionally uses ground-truth initial orientation and short evaluation windows.

5.3 Semantic Alignment

InrtlWrist achieves the highest mean IMU–language Recall@1 among the six matched methods under both protocols (Table 4). Frozen probing reaches 27.9, compared with 24.9 for LIMU-BERT, 23.2 for training from scratch and 9.5 for the random frozen encoder. Fine-tuning raises the mean to 38.8, with similar improvements in both retrieval directions.

Table 4: **IMU–language retrieval on HumanML3D.** Recall@1 (%↑), with I and T denoting IMU and text. Mean averages both retrieval directions over 200 sampled galleries of 100 candidates; the full gallery contains 3,701 clips.

	Frozen probing					Fine-tuning				
	Gallery of 100			Full gallery		Gallery of 100			Full gallery	
Encoder / system	Mean	I→T	T→I	I→T	T→I	Mean	I→T	T→I	I→T	T→I
InrtlWrist (Ours)	**27.9**	**28.3**	**27.5**	**3.6**	**3.4**	**38.8**	**39.5**	**38.1**	**5.1**	4.4
LIMU-BERT	24.9	25.3	24.5	3.1	2.5	36.6	37.2	36.0	4.4	3.9
SimMTM	17.2	17.4	17.0	1.7	1.2	32.4	32.7	32.1	4.0	3.4
FOCAL	8.6	9.1	8.0	0.7	0.7	22.4	22.7	22.0	2.1	1.9
RelCon	13.2	13.9	12.4	1.0	1.0	30.0	30.5	29.5	3.2	2.9
SSL-Wearables	24.8	25.6	24.0	2.5	2.5	34.5	34.9	34.1	4.2	**4.5**
InrtlWrist, plain attn.	26.3	26.7	26.0	3.4	2.4	37.0	37.7	36.3	4.6	4.1
InrtlWrist, accel. only	20.7	21.0	20.3	2.1	1.7	31.5	31.7	31.3	3.6	3.3
NormWear‡	15.5	15.9	15.2	1.4	1.2	15.7	15.8	15.6	1.0	1.3
Untrained frozen controls										
Random frozen encoder	9.5	10.0	9.1	0.6	0.4	–	–	–	–	–
Raw input, 0 layers	0.5	1.0	0.0	0.0	0.0	–	–	–	–	–
From-scratch references										
Same backbone, scratch	–	–	–	–	–	23.2	23.6	22.8	2.1	2.1
UniMTS (wrist-only)	–	–	–	–	–	13.2	24.9	1.5	2.3	0.0

The advantage over scratch persists in the full gallery, where fine-tuned InrtlWrist reaches 5.1 IMU-to-text and 4.4 text-to-IMU Recall@1, compared with 2.1 in both directions for scratch. SSL-Wearables achieves the highest text-to-IMU score at 4.5. Although absolute recall falls with the larger candidate set, InrtlWrist remains ahead of scratch in both directions and under both protocols.

The wrist-only UniMTS reimplementation, trained from scratch, reaches 24.9 IMU-to-text Recall@1 but 1.5 in the reverse direction on galleries of 100. This asymmetry is associated with retrieval hubness, detailed in Appendix B. All encoders receive paired motion–text supervision during downstream training, so these comparisons evaluate the pretrained representation's suitability for learning alignment rather than zero-shot language alignment.

6 Cross-Task Transferability

We evaluate whether supervision on one task produces representations useful for another. Each source encoder is trained from scratch on one of seven tasks, then evaluated on the other six with a new decoder under the benchmark's target splits and budgets (Figure 3). Frozen transfer is compared with a random encoder under the same probe, while transfer with fine-tuning is compared with target training from scratch. InrtlWrist, labeled MAE by its pretraining objective, provides the self-supervised reference.

	Random	MAE	HAR	Arm	Body	Gest.	Odom.	PD	Text
HAR F1 % ↑	16.1	27.9	–	27.0	27.4	21.7	25.0	21.3	25.1
Arm cm ↓	8.46	8.14	8.55	–	8.13	8.52	8.26	8.53	8.54
Body mm ↓	104.99	97.85	103.49	95.99	–	103.91	99.92	104.33	103.31
Gesture F1 % ↑	36.4	64.7	32.9	44.6	41.6	–	43.4	39.2	46.0
Odom. ATE m ↓	.2334	.1999	.2252	.1925	.1865	.2308	–	.2255	.2271
PD subj F1 % ↑	51.7	55.2	47.5	30.8	38.0	49.1	34.3	–	47.3
Text R@1 % ↑	9.5	27.9	8.3	11.1	11.8	9.3	11.7	9.4	–

(a) Frozen transfer.

	Scratch	MAE	HAR	Arm	Body	Gest.	Odom.	PD	Text
HAR F1 % ↑	27.8	36.8	–	32.4	32.7	27.5	32.4	29.7	29.2
Arm cm ↓	8.21	8.01	8.56	–	8.27	8.64	8.12	8.41	8.50
Body mm ↓	92.72	88.11	92.39	89.82	–	92.78	87.89	92.89	92.11
Gesture F1 % ↑	60.8	54.2	33.5	35.2	36.0	–	47.5	40.1	48.6
Odom. ATE m ↓	.1618	.1541	.1672	.1715	.1626	.1648	–	.1638	.1647
PD subj F1 % ↑	29.9	51.9	58.5	48.3	49.0	42.4	38.7	–	43.5
Text R@1 % ↑	23.2	38.8	24.9	29.4	30.1	25.7	30.6	25.8	–

(b) Transfer with fine-tuning.

Figure 3: **Cross-task transfer.** Source encoders (columns) are evaluated on each target (rows). Cells give raw scores; color indicates relative improvement over the first column, using a frozen random encoder in (a) and target training from scratch in (b). Green indicates improvement and orange-pink degradation, with color saturated at $\pm 40\%$. MAE denotes InrtlWrist, frozen in (a) and fine-tuned in (b). Dashes omit same-task evaluations. HAR denotes PRISM; PADS uses participant macro-F1.

Frozen transfer. Supervised source encoders outperform a random frozen encoder in 28 of 42 source–target pairs, including 15 gains above 10% (Figure 3a). Kinematic targets benefit most from other kinematic sources: full-body supervision, for example, reduces odometry ATE from 0.233 to 0.186 m. Because these tasks share Nymeria participants, we repeat the comparison on disjoint participant thirds; five of six kinematic transfers remain positive at roughly half their original magnitude, while MAE leads on full-body pose and odometry (Appendix E). All six sources raise PRISM macro-F1 from 16.1 to 21.3–27.4, near MAE's 27.9, but fall below the random encoder on PADS. MAE also leads on gesture and IMU–text (64.7 F1 and 27.9 Recall@1, versus at most 46.0 and 11.8 for supervised sources). Overall, 35 of 42 frozen transfers remain below target training from scratch; the exceptions are full-body-to-arm transfer and the six PADS transfers, for which scratch reaches only 29.9 participant F1.

Transfer with fine-tuning. With target fine-tuning, 22 source–target pairs improve over scratch and 20 decline (Figure 3b). Every source improves IMU–text and PADS, with PRISM supervision reaching 58.5 PADS F1 against 51.9 for direct MAE fine-tuning. Pose and odometry sources improve PRISM by 16–18% relative to scratch. Four of six sources also improve full-body pose; odometry initialization reaches 87.9 mm MPJPE, compared with 88.1 mm for MAE. No source improves odometry, only odometry supervision improves arm pose, and every source reduces gesture F1.

7 Architecture and Training Ablations

Decoder capacity. We compare frozen pretrained and random encoders as decoder capacity increases (Figure 4). Pretraining reduces body MPJPE by 9.7 mm with a linear readout and 7.1 mm with the benchmark decoder. Increasing decoder width changes this gap little, whereas adding temporal layers reduces it below 1 mm. A 12-layer odometry decoder similarly allows the random encoder to outperform every pretrained encoder (Appendix D). Stronger decoders therefore narrow or reverse the measured advantage of pretraining, motivating the light readouts used in the benchmark.

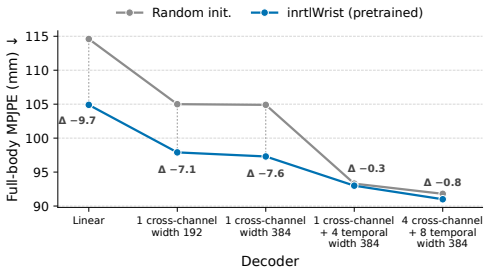

Figure 4: **Decoder capacity.** Frozen-encoder MPJPE (mm↓). Δ denotes pretrained minus random.

Pretraining data and motion coverage. Holding the architecture, probes and training budget fixed, we vary the amount and coverage of pretraining data. We sample uniformly from a five-source pool of 1,959,288 windows or restrict sampling to subsets of 1,000 motion clusters obtained by k-means on a separately trained MAE's features. At 50,000 updates, the 0.01% subsets trail the 0.1% subsets at every coverage level, while uniform draws from 0.1% to 100% of the pool perform alike. Increasing motion coverage generally improves performance, particularly when few clusters are available (Figure 5). These findings are conditional on the training budget and sampling scheme, since restricting clusters also changes source composition. Appendix H reports the full grid, normalization and probe protocols.

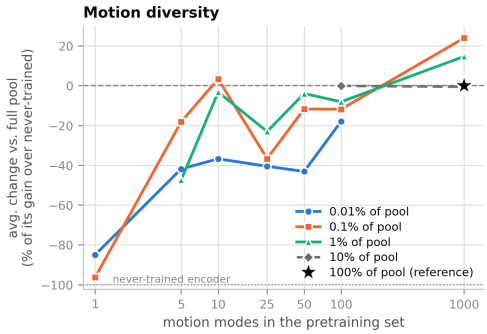

Figure 5: **Motion coverage.** Eight-probe mean after 50,000 updates. Per-probe changes from the full pool are normalized by its gain over random and clipped to ±100%.

Joint task supervision. We also compare single-task fine-tuning with joint supervision of full-body pose and either odometry or microgesture. Joint training improves odometry ATE from 0.154 to 0.149 m and gesture F1 from 54.2 to 65.0, but increases body MPJPE and reduces frozen IMU–text Recall@1. Gains on the partner tasks therefore accompany losses on both the jointly trained pose task and language retrieval. Appendix F reports the schedules and per-task checkpoint selection.

8 Conclusion

We present OmniIMU-Bench, ten wrist-IMU tasks evaluated under shared protocols, and the pretrained reference encoder InrtlWrist. Among the six matched methods, InrtlWrist achieves the best or joint-best frozen primary score on nine tasks and outperforms training from scratch on nine after fine-tuning. Source-task supervision transfers between kinematic tasks; on IMU–text retrieval no frozen supervised source approaches the pretrained encoder. Joint training improves the partner tasks at a cost to full-body pose and language retrieval. Decoder capacity can obscure differences between frozen representations. At the tested pretraining budget, additional broadly sampled windows provide no consistent downstream gain. These findings distinguish broad downstream usefulness from mutually beneficial task supervision.

Limitations. Our scaling study does not isolate motion coverage from data volume and source composition. It also does not establish how downstream performance scales with participant count, the balance of recorded and simulated data, or sensor placement. The benefit of larger corpora at longer training budgets remains unresolved.

Reproducibility statement

Every number in this paper comes from a registry-driven harness that we release with the benchmark: one task card per task fixing its inputs, decoder, split rule, metric and budget, and one encoder card per model. Appendix B gives each task's split rule, metric and subject counts, Appendix A the backbone, decoders and optimization settings. Appendix C records the corpus-against-evaluation-set audit, including the duplicated participants and mirrored recordings we had to merge before splitting. Code, task cards, splits and the per-cell result files will be released.

Statement on the use of AI

In this work, we used Anthropic's Claude through Claude Code to develop and review code, clean and reformat datasets, suggest experimental controls and parameters, analyze results, prepare figures and tables, and assist with the paper's structure, wording, literature review and references. It also provided feedback on the methodology and helped refine hypotheses. We did not use generative AI to produce synthetic datasets, translate text or conduct qualitative analysis; the simulated wrist IMU was computed from motion capture using a physical model. The authors defined the tasks and protocols, selected the experiments, reviewed the code and results, revised the text, and take responsibility for the final paper.

References

Salar Abbaspourazad, Oussama Elachqar, Andrew Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pC3WJHf51j.

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C. Fowlkes, Stefano Soatto, and Pietro Perona. Task2Vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6429–6438, 2019.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2016. ICLR 2017 Workshop Track.

Riku Arakawa, Hiromu Yakura, Vimal Mollyn, Suzanne Nie, Emma Russell, Dustin P. DeMeo, Haarika A. Reddy, Alexander K. Maytin, Bryan T. Carroll, Jill Fain Lehman, and Mayank Goel. PrISM-Tracker: A framework for multimodal procedure tracking using wearable sensors and state transition information with user-driven handling of errors and uncertainty. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4):1–27, 2022. doi: 10.1145/3569504.

Oresti Banos, Rafael Garcia, Juan A. Holgado-Terriza, Miguel Damas, Hector Pomares, Ignacio Rojas, Alejandro Saez, and Claudia Villalonga. mHealthDroid: A novel framework for agile development of mobile health applications. In *Ambient Assisted Living and Daily Activities (IWAAL 2014)*, volume 8868 of *Lecture Notes in Computer Science*, pp. 91–98. Springer, 2014. doi: 10.1007/978-3-319-13105-4_14.

Sarnab Bhattacharya, Rebecca Adaimi, and Edison Thomaz. Leveraging sound and wrist motion to detect activities of daily living with commodity smartwatches. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(2), 2022. doi: 10.1145/3534582.

Pietro Bonazzi, Youssef Ahmed, Daniel Eckert, Andrea Ronco, Junjie Zeng, Dengxin Dai, and Michele Magno. OpenWatch: A multimodal benchmark for hand gesture recognition on smartwatches. *arXiv preprint arXiv:2605.04791*, 2026.

Bones Studio. BONES-SEED: Skeletal everyday embodiment dataset. Dataset release, https://bones.studio/datasets/seed, 2025.

Brian M. Bot, Christine Suver, Elias Chaibub Neto, Michael Kellen, Arno Klein, Christopher Bare, Megan Doerr, Abhishek Pratap, John Wilbanks, E. Ray Dorsey, Stephen H. Friend, and Andrew D.

Trister. The mPower study, Parkinson disease mobile data collected using ResearchKit. *Scientific Data*, 3:160011, 2016. doi: 10.1038/sdata.2016.11.

Aidan Bradshaw, Riku Arakawa, Xin Liu, and Karan Ahuja. TransfHAR: Self-supervised wrist representations for on-demand activity recognition. *arXiv preprint arXiv:2608.15861*, 2026.

Robin Burchard, Kristina Kirsten, Marcel Miché, Philipp Scholl, Bert Arnrich, Kristof Van Laerhoven, Roselind Lieb, and Karina Wahl. A real-world dataset for detecting handwashing in daily life using wrist motion data from wearables. *Scientific Data*, 13(1):179, 2026. doi: 10.1038/s41597-026-06698-2.

Yize Cai, Rui Feng, Anlan Yu, Baoshen Guo, and Zhiqing Hong. BenchHAR: Benchmarking self-supervised learning for generalizable sensor-based activity recognition. *arXiv preprint arXiv:2605.08296*, 2026.

Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

Changhao Chen, Xiaoxuan Lu, Andrew Markham, and Niki Trigoni. IONet: Learning to cure the curse of drift in inertial odometry. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.12102.

Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *IEEE International Conference on Image Processing (ICIP)*, pp. 168–172, 2015. doi: 10.1109/ICIP.2015.7350781.

Arnav M. Das, Chi Ian Tang, Fahim Kawsar, and Mohammad Malekzadeh. PRIMUS: Pretraining IMU encoders with multimodal self-supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025. doi: 10.1109/ICASSP49660.2025.10888874.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M. Alabdulmohsin, Avital Oliver, Piotr Padlewski, Alexey Gritsenko, Mario Lucic, and Neil Houlsby. Patch n' pack: NaViT, a vision transformer for any aspect ratio and resolution. In *Advances in Neural Information Processing Systems*, volume 36, pp. 2252–2274, 2023.

Jiaxiang Dong, Haixu Wu, Haoran Zhang, Li Zhang, Jianmin Wang, and Mingsheng Long. SimMTM: A simple pre-training framework for masked time-series modeling. In *Advances in Neural Information Processing Systems*, volume 36, pp. 29996–30025, 2023. doi: 10.52202/075280-1306.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16115–16152, 2024.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3D human motions from text. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5142–5151, 2022. doi: 10.1109/CVPR52688.2022.00509.

Harish Haresamudram, Apoorva Beedu, Varun Agrawal, Patrick L. Grady, Irfan Essa, Judy Hoffman, and Thomas Plötz. Masked reconstruction based self-supervision for human activity recognition. In *Proceedings of the 2020 International Symposium on Wearable Computers*, pp. 45–49, 2020. doi: 10.1145/3410531.3414306.

Harish Haresamudram, Irfan Essa, and Thomas Plötz. Assessing the state of self-supervised human activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(3):116:1–116:47, 2022. doi: 10.1145/3550299.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, 2022. doi: 10.1109/CVPR52688.2022.01553.

Sachini Herath, Hang Yan, and Yasutaka Furukawa. RoNIN: Robust neural inertial navigation in the wild: Benchmark, evaluations, & new methods. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3146–3152, 2020. doi: 10.1109/ICRA40945.2020.9196860.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

Zhiqing Hong, Zelong Li, Shuxin Zhong, Wenjun Lyu, Haotian Wang, Yi Ding, Tian He, and Desheng Zhang. CrossHAR: Generalizing cross-dataset human activity recognition via hierarchical self-supervised pretraining. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(2), 2024. doi: 10.1145/3659597.

Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J. Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics*, 37(6):185:1–185:15, 2018. doi: 10.1145/3272127.3275108.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The Platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642, 2024.

Yifeng Jiang, Yuting Ye, Deepak Gopinath, Jungdam Won, Alexander W. Winkler, and C. Karen Liu. Transformer inertial poser: Real-time human motion reconstruction from sparse IMUs with simultaneous terrain generation. In *SIGGRAPH Asia 2022 Conference Papers*, 2022. doi: 10.1145/3550469.3555428.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, 2019.

Jennifer R. Kwapisz, Gary M. Weiss, and Samuel A. Moore. Activity recognition using cell phone accelerometers. *ACM SIGKDD Explorations Newsletter*, 12(2):74–82, 2011. doi: 10.1145/1964897.1964918.

Gierad Laput, Robert Xiao, and Chris Harrison. ViBand: High-fidelity bio-acoustic sensing using commodity smartwatch accelerometers. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pp. 321–333, 2016. doi: 10.1145/2984511.2984582.

Yuyang Leng, Renyuan Liu, Shaohan Hu, Peijun Zhao, Chun-Fu Chen, Songqing Chen, and Shuochao Yao. Physical self-supervised learning: IMU sensing without manual labels. In *Proceedings of the 24th Annual International Conference on Mobile Systems, Applications and Services (MobiSys)*, pp. 1011–1025, 2026. doi: 10.1145/3745756.3809252.

Zikang Leng, Amitrajit Bhattacharjee, Hrudhai Rajasekhar, Lizhe Zhang, Elizabeth Bruda, Hyeokhyen Kwon, and Thomas Plötz. IMUGPT 2.0: Language-based cross modality transfer for sensor-based human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(3), 2024. doi: 10.1145/3678545.

Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. In *Advances in Neural Information Processing Systems*, volume 36, pp. 25268–25280, 2023.

Shengzhong Liu, Tomoyoshi Kimura, Dongxin Liu, Ruijie Wang, Jinyang Li, Suhas Diggavi, Mani Srivastava, and Tarek Abdelzaher. FOCAL: Contrastive learning for multimodal time-series sensing signals in factorized orthogonal latent space. In *Advances in Neural Information Processing Systems*, volume 36, pp. 47309–47338, 2023. doi: 10.52202/075280-2049.

Wenxin Liu, David Caruso, Eddy Ilg, Jing Dong, Anastasios I. Mourikis, Kostas Daniilidis, Vijay Kumar, and Jakob Engel. TLIO: Tight learned inertial odometry. *IEEE Robotics and Automation Letters*, 5(4):5653–5660, 2020. doi: 10.1109/LRA.2020.3007421.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2024.

Yunfei Luo, Yuliang Chen, Asif Salekin, and Tauhidur Rahman. Toward foundation model for multivariate wearable sensing of physiological signals. *ACM Transactions on Computing for Healthcare*, 7(3), 2026. doi: 10.1145/3803808.

Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, Kevin Bailey, David Soriano Fosas, C. Karen Liu, Ziwei Liu, Jakob Engel, Renzo De Nardi, and Richard Newcombe. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *Computer Vision – ECCV 2024*, volume 15082 of *Lecture Notes in Computer Science*, pp. 445–465. Springer, 2024. doi: 10.1007/978-3-031-72691-0_25.

Vimal Mollyn, Karan Ahuja, Dhruv Verma, Chris Harrison, and Mayank Goel. SAMoSA: Sensing activities with motion and subsampled audio. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(3), 2022. doi: 10.1145/3550284.

Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. 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*, pp. 1–12, 2023. doi: 10.1145/3544548.3581392.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Aparajita Saraf, Amy Bearman, and Babak Damavandi. IMU2CLIP: Language-grounded motion sensor translation with multimodal contrastive learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 13246–13253. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.883.

Dan Morris, T. Scott Saponas, Andrew Guillory, and Ilya Kelner. RecoFit: Using a wearable sensor to find, recognize, and count repetitive exercises. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 3225–3234, 2014. doi: 10.1145/2556288.2557116.

Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam A. Tailor, Jacob Sunshine, Yun Liu, Tim Althoff, Shrikanth Narayanan, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Samy Abdel-Ghaffar, and Daniel McDuff. Scaling wearable foundation models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=yb4QE6b22f.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

Francisco Javier Ordóñez and Daniel Roggen. Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016. doi: 10.3390/s16010115.

Junyong Park, Oron Levy, Rebecca Adaimi, Asaf Liberman, Gierad Laput, and Abdelkareem Bedri. SlotFM: A motion foundation model with slot attention for diverse downstream tasks. *arXiv preprint arXiv:2509.21673*, 2025.

Plato. *The Republic of Plato*. Basic Books, 2 edition, 1991. Translated by Allan Bloom.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Yuheng Qiu, Can Xu, Yutian Chen, Shibo Zhao, Junyi Geng, and Sebastian Scherer. AirIO: Learning inertial odometry with enhanced IMU feature observability. *IEEE Robotics and Automation Letters*, 10(9):9368–9375, 2025. doi: 10.1109/LRA.2025.3581130.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *16th International Symposium on Wearable Computers (ISWC)*, pp. 108–109, 2012. doi: 10.1109/ISWC.2012.13.

Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. Multi-task self-supervised learning for human activity detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(2):1–30, 2019. doi: 10.1145/3328932.

Patrick Schäfer. The BOSS is concerned with time series classification in the presence of noise. *Data Mining and Knowledge Discovery*, 29(6):1505–1530, 2015.

Muhammad Shoaib, Stephan Bosch, Ozlem Durmaz Incel, Hans Scholten, and Paul J. M. Havinga. Complex human activity recognition using smartphone and wrist-worn motion sensors. *Sensors*, 16(4):426, 2016. doi: 10.3390/s16040426.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 19523–19536, 2022.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9120–9132, 2020.

Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen. Smart devices are different: Assessing and mitigating mobile sensing heterogeneities for activity recognition. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems (SenSys)*, pp. 127–140, 2015. doi: 10.1145/2809695.2809718.

Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, Soren Brage, Nick Wareham, and Cecilia Mascolo. SelfHAR: Improving human activity recognition through self-training with unlabeled data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5 (1), 2021. doi: 10.1145/3448112.

Prithviraj Tarale, Kiet Chu, Abhishek Varghese, Kai-Chun Liu, Maxwell A. Xu, Mohit Iyyer, and Sunghoon I. Lee. Bio-inspired self-supervised learning for wrist-worn accelerometer data. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Terry T. Um, Franz M. J. Pfister, Daniel Pichler, Satoshi Endo, Muriel Lang, Sandra Hirche, Urban Fietzek, and Dana Kulić. Data augmentation of wearable sensor data for Parkinson's disease monitoring using convolutional neural networks. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 216–220, 2017. doi: 10.1145/3136755.3136817.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Julian Varghese, Alexander Brenner, Michael Fujarski, Catharina Marie van Alen, Lucas Plagwitz, and Tobias Warnecke. Machine learning in the Parkinson's disease smartwatch (PADS) dataset. *npj Parkinson's Disease*, 10(1):9, 2024. doi: 10.1038/s41531-023-00625-7.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. Exploring and predicting transferability across NLP tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7882–7926, 2020. doi: 10.18653/v1/2020.emnlp-main.635.

Chunzhuo Wang, T. Sunil Kumar, Walter De Raedt, Guido Camps, Hans Hallez, and Bart Vanrumste. Eating speed measurement using wrist-worn IMU sensors towards free-living environments. *IEEE Journal of Biomedical and Health Informatics*, 28(10):5816–5828, 2024a. doi: 10.1109/JBHI.2024.3422875.

Chunzhuo Wang, T. Sunil Kumar, Walter De Raedt, Guido Camps, Hans Hallez, and Bart Vanrumste. Eat-radar: Continuous fine-grained intake gesture detection using FMCW radar and 3d temporal convolutional network with attention. *IEEE Journal of Biomedical and Health Informatics*, 28(2): 1000–1011, 2024b. doi: 10.1109/JBHI.2023.3339703.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025.

Gary M. Weiss, Kenichi Yoneda, and Thaier Hayajneh. Smartphone and smartwatch-based biometrics using activities of daily living. *IEEE Access*, 7:133190–133202, 2019. doi: 10.1109/ACCESS.2019.2940729.

Hongyi Wen, Julian Ramos Rojas, and Anind K. Dey. Serendipity: Finger gesture recognition using an off-the-shelf smartwatch. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 3847–3851, 2016. doi: 10.1145/2858036.2858466.

Huatao Xu, Pengfei Zhou, Rui Tan, Mo Li, and Guobin 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)*, pp. 220–233, 2021. doi: 10.1145/3485730.3485937.

Maxwell A. Xu, Jaya Narain, Gregory Darnell, Haraldur T. Hallgrimsson, Hyewon Jeong, Darren Forde, Richard Fineman, Karthik J. Raghuram, James M. Rehg, and Shirley You Ren. RelCon: Relative contrastive learning for a motion foundation model for wearable data. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=k2uUeLCrQq.

Vasco Xu, Chenfeng Gao, Henry Hoffmann, and Karan Ahuja. MobilePoser: Real-time full-body pose estimation and 3D human translation from IMUs in mobile consumer devices. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pp. 1–11, 2024. doi: 10.1145/3654777.3676461.

Zongzhe Xu, Aakarsh Anand, Sarah Jiang, Chuntung Zhuang, Zitao Shuai, Sriram Sankararaman, and Yuzhe Yang. Inertia-1: An open exploration of wearable motion foundation models. *arXiv preprint arXiv:2607.06617*, 2026.

Xinyu Yi, Yuxiao Zhou, and Feng Xu. TransPose: Real-time 3D human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics*, 40(4):1–13, 2021. doi: 10.1145/3450626.3459786.

Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. 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 (CVPR)*, 2022. doi: 10.1109/CVPR52688.2022.01282.

Hang Yuan, Shing Chan, Andrew P. Creagh, Catherine Tong, Aidan Acquah, David A. Clifton, and Aiden Doherty. Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. *npj Digital Medicine*, 7(1):91, 2024. doi: 10.1038/s41746-024-01062-3.

Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3712–3722, 2018.

Xiyuan Zhang, Diyan Teng, Ranak Roy Chowdhury, Shuheng Li, Dezhi Hong, Rajesh K. Gupta, and Jingbo Shang. UniMTS: Unified pre-training for motion time series. In *Advances in Neural Information Processing Systems*, volume 37, pp. 107469–107493, 2024. doi: 10.52202/079017-3414.

Yuhong Zhang, Jing Lin, Ailing Zeng, Guanlin Wu, Shunlin Lu, Yurong Fu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x++: A large-scale multimodal 3d whole-body human motion dataset, 2025a.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2023.

Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A. Ali Heydari, Girish Narayanswamy, Max Xu, Ahmed Metwally, Jinhua Xu, Jake Garrison, Xuhai Xu, Tim Althoff, Yun Liu, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Cecilia Mascolo, Xin Liu, Daniel McDuff, and Yuzhe Yang. SensorLM: Learning the language of wearable sensors. In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

A Implementation Details

Harness and provenance. Every number in this paper is produced by one harness from a task card (data, split, input window, decoder, recipe, selection rule, scorer) and an encoder card (checkpoint, architecture, input normalization, pretraining corpus). Each result file records the run, the selected checkpoint step and the scorer, and every table is generated from these files by a script released with the benchmark.

Input. All tasks use six channels (accelerometer in m/s^2, gyroscope in rad/s) from one wrist in a common forward–left–up (FLU) wrist frame. Each window is z-scored per channel before it enters any encoder, in pretraining and downstream alike. During pretraining only, every window is rotated by one of the 24 proper 90°-multiple rotations of both sensors (Um et al., 2017), drawn afresh at each visit; these rotations permute and negate axes, so they commute with per-channel z-scoring. Downstream training uses no augmentation except IMU–text alignment, which applies random 3-D rotations for every encoder alike.

Pretraining. The self-supervised encoders share one 12-layer, 384-wide ViT over 0.32 s patches (21 M parameters), one corpus and one optimizer recipe; they differ in the objective and, for the five external objectives, in using plain rather than alternating frame/global attention.

Controls. The *random frozen encoder* has the InrtlWrist architecture with random initialization (seed 42) and is never trained. The *raw-input* control is the same model with zero transformer layers, so the decoder reads a frozen random 384-dimensional linear projection of each 16-sample patch plus position embeddings, which preserves the input patch. Both are probed with exactly the decoders, recipes and selection rules of the other rows.

Compute. Every pretraining and downstream run uses a single NVIDIA A100 or H100 GPU.

B Tasks and Metrics

Splits and subjects. Evaluation is subject-disjoint for microgesture (43 OpenWatch subjects, canonical identities after the duplicated participants of Table 12 are merged), Parkinson's disease (469 PADS participants) and HHAR, MHEALTH and PRISM (9, 10 and 31 subjects); session-disjoint for the three Nymeria tasks, whose 235 participants recur across sessions; the dataset's own split for IMU–text retrieval; and split in time for TransfHAR (10 subjects), because most of its 60 activities were performed by a single participant (Appendix B.1).

Pose estimation (Nymeria). Input is the real right-wrist IMU of Nymeria, output one set of 6-D joint rotations per 0.32 s patch of a 10.24 s window. The arm task predicts five right-arm joints (collar, shoulder, elbow, wrist, hand), scored by the forward-kinematics position error of these joints with ground-truth ancestors, in cm; the full-body task predicts the 23 SMPL body joints, scored by root-relative MPJPE in mm. Nymeria is split by recording session (867 training and 101 evaluation sessions). All 82 participants of the evaluation sessions also appear in training, in 271 of the 867

Table 5: Arm pose: Nymeria real right-wrist IMU → 5-joint right-arm chain rotations. One training seed per cell. Task-specific systems are listed for orientation and are not ranked (Appendix B). MobilePoser and IMUPoser also read wrist orientation, and must predict the ancestor joints that our arm-pose metric supplies.

Encoder	pos err (cm)↓	ang err (°)↓
Frozen probe		
InrtlWrist (Ours)	8.14	9.89
LIMU-BERT	8.14	9.86
SimMTM	8.47	10.27
FOCAL	8.71	10.70
RelCon	8.50	10.41
SSL-Wearables	8.42	10.14
InrtlWrist, plain attention	8.36	10.17
InrtlWrist, accelerometer only	8.76	10.61
Random frozen encoder	8.46	10.35
Raw input, 0 layers	8.51	10.30
Fine-tune		
InrtlWrist (Ours)	8.01	9.86
LIMU-BERT	8.18	9.97
SimMTM	8.44	10.29
FOCAL	8.66	10.69
RelCon	8.55	10.53
SSL-Wearables	8.57	10.58
InrtlWrist, plain attention	8.19	10.06
InrtlWrist, accelerometer only	8.88	10.80
Scratch	8.21	9.89
Task-specific systems (not ranked)		
MobilePoser (Xu et al.)	9.74	24.28
IMUPoser (Mollyn et al.)	10.15	23.82

training sessions, so the pose and odometry tasks measure generalization to new recordings of seen people; we keep the session split (a participant-disjoint one would remove 31 % of the training sessions) and state its scope.

Inertial odometry (Nymeria). The model predicts pelvis velocity in the wrist frame, four values per 0.32 s patch of a 2.56 s window. For scoring, each predicted velocity is rotated to the world frame with the ground-truth wrist orientation at the window start propagated by integrating the raw gyroscope, the rotated velocities are summed into a trajectory, and the absolute trajectory error (m) is averaged over all evaluation windows. This uses the true initial orientation and is easier than full inertial navigation; numbers are not comparable to world-frame odometry systems.

Microgesture (OpenWatch). Five gestures and a background class from 1 s windows at 100 Hz, six windows per clip from a centered 2 s interval, zero-padded to 128 samples. A clip is labeled by the first run of three identical window predictions, otherwise by majority. The subject split is computed over the 43 canonical identities (Appendix C), stratified by posture (33/5/5 identities; 102 validation and 87 test clips); the test split is reported.

IMU–text retrieval (HumanML3D). Right-wrist IMU is simulated from the HumanML3D motions and aligned with the CLIP ViT-B/32 text space. Recall@1 is the mean of IMU→text and text→IMU retrieval over 200 random galleries of 100 test clips with the first caption of each clip, on the 3,701 of the official test split's 4,384 clip ids (2,192 motions and their mirrors, Appendix C) that have simulated wrist IMU and captions in the UniMTS release, carrying 2,721 distinct first captions.

Parkinson's disease (PADS). Each of the 469 participants is healthy, has Parkinson's disease or a differential diagnosis. PADS records both wrists during eleven scripted assessment steps; each (participant, step, wrist) recording is its own single-IMU session, 22 per participant, resampled to

Table 6: Full-body pose: Nymeria real right-wrist IMU → 23-joint SMPL body (root excluded). One training seed per cell. Task-specific systems are listed for orientation and are not ranked (Appendix B). MobilePoser and IMUPoser also read wrist orientation.

Encoder	MPJPE (mm)↓
Frozen probe	
InrtlWrist (Ours)	97.85
LIMU-BERT	98.55
SimMTM	103.65
FOCAL	108.41
RelCon	104.29
SSL-Wearables	101.67
InrtlWrist, plain attention	99.96
InrtlWrist, accelerometer only	102.79
Random frozen encoder	104.99
Raw input, 0 layers	106.36
Fine-tune	
InrtlWrist (Ours)	88.11
LIMU-BERT	89.74
SimMTM	92.91
FOCAL	95.92
RelCon	92.15
SSL-Wearables	90.61
InrtlWrist, plain attention	88.38
InrtlWrist, accelerometer only	94.42
Scratch	92.72
Task-specific systems (not ranked)	
MobilePoser (Xu et al.)	91.88
IMUPoser (Mollyn et al.)	96.87

50 Hz, so the encoder never sees the two wrists together. Windows of 2.56 s are classified independently and a participant's prediction is the argmax of the mean posterior over all of that participant's windows; the metric is the macro-F1 over the three classes on a subject-disjoint 80/20 split (375/94 participants; the 94 are 14 healthy, 58 Parkinson's and 22 differential). Every classification head in the benchmark is trained with inverse-frequency class-weighted cross-entropy; it matters most here because a participant's prediction averages about 180 window posteriors, which leaves whatever prior the loss wrote into the head, so an unweighted head assigns all 94 participants to the majority class (25.4 subject macro-F1). The subject metric is the clinically meaningful unit but a coarse one: the never-trained frozen control reaches 51.7, above six of the nine pretrained encoders, so encoders within a few points of each other are not ranked by it. Section 5 therefore draws its Parkinson's conclusion from the comparison with the dataset authors' pipeline, on which the window and subject units agree. That row is BOSS histograms (Schäfer, 2015) with the authors' gradient-boosted classifier (CatBoost (Prokhorenkova et al., 2018)), the smartwatch option their paper ranks best (Varghese et al., 2024) (48.1 subject macro-F1, 58.5 balanced accuracy); it differs from the publication in classifying single-wrist 2.56 s windows with per-participant averaging rather than one sample per participant on binary tasks.

Activity recognition. HHAR (6 classes), MHEALTH (12) and PRISM (5 global ADL classes; the 96 fine-grained procedure steps of PrISM-Tracker are a separate, much harder task we do not include) use subject-disjoint splits with 20 % of subjects held out, leaving 2, 2 and 6 evaluation subjects (2,661, 1,049 and 15,381 windows); TransfHAR (60 classes) is split in time. All use 2.56 s windows and report window-level macro-F1 pooled over the evaluation split; Section 5 also reports the mean over the four sets, and Table 11 gives every metric of every encoder on them.

Task-specific systems. Published systems (MobilePoser and IMUPoser (Mollyn et al., 2023) for pose, TLIO and Air-IO for odometry, Mix-Token (Bonazzi et al., 2026) and the first-party NormWear-Base result (Luo et al., 2026) for OpenWatch gestures, UniMTS for IMU–text, Deep-

Table 7: Odometry (wrist frame): Nymeria real right-wrist IMU → pelvis velocity in the wrist (FLU) frame. Task-specific systems are listed for orientation and are not ranked (Appendix B).

Encoder	ATE (m)↓	vel MSE↓
Frozen probe		
InrtlWrist (Ours)	0.1999	0.02732
LIMU-BERT	0.2078	0.02856
SimMTM	0.2240	0.03064
FOCAL	0.2404	0.03334
RelCon	0.2153	0.02893
SSL-Wearables	0.2205	0.03078
InrtlWrist, plain attention	0.2104	0.02892
InrtlWrist, accelerometer only	0.2296	0.03232
Random frozen encoder	0.2334	0.03217
Raw input, 0 layers	0.2558	0.03720
Fine-tune		
InrtlWrist (Ours)	0.1541	0.01824
LIMU-BERT	0.1576	0.01885
SimMTM	0.1629	0.01946
FOCAL	0.1767	0.02218
RelCon	0.1672	0.02037
SSL-Wearables	0.1656	0.02030
InrtlWrist, plain attention	0.1569	0.01880
InrtlWrist, accelerometer only	0.1898	0.02558
Scratch	0.1618	0.01940
Task-specific systems (not ranked)		
TLIO	0.1874	0.02978
Air-IO	0.1803	0.03148

Table 8: Microgesture recognition on OpenWatch (five gestures plus background, real recordings, held-out test split), our encoder against training from scratch and the task-specific systems.

Encoder	clip F1 (test) (%)↑	clip acc (test) (%)↑	win F1 (test) (%)↑	win acc (test) (%)↑
Frozen probe				
InrtlWrist (Ours)	64.7	67.8	49.4	53.6
LIMU-BERT	61.7	60.9	49.8	51.9
SimMTM	55.9	57.5	48.3	50.6
FOCAL	48.1	47.1	36.9	38.3
RelCon	38.5	39.1	39.9	42.5
SSL-Wearables	54.7	56.3	45.6	48.9
InrtlWrist, plain attention	63.9	66.7	49.4	53.3
InrtlWrist, accelerometer only	62.2	64.4	49.1	51.5
Random frozen encoder	36.4	40.2	31.9	35.4
Raw input, 0 layers	11.4	51.7	11.4	51.7
Fine-tune				
InrtlWrist (Ours)	54.2	65.5	48.0	59.4
LIMU-BERT	57.3	67.8	49.1	59.4
SimMTM	61.9	69.0	50.5	59.8
FOCAL	39.3	49.4	39.5	47.5
RelCon	42.7	57.5	38.3	51.0
SSL-Wearables	58.3	66.7	53.1	60.7
InrtlWrist, plain attention	70.3	74.7	56.2	63.6
InrtlWrist, accelerometer only	60.6	67.8	49.7	59.2
Scratch	60.8	69.0	45.1	56.5
Task-specific systems (not ranked)				
Mix-Token (reimpl.)	50.0	62.1	49.0	59.2
NormWear-Base (first-party)	45.3	51.7	42.5	50.4

ConvLSTM (Ordóñez & Roggen, 2016) for activity recognition) use their own inputs, output rates, auxiliary components and recipes, so we report them for orientation and do not rank them against the controlled encoders. The pose systems are retrained on the benchmark's Nymeria training sessions and read wrist orientation, which the controlled encoders never see. TLIO and Air-IO are retrained for the wrist-frame odometry protocol and scored by the benchmark's trajectory reconstruction on the

Table 9: IMU–text retrieval on HumanML3D (simulated wrist IMU, test split), where the task-specific row is a wrist-only UniMTS retrained from scratch whose text-to-IMU direction is at chance because of hubness.

		Gallery of 100				Full gallery			
Encoder	R@1 mean↑	i2t R@1↑	t2i R@1↑	i2t MRR↑	t2i MRR↑	i2t R@1↑	t2i R@1↑	i2t MRR↑	t2i MRR↑
Frozen probe									
InrtlWrist (Ours)	27.9	28.3	27.5	42.5	41.8	3.6	3.4	9.0	8.2
LIMU-BERT	24.9	25.3	24.5	39.2	38.4	3.1	2.5	7.8	6.6
SimMTM	17.2	17.4	17.0	30.3	29.8	1.7	1.2	5.0	4.2
FOCAL	8.6	9.1	8.0	19.9	18.6	0.7	0.7	2.3	2.2
RelCon	13.2	13.9	12.4	26.9	25.0	1.0	1.0	3.6	3.1
SSL-Wearables	24.8	25.6	24.0	39.2	37.9	2.5	2.5	7.4	6.6
InrtlWrist, plain attention	26.3	26.7	26.0	40.4	39.7	3.4	2.4	8.3	7.0
InrtlWrist, accelerometer only	20.7	21.0	20.3	34.4	33.7	2.1	1.7	6.1	5.4
Random frozen encoder	9.5	10.0	9.1	21.3	19.7	0.6	0.4	2.6	2.1
Raw input, 0 layers	0.5	1.0	0.0	5.2	2.0	0.0	0.0	0.3	0.1
Fine-tune									
InrtlWrist (Ours)	38.8	39.5	38.1	54.9	53.9	5.1	4.4	12.6	11.0
LIMU-BERT	36.6	37.2	36.0	52.5	51.4	4.4	3.9	11.5	10.3
SimMTM	32.4	32.7	32.1	48.1	47.4	4.0	3.4	10.3	9.0
FOCAL	22.4	22.7	22.0	37.2	36.5	2.1	1.9	6.2	5.7
RelCon	30.0	30.5	29.5	45.5	44.5	3.2	2.9	8.9	8.1
SSL-Wearables	34.5	34.9	34.1	50.4	49.5	4.2	4.5	11.1	10.5
InrtlWrist, plain attention	37.0	37.7	36.3	53.0	51.8	4.6	4.1	11.7	10.6
InrtlWrist, accelerometer only	31.5	31.7	31.3	46.9	46.3	3.6	3.3	9.6	8.8
Scratch	23.2	23.6	22.8	38.5	37.6	2.1	2.1	6.5	6.0
Task-specific systems (not ranked)									
UniMTS (from scratch, wrist-only)	13.2	24.9	1.5	39.3	6.5	2.3	0.0	6.9	0.3

Table 10: Parkinson's classification on PADS (healthy, Parkinson's or differential diagnosis, all values in %), with the dataset authors' BOSS classifier (Varghese et al., 2024) re-run under this protocol as the task-specific row.

	Per subject		Per window		
Encoder	macro-F1↑	bal-acc↑	bal-acc↑	macro-F1↑	acc↑
Frozen probe					
InrtlWrist (Ours)	55.2	60.0	48.3	44.1	47.9
LIMU-BERT	50.5	58.4	49.0	43.8	46.7
SimMTM	48.7	59.3	48.0	40.3	41.6
FOCAL	40.7	53.9	39.2	31.3	31.5
RelCon	40.5	51.5	40.7	34.8	42.1
SSL-Wearables	57.4	61.7	48.5	43.9	47.3
InrtlWrist, plain attention	48.1	59.7	48.8	41.4	42.6
InrtlWrist, accelerometer only	48.2	61.6	47.6	40.1	41.2
Random frozen encoder	51.7	55.9	42.7	38.4	42.3
Raw input, 0 layers	12.6	33.3	33.3	12.6	23.4
Fine-tune					
InrtlWrist (Ours)	51.9	63.9	51.0	43.6	45.3
LIMU-BERT	46.0	55.0	49.3	43.0	48.9
SimMTM	46.8	59.3	49.1	40.6	41.5
FOCAL	46.3	57.7	45.1	37.0	37.8
RelCon	49.1	58.1	46.9	39.9	41.4
SSL-Wearables	48.7	59.0	50.0	41.9	43.2
InrtlWrist, plain attention	61.0	68.3	50.1	44.9	49.7
InrtlWrist, accelerometer only	59.9	67.7	48.7	42.6	44.7
Scratch	29.9	45.5	43.9	31.8	31.9
Task-specific systems (not ranked)					
PADS pipeline, BOSS + CatBoost (Varghese et al.)	48.1	58.5	42.1	35.8	37.2

same 73,515 evaluation windows. DeepConvLSTM is trained from scratch on the same windows and splits as each activity set's scratch row. The official UniMTS checkpoint is left out because it was pretrained on HumanML3D; the row is its wrist-only reimplementation trained from scratch and scored by the same retrieval script. Its text-to-IMU direction is at chance because of hubness (its text embeddings are nearly collinear, mean pairwise cosine 0.996), not a scoring fault. The secondary metrics of every other cell (angular error, velocity MSE, window and clip accuracy, both retrieval directions at Recall@1, 5 and 10) are in the released results.

Table 11: Activity recognition (macro-F1, %) on the HHAR, MHEALTH and PRISM, which are subject-disjoint, and TransfHAR (Bradshaw et al., 2026), which is split in time, with their mean.

	HHAR			MHEALTH			PRISM			TransfHAR			Mean
Encoder	macro-F1↑	bal-acc↑	acc↑	macro-F1↑	bal-acc↑	acc↑	macro-F1↑	bal-acc↑	acc↑	macro-F1↑	bal-acc↑	acc↑	macro-F1↑
Frozen probe													
InrtlWrist (Ours)	62.1	62.7	61.9	68.6	70.0	68.3	27.9	27.2	67.0	72.7	72.2	73.0	57.8
LIMU-BERT	61.2	60.8	59.9	67.6	68.1	66.3	25.1	25.1	66.0	71.1	70.3	71.4	56.2
SimMTM	57.5	57.8	56.8	61.4	63.1	62.5	26.7	26.3	66.9	66.8	67.2	68.7	53.1
FOCAL	44.9	46.7	45.8	63.0	63.0	63.4	21.9	23.2	65.9	47.0	48.1	48.9	44.2
RelCon	33.3	35.5	35.5	38.4	40.4	42.9	18.3	21.2	64.8	51.6	53.1	53.3	35.4
SSL-Wearables	59.2	59.7	59.5	66.3	67.0	66.3	27.6	27.1	66.3	67.4	66.6	67.2	55.1
InrtlWrist, plain attention	60.2	60.4	60.3	68.1	69.6	68.5	26.5	26.1	66.4	73.0	71.4	72.7	56.9
InrtlWrist, accelerometer only	57.8	58.3	58.8	68.5	70.8	69.3	28.0	27.2	66.5	69.5	69.0	70.0	55.9
Random frozen encoder	32.4	36.2	37.7	46.6	48.9	51.8	16.1	20.2	64.1	9.7	12.1	17.7	26.2
Raw input, 0 layers	5.3	16.7	19.0	1.4	8.3	9.2	15.6	20.0	64.2	0.1	1.7	4.6	5.6
Fine-tune													
InrtlWrist (Ours)	60.3	61.0	60.2	72.4	74.9	73.7	36.8	36.2	67.8	86.1	85.4	85.9	63.9
LIMU-BERT	56.3	58.5	57.2	62.9	69.2	68.1	29.5	29.2	64.9	75.8	76.3	78.3	56.1
SimMTM	47.0	48.1	48.4	56.8	62.8	62.3	30.8	30.6	67.3	73.7	74.3	74.6	52.1
FOCAL	61.3	61.7	61.4	65.9	68.2	67.8	28.8	27.9	68.5	68.3	67.5	68.9	56.1
RelCon	42.9	46.3	46.7	60.3	63.4	62.2	28.9	28.4	68.4	79.0	78.3	79.1	52.8
SSL-Wearables	53.5	55.7	56.0	64.8	68.5	68.9	32.8	32.5	67.4	75.9	75.6	78.0	56.7
InrtlWrist, plain attention	52.6	54.9	55.2	68.6	72.7	71.4	36.2	35.3	70.0	78.6	78.1	79.8	59.0
InrtlWrist, accelerometer only	56.8	57.9	58.0	68.3	70.7	68.8	29.8	29.5	67.5	77.2	76.6	77.8	58.0
Scratch	48.8	50.3	49.6	63.2	66.2	65.2	27.8	27.2	68.5	73.5	73.4	74.4	53.3
Task-specific systems (not ranked)													
DeepConvLSTM (Ordóñez & Roggen)	44.8	46.6	44.8	55.6	59.5	60.2	35.6	34.0	70.6	71.6	72.9	72.8	51.9

B.1 TransfHAR smartwatch dataset

TransfHAR (Bradshaw et al., 2026) contains recordings of ten adult participants (anonymous codes p1–p10) wearing a smartwatch that logged device-motion data at about 50 Hz: user acceleration and gravity (in g, summed and converted to m/s^2) and rotation rate (rad/s). Each of 70 recordings contains one activity (2.0 h in total; 5,850 s used for training and 1,284 s for evaluation). Its file names carry 69 raw activity names; merging spelling variants leaves 60 everyday hand activities (writing, typing, tying a shoelace, stirring coffee, shuffling cards and wiping among them; the dataset's label map lists all 60). Most were performed by a single participant (53 of 60 classes; the other 7 by two to four), so a participant-disjoint split is impossible. The benchmark therefore splits each recording in time, the first 80 % for training and the last 20 % for evaluation, with one 2.56 s window removed at the boundary. This measures recognition of an activity later in the same recording, not generalization to new people, and results on this set complement the subject-disjoint sets. The dataset provides the six channels, timestamps, anonymous participant codes and the label map.

C Data Audit

Two kinds of leakage. *Inductive* leakage places evaluation participants in downstream training and leaks labels; *transductive* exposure shows unlabeled evaluation inputs to self-supervised pretraining and leaks only the test participants' input distribution. Both inflate evaluation numbers.

Pretraining pools. The reference corpus contains no OmniIMU-Bench evaluation set. Commonly used activity pools, ours included, bundle WISDM (Weiss et al., 2019), PAMAP2 (Reiss & Stricker, 2012), RecoFit (Morris et al., 2014), SAMoSA (Mollyn et al., 2022), Shoaib (Shoaib et al., 2016), UT-Watch (Bhattacharya et al., 2022) and UTD-MHAD (Chen et al., 2015), so we exclude these datasets from evaluation; UTD-MHAD is listed in our pool but its clips are shorter than one window and contribute nothing to the corpus, which is why Table 1 omits it. An earlier eleven-source corpus that we used during development shows how easily exposure happens: it contained the Nymeria and HumanML3D training splits and OpenWatch recordings under the release's random 80/10/10 partition, whose training part contains all five OmniIMU-Bench gesture test participants. That corpus touches this paper only in Appendix H: in the session-level sweep, reported for reconstruction loss alone; in the encoder whose embeddings define the motion modes; and in the grid's validation set, used only for reconstruction loss and checkpoint selection. The reference encoder is validated on the five pretraining sources alone.

Duplicated participants in OpenWatch. Grouping the release's files by content hash reveals byte-identical recordings under different participant ids (Table 12): 50 ids correspond to 43 distinct identities, and every one of the 35 motion-augmented ids is a wrist-mirrored copy of a real participant, identified by the correlation of acceleration magnitudes. Any split over the release's ids can therefore place the same person in training and test; the original benchmark's split places a copy of training participant 46 in its test set as id 48, and that participant is the best-recognized test participant for every model we trained (clip accuracy 0.88–1.00 against 0.38–0.88 for the other five). With only 8 of 99 test clips, removing it moves clip macro-F1 by just 0.01–0.02, so the score differences in Table 12 should not be read as the size of the leak. OmniIMU-Bench splits over canonical identities and removes exact duplicates.

Table 12: OpenWatch release audit. Duplicates are exact: files are grouped by content hash. Bottom: clip macro-F1 (%) on the release's participant-id split, where test id 48 is a byte-identical copy of training id 46, and on the canonical-identity split used by OmniIMU-Bench. The two splits have different test participants, so the difference mixes leakage with test-set composition (per-participant analysis in the text).

Participant ids in the release	50
Distinct identities after merging duplicate ids	43
Duplicate-identity groups (ids)	39, 41, 42, 46, 47, 48, 49; 40, 43
Byte-identical files / groups / ids involved	254 / 102 / 9
Recordings before / after exact de-duplication	2,330 / 2,178
Augmented ids that are mirrored copies of real ids	35 of 35 (ids 51–85; 32 distinct sources)
Clip macro-F1 (%)	release split / canonical split
InrtlWrist, frozen probe	57.0 / 64.7
InrtlWrist, fine-tune	64.7 / 54.2
Scratch	54.8 / 60.8

Motion-capture provenance. HumanML3D, whose simulated IMU we use for retrieval, is built from 17 AMASS subsets and HumanAct12. The reference corpus contains neither: its Motion-X++ subsets are recovered from video and multi-view capture, and BONES-SEED is studio motion capture (half of the released clips are mirrored; we count only the originals). A direct search for copies, matching every two-second snippet by its bone directions, recovers same-recording HumanML3D clips at a median 4.4° (positive control) but finds no Motion-X++ clip within 5° of any test clip (closest 8.5°). BONES-SEED was checked by provenance only; its license requires the attribution that follows: this work uses Motion Data by Bones Studio (https://bones.studio/); the released encoders' training data includes Motion Data by Bones Studio, and use of the underlying dataset is subject to the BONES Motion Capture Dataset License Agreement. None of the five sources draws on Nymeria. The only overlap we cannot exclude is PRISM with SAMoSA, collected by the same group; identifiers are anonymized per dataset, so shared participants cannot be checked.

Sensor units, and overlap inside HumanML3D. EatingSpeed stores gyroscope data in deg/s while all other sources use rad/s; per-window z-scoring makes our encoders invariant to this scale, but a pipeline keeping physical units would mix scales by a factor of 57, so the harness converts units on load and rejects a source whose still windows have a median gyroscope magnitude above 0.3 rad/s or whose 99th percentile exceeds 40 rad/s. Every HumanML3D split list pairs each clip with a mirrored copy; Table 1 counts unique motion only. The official split also cuts some AMASS recordings into several clips, so 437 of the 2,192 test clips share frames of a source recording with a training or validation clip; this affects every encoder equally but makes the retrieval task partly transductive.

D Controls and Decoder Capacity

We probed the same frozen encoders on full-body pose with decoders of increasing capacity: a linear readout, the benchmark decoder (one cross-channel attention layer of width 192), one cross-channel layer of width 384, the same plus four temporal layers, and a four-plus-eight-layer decoder of width 384. A never-trained encoder with the InrtlWrist architecture falls 9.7 mm behind InrtlWrist with the linear readout, 7.1 mm with the benchmark decoder and 7.6 mm with the wider single layer, but

only 0.3 mm once four temporal layers are added and 0.8 mm with the deep decoder; on odometry a 12-layer decoder lets the untrained encoder beat every pretrained one. Deep decoders learn the task from whatever the frozen encoder passes through, so the kinematic probes use the light decoders of the original task configurations (arm pose 2 cross-channel plus 2 temporal layers, full-body pose 1 cross-channel layer, odometry 1 temporal layer, all of width 192). The raw-input control is sensitive to token pooling on recognition and retrieval; its near-chance scores therefore characterize this readout configuration rather than learning from raw IMU in general.

E Cross-Task Transfer on Disjoint Participants

The three kinematic tasks share one Nymeria split that is session-disjoint but not participant-disjoint (Appendix B), so a source encoder trained from scratch on one of them (Figure 3) has seen the participants of another's evaluation sessions. To separate that exposure from the label supervision we repeated the kinematic block of the figure on three disjoint thirds of the participants, one per task, with every cell (from-scratch source, never-trained control, InrtlWrist probe and transfer probes) retrained inside its third under the benchmark's own decoders, splits and budgets. A third carries a third of the data, so the raw numbers are not comparable with Table 3; the gains over the never-trained encoder of the same third are, and Table 13 lists both. Five of the six frozen transfers survive at about half their size (arm→full-body $+8.6$ to $+5.2$ %, arm→odometry $+17.5$ to $+10.8$, full-body→arm $+3.8$ to $+2.3$, full-body→odometry $+20.1$ to $+11.3$, odometry→full-body $+4.8$ to $+2.5$); only odometry→arm changes sign ($+2.4$ to -2.0). What does not survive is the comparison with the InrtlWrist: on the shared split a kinematic source matches or beats the frozen InrtlWrist on every kinematic target, whereas on disjoint thirds only full-body→arm still does (8.62 against 8.68 cm, inside the run-to-run spread) and the InrtlWrist leads on full-body pose (101.9 against 102.2 mm) and odometry (0.212 against 0.223 m); its own margin over the never-trained encoder is $+1.6$, $+5.5$ and $+15.8$ % on the thirds against $+3.8$, $+6.8$ and $+14.4$ % on the shared split, as expected of a corpus with no Nymeria recording. The sources are also trained on a third of the data, so this control does not separate data volume from participant exposure.

Table 13: The kinematic block of Figure 3 repeated on three disjoint thirds of the Nymeria participants, one per task, so that no encoder has seen a participant that another task is evaluated on. Frozen probes only. *Gain* is the improvement over the never-trained encoder *of the same split*; the raw metrics of the two splits are not comparable, because each third carries a third of the data and every cell, source encoders included, is retrained inside it.

		session split (Fig. 3)		participant split	
Target	Frozen encoder	metric	gain	metric	gain
Arm pose (cm ↓)	never-trained encoder	8.46	—	8.82	—
	InrtlWrist	8.14	+3.8	8.68	+1.6
	source: full-body pose	8.13	+3.8	8.62	+2.3
	source: odometry	8.26	+2.4	9.00	−2.0
Full-body pose (mm ↓)	never-trained encoder	105.0	—	107.9	—
	InrtlWrist	97.9	+6.8	101.9	+5.5
	source: arm pose	96.0	+8.6	102.2	+5.2
	source: odometry	99.9	+4.8	105.1	+2.5
Odometry (ATE m ↓)	never-trained encoder	0.233	—	0.252	—
	InrtlWrist	0.200	+14.4	0.212	+15.8
	source: arm pose	0.193	+17.5	0.224	+10.8
	source: full-body pose	0.186	+20.1	0.223	+11.3

F Joint Task Supervision

We vary how the pretrained representation is adapted, asking whether supervising two tasks at once changes the cross-task trade-offs observed in Section 6. Starting from the same MAE, we alternate updates on full-body pose and either odometry or microgesture, using separate decoders. Table 14 shows gains for both partner tasks; gesture F1 also exceeds training from scratch. Full-body MPJPE, however, rises from 88.1 mm alone to 91.5 mm with odometry and 94.4 mm with microgesture. Frozen IMU–text Recall@1 falls from 27.9 to 21.0 and 17.2, respectively. These partner-task gains accompany lower pose and language scores. The task pairs use different training schedules, and each target selects its own checkpoint; the reported scores need not occur at one shared checkpoint. Two pairs were run, each starting from the same reference InrtlWrist as every single-task fine-tune, with each task keeping the benchmark's own loader, decoder and loss and one task per optimizer step in a fixed cycle: full-body pose with odometry at a 1:1 ratio for 40,000 steps, and full-body pose with microgesture at 10:1 for 22,000 steps because the gesture training set is 13 batches.

Table 14: **Joint fine-tuning with full-body pose.** Both MAE conditions start from the same pretrained encoder. Scores use per-task checkpoints. Bold: best per row.

Target	Scratch	MAE alone	MAE + pose
Odometry (ATE, m↓)	0.162	0.154	**0.149**
Microgesture (clip F1↑)	60.8	54.2	**65.0**

G Embedding Space of the Reference Encoder

Figure 6 projects with t-SNE (van der Maaten & Hinton, 2008) the reference encoder's frozen pooled embeddings of 12,961 windows from eighteen sources (twelve labeled activity datasets, the three mocap-derived collections, Nymeria, OCDetect and EatingSpeed). Every number below is the mean fraction of a window's ten nearest neighbors (cosine) that share a property with it, so it measures local geometry, not the projection's layout; the labels are the datasets' own, 261 activities in all.

Activity structure is local and carried by whole-body motion. 38.8 % of a window's ten nearest neighbors share its activity label (chance 1.7 %) and 65.6 % its semantic family (chance 14.5 %), yet neither labeling clusters globally (negative cosine silhouettes), so the projection is not a taxonomy. Restricting the neighborhood to windows from *other* datasets separates what travels between corpora:

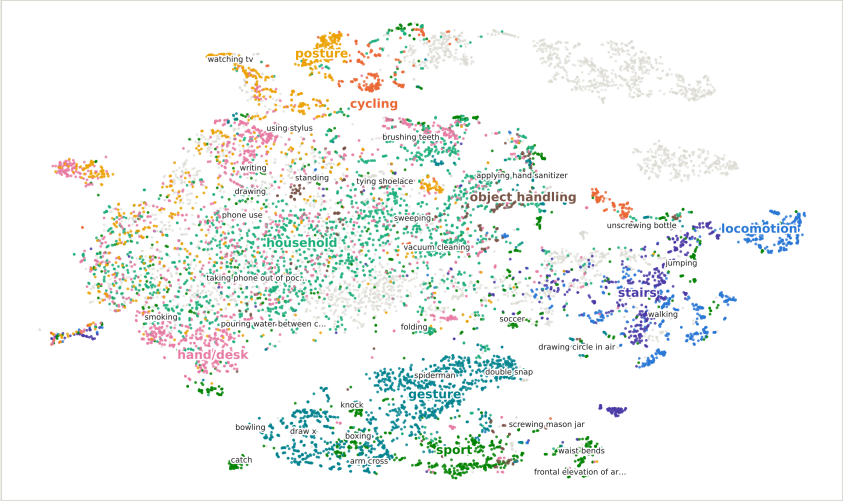

Figure 6: t-SNE (van der Maaten & Hinton, 2008) of the reference InrtlWrist encoder's frozen embeddings (cosine distances): 12,961 windows from eighteen sources, at most 800 each. The 9,139 windows with one of 261 activity labels are colored by semantic family and labeled by family and, where names fit, by activity; gray is the rest, including OCDetect and EatingSpeed, two pretraining sources without activity labels. The projection is illustrative; every statement in the text rests on neighborhood statistics.

locomotion 0.512 (chance 0.059), cycling 0.466 (0.026) and stairs 0.415 (0.042), 8.7 to 18 times chance, against household, hand and desk work, posture, gestures and gym exercise at 1.7 to 4.5 times chance. By activity, cycling, running, jogging, walking and stairs travel, while sitting, cooking, typing, washing dishes and folding stay inside the dataset that recorded them, partly because per-window standardization removes the gravity offset that distinguishes static postures.

Synthesis separates the space; pretraining exposure barely does. The three sources synthesized from motion capture are isolated regardless of exposure (BONES-SEED 0.981 and Motion-X++ 0.899, in the pool; HumanML3D 0.942, held out): mocap-derived windows have 0.96 of their ten nearest neighbors inside their own dataset against 0.59 for recorded windows, a contrast of $+0.364$. The exposure contrast over the recorded windows is $+0.027$ and changes sign under leave-one-source-out (from -0.060 without MHEALTH to $+0.090$ without Nymeria); the two least isolated sources are EatingSpeed at 0.335, in the corpus, and Nymeria at 0.338, never seen. Recorded corpora therefore occupy the same region whether or not they were pretrained on, while this objective keeps synthesized inertial data apart even when it trains on a large amount of it, a caution for enlarging a corpus with motion-capture synthesis. UTD-MHAD (Chen et al., 2015) (0.980) and OpenWatch (0.865) are isolated by an artifact: their short clips are wrap-padded 2.91 and 3.41 times to reach 10.24 s, a periodicity no other source has, so they are excluded from every exposure contrast.

What this supports. One self-supervised encoder places heterogeneous recorded corpora, seen or unseen, in a shared region with activity structure inside it at about 23 times chance, strongest for gross locomotion and weakest for fine manipulation; it does not support a claim of a dataset-invariant representation, since dataset identity remains the strongest neighborhood signal (0.632 against chance 0.058), and only the task measurements of Sections 5 and 6 establish what the representation is worth.

H Pretraining Volume and Motion Coverage

Coverage saturates within the first percent. A uniform draw of 0.1% of the pool (1,959 windows) already contains 82% of the motion modes, a 0.5% draw 99.7%, and every draw of 5% or more all of them; the reference corpus, at 10% of the pool, covers every mode.

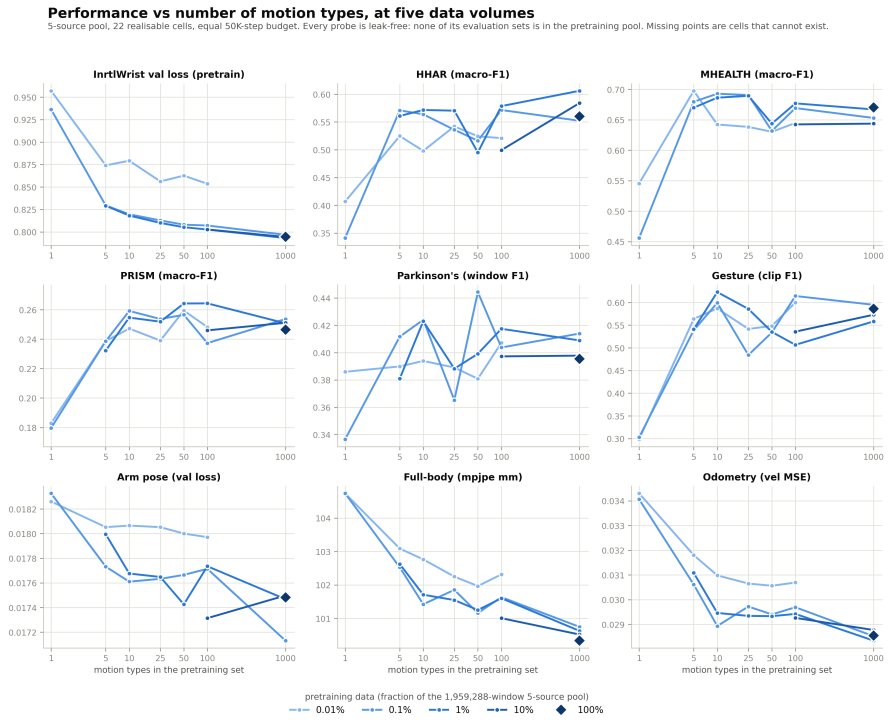

Figure 7: Each probe of the grid against the number of motion modes at five pretraining volumes on the five-source pool, every model trained for 50,000 steps; Figure 5 is its average. Modes are added largest first ($K{=}1$ is 98 % EatingSpeed). Activity, Parkinson's (window macro-F1) and gesture panels use the benchmark probes on evaluation sets outside the pool; the arm, full-body and odometry panels use the grid's own lighter probes and recipe, so they show trends within the grid only.

Downstream performance is flat across a thousand-fold range of data. With all modes available, pretraining on 1,959 windows gives a reconstruction loss within 0.3% of the full pool and an arm-pose probe indistinguishable from the other volumes; full-body error varies by 0.4 mm, odometry velocity error by 1.5% and MHEALTH, PRISM and gesture F1 by at most 4 points, without a trend. No classification probe improves from 1% to 100% of the pool (per task in Figure 7).

Restricting motion modes matters most at the extreme. At 1,959 windows, training on a single mode raises InrtlWrist loss from 0.797 to 0.936 and takes the eight-probe average to -96% of the full pool's gain, almost back to a never-trained encoder, collapsing every classification probe (gesture F1 30.3 against 59.5 with all modes). From ten modes onward the kinematic probes are within 5% of the 1,000-mode model in raw error, which for full-body pose and odometry is still 10 to 41% of the full pool's gain over a never-trained encoder; from five modes onward the activity and gesture probes fluctuate without a trend (Figure 5). Because modes are added largest first, the mode-count axis also shifts source composition (the single-mode subset is 98 % EatingSpeed, the 100-mode subsets about 90 % OCDetect), so this comparison bounds the effect of the number of modes rather than isolating it.

Scope and caveats. These results hold for one architecture and one data pool at 50,000 steps, and they do not separate diversity from volume: the uniform draws change size and coverage together, and the mode-restricted draws change source composition with the mode count. Enlarging the corpus brings no consistent downstream gain at this budget, coverage of the motion modes is complete long before the corpus is, and whether more data pays at larger budgets is open. The motion modes are k-means clusters of embeddings from an InrtlWrist trained on the earlier eleven-source corpus, re-clustered on the five-source pool: a data-driven partition, not a taxonomy of movements. An earlier session-level sweep on the eleven-source corpus (5 %, 25 % and 100 % of 3.56 million windows,

300,000 steps each) lowers reconstruction loss by only 0.4 % (0.7604 to 0.7572); because that corpus contains evaluation participants (Appendix C), we do not use its downstream results.