MOTION CAPTURE IS NOT THE TARGET DOMAIN: SCALING SYNTHETIC DATA FOR LEARNING MOTION REPRESENTATIONS (DATA-FM @ ICLR 2026)

Firas Darwish *
Department of Statistics
University of Oxford
Oxford, UK
{firas.darwish}@worc.ox.ac.uk

George Nicholson
Department of Statistics
Big Data Institute
University of Oxford
Oxford, UK
{george.nicholson}@stats.ox.ac.uk

Aiden Doherty†
Nuffield Department of Population Health
Big Data Institute
University of Oxford
Oxford, UK
{aiden.doherty}@ndph.ox.ac.uk

Hang Yuan †
Nuffield Department of Population Health
Big Data Institute
University of Oxford
Oxford, UK
{hang.yuan}@ndph.ox.ac.uk

ABSTRACT

Synthetic data offers a compelling path to scalable pretraining when real-world data is scarce, but models pretrained on synthetic data often fail to transfer reliably to deployment settings. We study this problem in full-body human motion, where large-scale data collection is infeasible but essential for wearable-based Human Activity Recognition (HAR), and where synthetic motion can be generated from motion-capture–derived representations. We pretrain motion time-series models using such synthetic data and evaluate their transfer across diverse downstream HAR tasks. Our results show that synthetic pretraining improves generalisation when mixed with real data or scaled sufficiently. We also demonstrate that large-scale motion-capture pretraining yields only marginal gains due to domain mismatch with wearable signals, clarifying key sim-to-real challenges and the limits and opportunities of synthetic motion data for transferable HAR representations.

1 INTRODUCTION

Synthetic data holds significant promise for addressing data scarcity in machine learning, particularly in regimes where collecting large-scale, diverse real-world datasets is expensive, slow, or impractical (Oneto et al., 2025; Broda et al., 2025; Lu et al., 2023). By complementing (or in some cases substituting) real data, synthetic data enables scalable training across tasks and environments while offering greater control over data diversity and coverage.

Recent progress across domains highlights the utility of synthetic data, with successful applications in robotics (Zhao et al., 2020; Cheng et al., 2025), autonomous driving (Song et al., 2024; Hu et al., 2023), and Large Language Model (LLM) pretraining (Kang et al., 2025; Alcoba Inciarte et al., 2024). Despite these successes, models trained primarily on synthetic data often struggle to transfer reliably to real-world settings due to the sim-to-real gap, where mismatches in perceptual or dynamical details lead to degraded performance at deployment (Hu et al., 2023; Zhao et al., 2020).

*Corresponding Author
†Joint supervision.

Human motion (particularly full-body human motion) is another domain where scaling data collection is essential. Human Activity Recognition (HAR) models use motion data from wearable devices to classify activities over time, but models trained and evaluated on the same datasets often generalise poorly due to systematic variations in sensors, subjects, and environments (Chen et al., 2021). Consequently, pretraining on large and diverse motion data is critical for learning representations that transfer effectively across HAR tasks.

However, collecting human motion data at scale remains infeasible in practice (Lin et al., 2024). Motion capture datasets, while accurate, are limited in both diversity and volume and are largely restricted to controlled laboratory environments (Wang et al., 2014), while large-scale wearable data collection in the wild is limited to wrist motion (Ma et al., 2024).

At the same time, human motion is a natural candidate for synthetic generation via simulation or generative modelling because it is governed by well-defined physical and kinematic constraints that must be maintained over extended temporal horizons, such as balance, contact dynamics, and joint limits (Yuan et al., 2023).

In this paper, we investigate the utility of using synthetic data in pretraining motion time-series models that are used downstream in different HAR tasks. Our core contributions in this paper can be summarised as:

1. We offer a framework to apply the sim-to-real paradigm in the context of wearables, and formulate the core challenges that arise when representing human motion.
2. We demonstrate that pretraining motion time-series models on synthetic data improves generalisation to downstream HAR tasks; however, we do not observe the same scaling laws compared to training on real data (Narayanswamy et al., 2024).
3. We show that scaling motion-capture-based pretraining leads to marginal gains in HAR tasks, exhibiting evidence of a domain mismatch with the downstream wearable sensor signals that limits the transferability of motion-capture representations.

2 Related Work

2.1 Pretraining Motion Time-Series Models

Pretraining motion time-series models has gained momentum as a means of learning transferable representations for HAR, typically through self-supervised objectives such as masked modelling (Liu et al., 2025; Senane et al., 2024), contrastive learning (Zhang et al., 2024; Xu et al., 2025), multimodal alignment (Girdhar et al., 2023), or transformation discrimination (Yuan et al., 2024). Despite this progress, generalisation across HAR datasets remains challenging due to systematic variations in subjects, activities, and sensing setups: physiological and biomechanical differences between wearers alter the mapping from latent activity to sensor signals (Jimale & Noor, 2021; Calatrava-Nicolás et al., 2025), activity sets vary across datasets with only partial overlap, and device heterogeneity introduces additional distribution shifts (Kunze & Lukowicz, 2014; Stisen et al., 2015).

In this work, we adopt the UniMTS pretraining and evaluation framework (Zhang et al., 2024) because its contrastive formulation achieves strong performance while explicitly addressing these sources of variation through full-body Inertial Measurement Unit (IMU) pretraining, random joint masking, and rotation-based augmentation for device invariance; moreover, its text–motion alignment enables compatibility with text-conditioned motion generation, and its evaluation across 18 diverse HAR datasets provides a rigorous test of generalisation.

2.2 Human Motion Generation

Human motion generation has been extensively studied across a wide range of generative paradigms, conditioning modalities, and evaluation protocols (Zhu et al., 2023; Khani et al., 2025). Prior work explores diverse modelling approaches, including GANs (Wang et al., 2021; Yamamoto & Murakami, 2024), VAEs (Bie et al., 2022; Diomataris et al., 2024; Guo et al., 2022), diffusion models (Zhang et al., 2022; Tevet et al., 2022), and autoregressive architectures (Wang et al., 2025; Pinyoanuntapong et al., 2024), with generation conditioned on text (Tevet et al., 2022; Zhang et al., 2022;

Guo et al., 2022), discrete action labels (Guo et al., 2020), audio signals such as music or speech (Qi et al., 2023; Ferstl & McDonnell, 2018), or scene representations to ensure physical and environmental consistency (von Marcard et al., 2018; Zhang et al., 2021).

Evaluation typically combines qualitative human evaluations (Andreou et al., 2024) with quantitative fidelity metrics that measure similarity to ground-truth motion (Raab et al., 2023), assessments of physical plausibility (like ground contact, penetration, and foot sliding) (Li et al., 2023), measures of motion diversity across and within activities (Barquero et al., 2022), and alignment with conditioning inputs (Li et al., 2024). As a result, contemporary motion generators produce data of sufficient fidelity and coverage to serve as a viable source of human motion data.

2.3 Synthetic Data in Wearables

Prior work has explored synthetic data generation for wearable-based HAR using a range of generative approaches. GAN-based methods have been shown to effectively substitute or augment real data for gesture and activity recognition, improving performance in low-data and class-imbalanced settings (Kontogiannis et al., 2024; Li et al., 2020; Alharbi et al., 2020; de Souza et al., 2023). More recent work demonstrates that synthetic multivariate time-series generated from wearable sensors can improve HAR accuracy in data-scarce regimes and across larger activity sets (Silva et al., 2025; Lupión et al., 2024), while augmentation with in-the-wild smartwatch data further improves performance (DeSmet et al., 2024).

However, these works focus exclusively on supervised learning settings, and do not examine the role of synthetic data in pretraining motion time-series models or its impact on representation learning and transfer across HAR tasks.

3 Problem Formulation

3.1 Motion Representation

We model human actions as continuous motion of a deformable body in physical space. Let $\mathcal{B}$ denote the human body, represented as a compact three-dimensional material manifold whose points correspond to material particles of the body. At any time $t \in [0, T]$, a body configuration is given by a smooth embedding

$$q(t) : \mathcal{B} \to \mathbb{R}^3,$$

which assigns a spatial position to every body point.

An action is then defined as a time-indexed family of such configurations (a curve in configuration space). Formally, an action is a function

$$a \in C^k([0, T],\ \mathrm{Emb}(\mathcal{B}, \mathbb{R}^3)),$$

where C^k denotes k-times continuously differentiable functions in time, and $\mathrm{Emb}(\mathcal{B}, \mathbb{R}^3)$ is the space of smooth embeddings of the body into three-dimensional physical space. The regularity condition $k \geq 2$ ensures that velocity and acceleration fields exist and vary continuously for every body point.

The space of all human actions is therefore the infinite-dimensional function space

$$\mathcal{A} = C^k([0, T],\ \mathrm{Emb}(\mathcal{B}, \mathbb{R}^3)).$$

While the action space $\mathcal{A}$ provides a holistic representation of human motion, real-world sensing and learning systems such as motion capture operate on finite, discretised observations that partially captures this action space.

Let $N \in \mathbb{N}$ denote the number of discrete time steps and let a fixed set of 24 anatomical landmarks (joints) be selected on the body, following the SMPL representation (Loper et al., 2023; Zhang et al., 2024; Guo et al., 2022). At each time step, the spatial configuration of the body is represented by the three-dimensional positions of these joints. A discretised action is then defined as a mapping

$$q : \{0, \ldots, N\} \to \mathbb{R}^{3\times 24},$$

where $q(n)$ specifies the stacked x, y, z coordinates of the 24 tracked joints at discrete time index n.

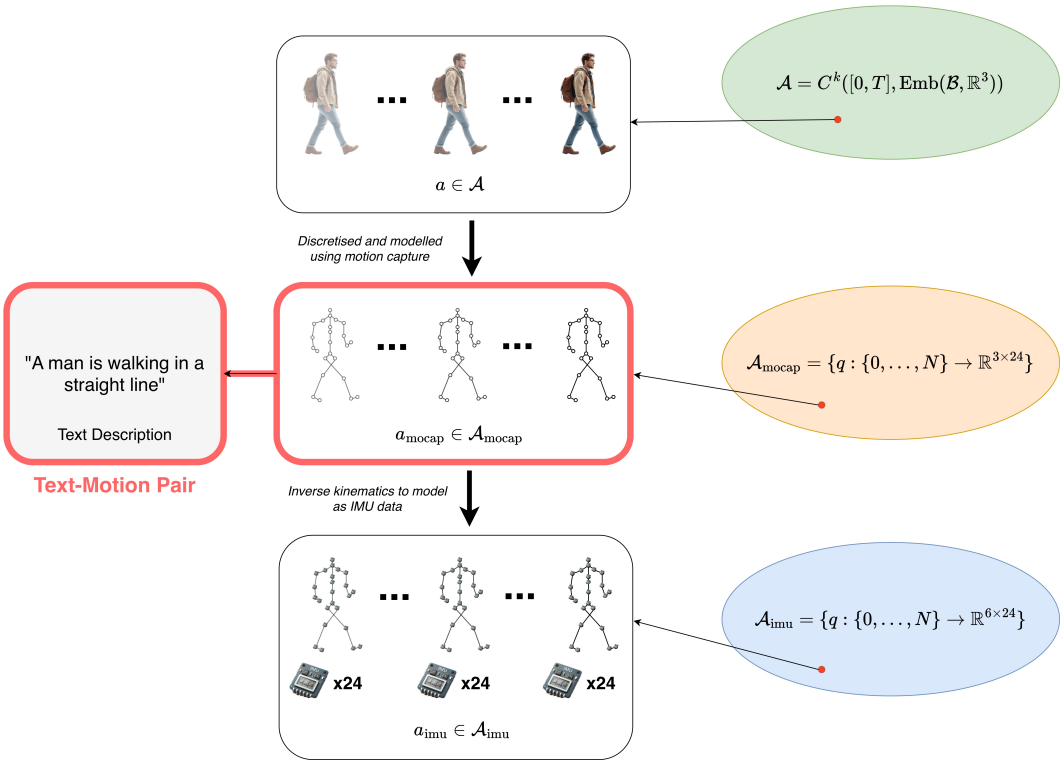

Figure 1: Modelling Human Motion. Human motion modelling relies on selecting a number of joints to track over a period of time, forming a discretised representation of human motion. This motion can be labelled and annotated with text. Using inverse kinematics, one can simulate accelerometer and gyroscope readings for each of those joints over time that can be used for wearables pretraining.

While $\mathcal{A}_{\text{mocap}}$ provides a discretised representation of human motion in terms of joint positions, many downstream HAR systems operate on data collected from wearable inertial sensors (Serpush et al., 2022). To bridge this gap, we introduce an IMU-based action space that aligns with the measurement modality of such sensors.

Specifically, we define the IMU action space as

$$\mathcal{A}_{\text{imu}} = \{q : \{0, \ldots, N\} \to \mathbb{R}^{6\times 24}\},$$

where each of the 24 joints is associated with a six-dimensional inertial measurement vector consisting of tri-axial linear acceleration and tri-axial angular velocity. For each discrete time index n, $q(n)$ represents the stacked inertial measurements across all joints.

The mapping from $\mathcal{A}_{\text{mocap}}$ to $\mathcal{A}_{\text{imu}}$ is obtained via inverse kinematics and differentiation, converting joint position trajectories into joint-centric accelerometer and gyroscope signals (Zhang et al., 2024). This transformation preserves the underlying motion while expressing it in a sensor-centric coordinate system consistent with IMUs.

3.2 Text-to-Motion Generation

We consider a dataset of paired natural language descriptions and discretised motion capture sequences, which can be used to train a text-conditioned generative model that learns a distribution over motions in the motion capture action space $\mathcal{A}_{\text{mocap}}$ (Guo et al., 2022). Because this mapping is inherently non-deterministic, a single text description can correspond to multiple valid motion realisations, and sampling from the learned distribution produces diverse motion trajectories consistent with the same prompt (Petrovich et al., 2022; Zhang et al., 2024).

Given a trained text-conditioned motion model and a predefined set of text prompts, we generate a synthetic motion capture dataset by sampling K motion trajectories per prompt. The resulting motion sequences are then converted into simulated IMU signals using the same inverse kinematics procedure described earlier, expressing each sample in the IMU action space $\mathcal{A}_{\text{imu}}$, which can be used to pretrain wearable motion time-series models (Zhang et al., 2024).

3.3 Pretraining Motion Time-Series Models

3.3.1 Pretraining

To learn motion representations that generalise across downstream HAR tasks, we pretrain a motion encoder using contrastive learning on paired text–motion data, following Zhang et al. (2024). A dual-encoder architecture maps text descriptions and their corresponding IMU motion sequences (drawn from either real datasets or synthetically generated pairs) into a shared embedding space, and training optimises a contrastive objective that increases similarity between matched text–motion pairs while reducing similarity for mismatched pairs within a batch.

3.3.2 Evaluation

Given multiple HAR datasets containing labelled motion sequences in the IMU action space $\mathcal{A}_{\text{imu}}$ and a pretrained motion encoder, we evaluate downstream performance under both 0-shot and k-shot settings following UniMTS (Zhang et al., 2024). Because HAR datasets vary in sensor placement and number, we align each dataset with the pretrained 24-joint motion encoder by retaining channels corresponding to observed joints and zeroing out unobserved ones.

In the 0-shot setting, the motion encoder is frozen and classification is performed by matching motion embeddings to text embeddings of activity labels via similarity scores. In the k-shot setting, we fine-tune the pretrained motion encoder with a linear classification head using k labelled examples per activity class from the target dataset, and then evaluate performance on the dataset's test split.

4 Experiments

4.1 Implementation Details

We generate our synthetic motion data using state of the art text-to-motion models that were all trained on labelled motion capture sequences from HumanML3D (Guo et al., 2022): (1) Human Motion Diffusion Model (MDM) (Tevet et al., 2022), (2) Text-to-motion (T2M) (Guo et al., 2022), and (3) MotionDiffuse (Zhang et al., 2022).

Our real-world data consists of 24,661 motion sequences extracted from HumanML3D motion capture dataset and transformed into A_{imu} (Guo et al., 2022). Each of these motion sequences is paired with a text file containing around 6 sentences, with each sentence providing an individual annotated description of the motion (Zhang et al., 2024). Using each text-to-motion model, we randomly sample one description from the associated text file and generate a motion conditioned on that sentence. This procedure is applied to all 24,661 text files, yielding 24,661 generated motion sequences per pass. We repeat this process multiple times to produce synthetic datasets at increasing multiples of the real-world data volume. The resulting synthetic dataset sizes can be found in Table 3 of Appendix A.

Following from Zhang et al. (2024), our motion encoder is implemented as a 4.94M parameter spatio-temporal graph convolutional network that processes joint-level accelerometry signals over a fixed temporal window. We pretrain the model for 100 epochs using a batch size of 64, learning rate of 10^{-4}, and temperature parameter of 0.1. We also apply data augmentation in the form of random rotations of the accelerometry signals and random masking of a subset of joints during each training iteration. For the text modality, we use the pretrained CLIP ViT-B/32 text encoder (Radford et al., 2021), which is fine-tuned jointly during pretraining.

To evaluate the pretrained models, we compute the macro-averaged F1 score within each dataset, and then report the mean of these scores across all 18 HAR datasets from Zhang et al. (2024) for the 0-shot, 1-shot, 2-shot, 3-shot, 5-shot, and 10-shot settings.

Table 1: Synthetic data effectively complements real motion data during pretraining. Across downstream HAR tasks, pretraining on mixed real and synthetic data tends to improve generalisation relative to real-world data alone, and improvements are most evident in higher-shot regimes. Reported scores are mean macro-averaged F1 across datasets; **bold** and underlined values denote best and second-best results per column. All results are obtained using a fixed random seed across runs.

Method	0-shot	1-shot	2-shot	3-shot	5-shot	10-shot
Real-World Baseline	**0.3070**	0.5445	**0.6243**	0.6611	0.7041	0.7667
Purely Synthetic Data						
MDM	0.2460	0.5120	0.6160	0.6559	0.6974	0.7653
T2M	0.2606	0.5229	0.6100	0.6551	0.7024	0.7649
MotionDiffuse	0.2738	0.5363	0.6060	0.6548	**0.7053**	0.7739
Mixed Data						
MDM	0.3002	0.5337	0.6014	0.6738	0.7049	0.7702
T2M	0.2901	**0.5539**	0.6139	**0.6931**	0.6970	0.7687
MotionDiffuse	0.2732	0.5342	0.6099	0.6707	0.7004	**0.7931**

All synthetic data generation, pretraining, and evaluation experiments are conducted using approximately 12 NVIDIA A100 GPUs, for a total of 60 GPU-days.

4.2 Results

4.2.1 Pretraining with Mixed Data

We study the effect of pretraining data composition by comparing real-world data, synthetic data, and their mixtures while holding the total pretraining volume fixed, allowing us to evaluate both the standalone utility of synthetic motions and their ability to complement real-world data.

Across all settings, we pretrain on a fixed total of 24,660 text–motion pairs, with the following configurations:

- **Real-World Baseline:** 24,660 real-world text–motion pairs, matching the setup in Zhang et al. (2024).
- **Purely Synthetic Data:** 24,660 synthetic text–motion pairs generated by each text-to-motion model (Table 3).
- **Mixed Data:** A randomly sampled 50/50 mixture of real-world and synthetic text–motion pairs.

Results are reported in Table 1. The overall gains from mixed pretraining relative to the real-world baseline align with what Kang et al. (2025) demonstrated in LLM pretraining. Beyond that, models pretrained on purely synthetic data achieve performance close to the real-world baseline. Both synthetic-only and mixed pretraining perform particularly well in higher-shot regimes (3-, 5-, and 10-shot). Across synthetic pretraining settings, we further observe that no single text-to-motion generative model consistently outperforms the others across all evaluation regimes. Performance differences between synthetic datasets are relatively small, indicating that different generators produce motion data of comparable utility for motion time-series pretraining.

4.2.2 Pretraining with Synthetic Data

We examine the effect of scaling the amount of synthetic data used during pretraining. For each text-to-motion model, we pretrain using synthetic datasets at $1\times$, $2\times$, $4\times$, and $8\times$ the size of the real-world dataset, and report downstream evaluation results in Table 2.

While synthetic pretraining benefits from scale, a consistent gap remains in the 0-shot setting: even at $8\times$ scale, synthetic-pretrained models do not surpass the real-world baseline without fine-tuning. We analyse how 0-shot evaluation behaves more extensively in Section 4.2.3. However, this result

Table 2: Scaling synthetic data improves downstream HAR performance. At sufficient scale (8×), synthetic pretraining outperforms real-world pretraining in all fine-tuned settings, with the exception of the 0-shot regime. Reported scores are mean macro-averaged F1 across datasets; **bold** and underlined values denote best and second-best results per column, respectively. All results are obtained using a fixed random seed across runs.

Method	0-shot	1-shot	2-shot	3-shot	5-shot	10-shot
Real-World Baseline	**0.3070**	0.5445	0.6243	0.6611	0.7041	0.7667
MDM (Synthetic)						
1×	0.2460	0.5120	0.6160	0.6559	0.6974	0.7653
2×	0.2928	0.5334	0.6089	0.6505	0.7008	0.7368
4×	0.2546	0.5299	0.6344	0.6749	0.7023	0.7663
8×	0.2654	**0.5517**	0.6266	0.6704	0.7165	0.7792
T2M (Synthetic)						
1×	0.2606	0.5229	0.6100	0.6551	0.7024	0.7649
2×	0.2462	0.5132	0.6062	0.6669	**0.7226**	0.7635
4×	0.2482	0.5256	0.6114	0.6766	0.7024	0.7719
8×	0.2123	0.5491	0.6372	0.6631	0.7050	0.7891
MotionDiffuse (Synthetic)						
1×	0.2738	0.5363	0.6060	0.6548	0.7053	0.7739
2×	0.2739	0.5233	0.6186	0.6725	0.6995	0.7907
4×	0.2566	0.5423	0.6472	0.6700	0.7150	**0.7959**
8×	0.2427	0.5460	**0.6550**	**0.6893**	0.7213	0.7956

seems to suggest that real-world data induces representations that generalise more effectively in the absence of task-specific supervision, whereas synthetic pretraining derives greater benefit from fine-tuning when labelled data is available.

While performance generally improves as synthetic pretraining data is scaled, the trend is not strictly monotonic. This contrasts with prior work showing monotonic gains when scaling real-world pretraining data Yuan et al. (2024). The clear cause for this variability in the synthetic setting is unclear to us, but it may stem from stochasticity in synthetic data generation, such as differences in the semantic expressiveness of sampled text prompts or randomness in the sampled motion lengths.

4.2.3 Scaling Real-World Data

We study the effect of scaling real-world pretraining data by randomly subsampling 10%, 20%, 40%, and 80% of the available text–motion pairs and pretraining models on each subset, with three random seeds per configuration. This analysis serves two purposes: first, to examine the stability of k-shot evaluations by quantifying variability across runs, motivated by the persistent 0-shot performance gap between real and synthetic pretraining; and second, to assess whether the commonly held expectation that increased pretraining data volume leads to improved downstream performance (Narayanswamy et al., 2024) holds when scaling real-world text–motion data in our setting, given that this trend did not hold for synthetic data in Section 4.2.2. Results are shown in Figure 2.

Given the high performance variability across trained models and inconsistent scaling results in the 0-shot setting, out-of-the-box generalisation may not be the best metric to fully capture the quality of learned representations for HAR tasks. It seems comparisons between pretrained motion time-series models are more reliable when some degree of fine-tuning is undertaken.

More critically, even in the fine-tuning setting, scaling real-world data yields only marginal gains, with pretraining on 40% of the data nearly matching performance obtained using the full dataset. This stands in contrast to Yuan et al. (2024); Narayanswamy et al. (2024), which reports substantially larger improvements when scaling real-world pretraining data, but does so using wearable sensor data rather than motion capture. Taken together, these results suggest a domain mismatch: pretraining on motion capture may emphasise factors that are weakly aligned with the signal char-

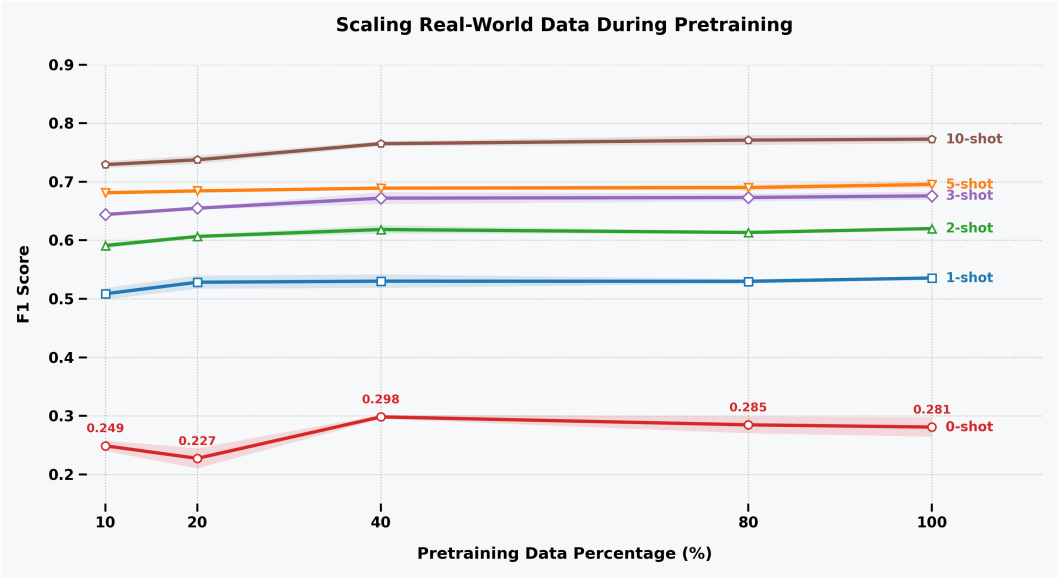

Figure 2: Fine-tuning stabilises gains from real-world pretraining. While increasing real-world pretraining data yields small but consistent improvements in fine-tuned (k-shot) performance, zero-shot performance remains unstable and does not scale reliably with data volume. Curves show results from three pretrained models per data scale, using the same set of random seeds at each percentage; shaded regions indicate standard error. Scores are mean macro-averaged F1 across downstream datasets.

acteristics required for wearable-based HAR, indicating that motion capture may not be the most appropriate representation for learning transferable motion representations (Figure 1).

5 Conclusion

In this paper, we investigate the impact of incorporating synthetic motion data into the pretraining of motion time-series models for HAR. Our results indicate that generalisation to downstream tasks can be improved either by mixing real and synthetic data during pretraining or by using purely synthetic data at sufficiently large scales.

Our reliance on motion capture as a proxy for human motion during pretraining is a key bottleneck in motion pretraining (Figure 1). Motion capture may not constitute an optimal surrogate or representation for human motion when the downstream objective is HAR, which may in turn constrain the benefits of synthetic data designed to approximate the motion capture domain. As a result, the domain over which our representations are learned during pretraining may be misaligned with the domain used downstream.

Looking ahead, synthetic data generation offers the potential for even larger gains if it is not restricted to the real-world text descriptions used to train existing text-to-motion models. The strength of text-to-motion models lies in their ability to interpolate over a learned motion manifold (Figure 3), enabling the synthesis of diverse and novel motion patterns that are absent from motion capture datasets but potentially beneficial for HAR. Exploring richer and more targeted prompt distributions represents a promising direction for future work.

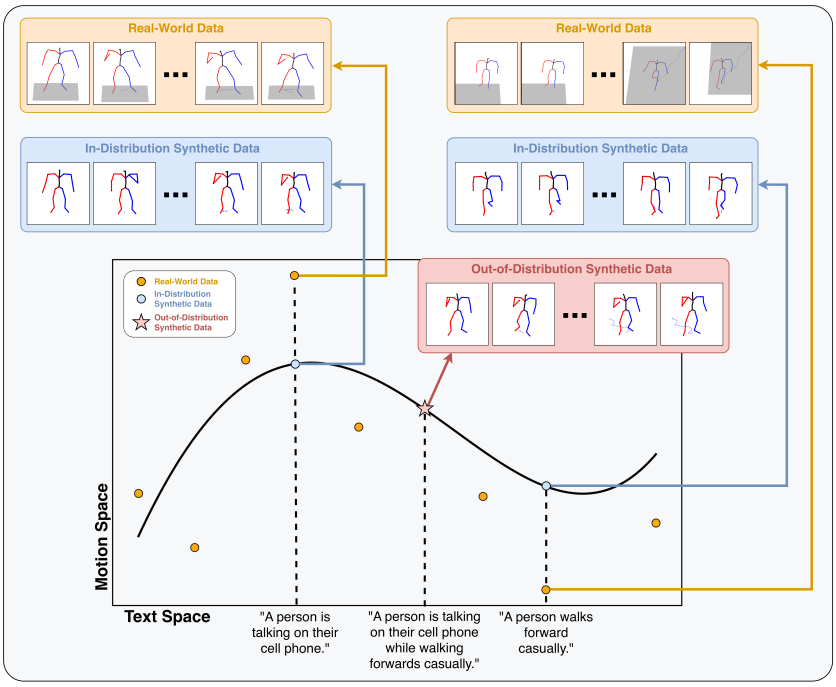

Figure 3: Using the Text-to-Motion Model. Assuming text and motion samples are embedded and projected into a shared low-dimensional space for visualisation, where the black curve represents the text-to-motion model (shown as deterministic mapping here). In-distribution synthetic data is obtained by sampling the text-to-motion model using matching text inputs from our real-world data. Generating out-of-distribution synthetic data consists of sampling motion from text unseen in our real-world data.

ACKNOWLEDGMENTS

We would like to acknowledge the Oxford Biomedical Research Computing (BMRC) facility, which we used to run the various experiments needed for this paper. This research was funded by the Rhodes Trust, and the Engineering and Physical Sciences Research Council (EPSRC). AD is supported by a range of grants from the Wellcome Trust [223100/Z/21/Z, 227093/Z/23/Z], Swiss Re, GSK, Boehringer Ingelheim, Google, National Institutes of Health's Oxford Cambridge Scholars Program, EPSRC Centre for Doctoral Training in Health Data Science (EP/S02428X/1), British Heart Foundation Centre of Research Excellence (grant number RE/18/3/34214), Cancer Research UK, and funding administered by the Danish National Research Foundation in support of the Pioneer Centre for SMARTbiomed. For the purpose of open access, the author(s) has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

REFERENCES

Alcides Alcoba Inciarte, Sang Yun Kwon, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. On the utility of pretraining language models on synthetic data. In Nizar Habash, Houda Bouamor, Ramy Eskander, Nadi Tomeh, Ibrahim Abu Farha, Ahmed Abdelali, Samia Touileb, Injy Hamed, Yaser Onaizan, Bashar Alhafni, Wissam Antoun, Salam Khalifa, Hatem Haddad, Imed Zitouni, Badr AlKhamissi, Rawan Almatham, and Khalil Mrini (eds.), *Proceedings of the Second Arabic Natural Language Processing Conference*, pp. 265–282, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.arabicnlp-1.23. URL https://aclanthology.org/2024.arabicnlp-1.23/.

Fayez Alharbi, Lahcen Ouarbya, and Jamie A Ward. Synthetic sensor data for human activity recognition. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9. IEEE, 2020.

Nefeli Andreou, Xi Wang, Victoria Fernández Abrevaya, Marie-Paule Cani, Yiorgos Chrysanthou, and Vicky Kalogeiton. Lead: Latent realignment for human motion diffusion. *Computer Graphics Forum*, 44, 2024. URL https://api.semanticscholar.org/CorpusID:273482704.

Germán Barquero, Sergio Escalera, and Cristina Palmero. Belfusion: Latent diffusion for behavior-driven human motion prediction. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2317–2327, 2022. URL https://api.semanticscholar.org/CorpusID:254017723.

Xiaoyu Bie, Wen Guo, Simon Leglaive, Lauren Girin, Francesc Moreno-Noguer, and Xavier Alameda-Pineda. Hit-dvae: Human motion generation via hierarchical transformer dynamical vae. *arXiv preprint arXiv:2204.01565*, 2022.

Rafal Broda, Alexander Schnerring, Dominik Schnaus, Michael Nieslony, Julian J. Krauth, Marc Röger, Sonja Kallio, Rudolph Triebel, and Robert Pitz-Paal. Bridging the sim2real gap: Training deep neural networks for heliostat detection with purely synthetic data. *Solar Energy*, 300:113728, 2025. ISSN 0038-092X. doi: https://doi.org/10.1016/j.solener.2025.113728. URL https://www.sciencedirect.com/science/article/pii/S0038092X25004918.

Francisco M. Calatrava-Nicolás, Shoko Miyauchi, Vítor Fortes Rey, Paul Lukowicz, Todor Stoyanov, and Óscar Martínez Mozos. Embedded inter-subject variability in adversarial learning for inertial sensor-based human activity recognition. *2025 IEEE 35th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2025. URL https://api.semanticscholar.org/CorpusID:282324303.

Kaixuan Chen, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, and Yunhao Liu. Deep learning for sensor-based human activity recognition: Overview, challenges and opportunities, 2021. URL https://arxiv.org/abs/2001.07416.

Shuo Cheng, Liqian Ma, Zhenyang Chen, Ajay Mandlekar, Caelan Garrett, and Danfei Xu. Generalizable domain adaptation for sim-and-real policy co-training. *arXiv preprint arXiv:2509.18631*, 2025.

Maynara Donato de Souza, Clesson Roberto Silva Junior, Jonysberg Quintino, André Luis Santos, Fabio QB da Silva, and Cleber Zanchettin. Exploring the impact of synthetic data on human activity recognition tasks. *Procedia Computer Science*, 222:656–665, 2023.

Chance DeSmet, Colin Greeley, and Diane J Cook. Hydra-ts: Enhancing human activity recognition with multi-objective synthetic time series data generation. *IEEE Sensors Journal*, 2024.

Markos Diomataris, Nikos Athanasiou, Omid Taheri, Xi Wang, Otmar Hilliges, and Michael J. Black. Wandr: Intention-guided human motion generation, 2024. URL https://arxiv.org/abs/2404.15383.

Ylva Ferstl and Rachel McDonnell. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, IVA '18, pp. 93–98, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360135. doi: 10.1145/3267851.3267898. URL https://doi.org/10.1145/3267851.3267898.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all, 2023. URL https://arxiv.org/abs/2305.05665.

Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, pp. 2021–2029. ACM, October 2020. doi: 10.1145/3394171.3413635. URL http://dx.doi.org/10.1145/3394171.3413635.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5142–5151, 2022. URL https://api.semanticscholar.org/CorpusID:250602257.

Xuemin Hu, Shen Li, Tingyu Huang, Bo Tang, Rouxing Huai, and Long Chen. How simulation helps autonomous driving:a survey of sim2real, digital twins, and parallel intelligence, 2023. URL https://arxiv.org/abs/2305.01263.

Ali Olow Jimale and Mohd Halim Mohd Noor. Subject variability in sensor-based activity recognition. *Journal of Ambient Intelligence and Humanized Computing*, 14:3261–3274, 2021. URL https://api.semanticscholar.org/CorpusID:239679069.

Feiyang Kang, Newsha Ardalani, Michael Kuchnik, Youssef Emad, Mostafa Elhoushi, Shubhabrata Sengupta, Shang-Wen Li, Ramya Raghavendra, Ruoxi Jia, and Carole-Jean Wu. Demystifying synthetic data in llm pre-training: A systematic study of scaling laws, benefits, and pitfalls, 2025. URL https://arxiv.org/abs/2510.01631.

Aliasghar Khani, Arianna Rampini, Bruno Roy, Larasika Nadela, Noa Kaplan, Evan Atherton, Derek Cheung, and Jacky Bibliowicz. Motion generation: A survey of generative approaches and benchmarks. *ArXiv*, abs/2507.05419, 2025. URL https://api.semanticscholar.org/CorpusID:280148631.

George Kontogiannis, Pantelis Tzamalis, and Sotiris Nikoletseas. Exploring the impact of synthetic data on human gesture recognition tasks using gans. In *2024 20th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)*, pp. 384–391. IEEE, 2024.

Kai Kunze and Paul Lukowicz. Sensor placement variations in wearable activity recognition. *IEEE Pervasive Computing*, 13(4):32–41, 2014. doi: 10.1109/MPRV.2014.73.

Tianyu Li, Calvin Qiao, Guanqiao Ren, KangKang Yin, and Sehoon Ha. Aamdm: Accelerated auto-regressive motion diffusion model. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1813–1823, 2023. URL https://api.semanticscholar.org/CorpusID:266977570.

Xi'ang Li, Jinqi Luo, and Rabih Younes. Activitygan: Generative adversarial networks for data augmentation in sensor-based human activity recognition. In *Adjunct proceedings of the 2020 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2020 ACM international symposium on wearable computers*, pp. 249–254, 2020.

Zhe Li, Weihao Yuan, Yisheng He, Lingteng Qiu, Shenhao Zhu, Xiaodong Gu, Weichao Shen, Yuan Dong, Zilong Dong, and Laurence T. Yang. Lamp: Language-motion pretraining for motion generation, retrieval, and captioning. *ArXiv*, abs/2410.07093, 2024. URL https://api.semanticscholar.org/CorpusID:273233776.

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, 2024. URL https://arxiv.org/abs/2307.00818.

Qian Liu, Junchen Ye, Haohan Liang, Leilei Sun, and Bowen Du. Ts-mae: A masked autoencoder for time series representation learning. *Information Sciences*, 690:121576, 2025. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2024.121576. URL https://www.sciencedirect.com/science/article/pii/S0020025524014907.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. *SMPL: A Skinned Multi-Person Linear Model*. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023. ISBN 9798400708978. URL https://doi.org/10.1145/3596711.3596800.

Yingzhou Lu, Lulu Chen, Yuanyuan Zhang, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, Tianfan Fu, and Wenqi Wei. Machine learning for synthetic data generation: a review. *arXiv preprint arXiv:2302.04062*, 2023.

Marcos Lupión, Federico Cruciani, Ian Cleland, Chris Nugent, and Pilar M Ortigosa. Data augmentation for human activity recognition with generative adversarial networks. *IEEE Journal of Biomedical and Health Informatics*, 28(4):2350–2361, 2024.

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, 2024. URL https://arxiv.org/abs/2406.09905.

Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam Tailor, Jake 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, 2024. URL https://arxiv.org/abs/2410.13638.

Luca Oneto, Sandro Ridella, and Davide Anguita. Informed machine learning: Excess risk and generalization. *Neurocomputing*, 646:130521, 2025. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2025.130521. URL https://www.sciencedirect.com/science/article/pii/S0925231225011932.

Mathis Petrovich, Michael J. Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions, 2022. URL https://arxiv.org/abs/2204.14109.

Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model, 2024. URL https://arxiv.org/abs/2312.03596.

Qiaosong Qi, Le Zhuo, Aixi Zhang, Yue Liao, Fei Fang, Si Liu, and Shuicheng Yan. Diffdance: Cascaded human motion diffusion model for dance generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pp. 1374–1382. ACM, October 2023. doi: 10.1145/3581783.3612307. URL http://dx.doi.org/10.1145/3581783.3612307.

Sigal Raab, Inbal Leibovitch, Guy Tevet, Moab Arar, Amit H. Bermano, and Daniel Cohen-Or. Single motion diffusion. *ArXiv*, abs/2302.05905, 2023. URL https://api.semanticscholar.org/CorpusID:256827051.

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, 2021. URL https://arxiv.org/abs/2103.00020.

Zineb Senane, Lele Cao, Valentin Leonhard Buchner, Yusuke Tashiro, Lei You, Pawel Andrzej Herman, Mats Nordahl, Ruibo Tu, and Vilhelm von Ehrenheim. Self-supervised learning of time series representation via diffusion process and imputation-interpolation-forecasting mask. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 2560–2571, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671673. URL https://doi.org/10.1145/3637528.3671673.

Fatemeh Serpush, Mohammad Bagher Menhaj, Behrooz Masoumi, and Babak Karasfi. Wearable sensor-based human activity recognition in the smart healthcare system. *Computational Intelligence and Neuroscience*, 2022, 2022. URL https://api.semanticscholar.org/CorpusID:247136158.

Jonathan CF Silva, Mateus C Silva, Vicente JP Amorim, Pedro SO Lazaroni, and Ricardo AR Oliveira. Synthetic data generation with few-shot adaptation for har using wearable sensors via conditional seriesgan. In *2025 XV Symposium on Computing Systems Engineering (SBESC)*, pp. 1–6. IEEE, 2025.

Zhihang Song, Zimin He, Xingyu Li, Qiming Ma, Ruibo Ming, Zhiqi Mao, Huaxin Pei, Lihui Peng, Jianming Hu, Danya Yao, and Yi Zhang. Synthetic datasets for autonomous driving: A survey. *IEEE Transactions on Intelligent Vehicles*, 9(1):1847–1864, January 2024. ISSN 2379-8858. doi: 10.1109/tiv.2023.3331024. URL http://dx.doi.org/10.1109/TIV.2023.3331024.

Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor S. Prentow, Mikkel Baun Kjærgaard, Anind K. Dey, Tobias Sonne, and Mads Møller Jensen. Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition. *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*, 2015. URL https://api.semanticscholar.org/CorpusID:5687667.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model, 2022. URL https://arxiv.org/abs/2209.14916.

Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision*, 2018. URL https://api.semanticscholar.org/CorpusID:273160019.

Xin Wang, Qiudi Chen, and Wanliang Wang. 3d human motion editing and synthesis: A survey. *Computational and Mathematical methods in medicine*, 2014(1):104535, 2014.

Yilin Wang, Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Xinxin Zuo, Juwei Lu, Hai Jiang, and Li Cheng. Motiondreamer: One-to-many motion synthesis with localized generative masked transformer, 2025. URL https://arxiv.org/abs/2504.08959.

Zhiyong Wang, Jinxiang Chai, and Shihong Xia. Combining recurrent neural networks and adversarial training for human motion synthesis and control. *IEEE Transactions on Visualization and Computer Graphics*, 27(1):14–28, 2021. doi: 10.1109/TVCG.2019.2938520.

Maxwell A. Xu, Jaya Narain, Gregory Darnell, Haraldur Hallgrimsson, Hyewon Jeong, Darren Forde, Richard Fineman, Karthik J. Raghuram, James M. Rehg, and Shirley Ren. Relcon: Relative contrastive learning for a motion foundation model for wearable data, 2025. URL https://arxiv.org/abs/2411.18822.

Kazuki Yamamoto and Makoto Murakami. Human motion generation with stylegan. In *Other Conferences*, 2024. URL https://api.semanticscholar.org/CorpusID:269809173.

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.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model, 2023. URL https://arxiv.org/abs/2212.02500.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model, 2022. URL https://arxiv.org/abs/2208.15001.

Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European Conference on Computer Vision*, 2021. URL https://api.semanticscholar.org/CorpusID:251040567.

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, 2024. URL https://arxiv.org/abs/2410.19818.

Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. *CoRR*, abs/2009.13303, 2020. URL https://arxiv.org/abs/2009.13303.

Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:2430–2449, 2023. URL https://api.semanticscholar.org/CorpusID:263796023.

A APPENDIX

Table 3: Synthetic Data Volume. Synthetic data volume scaling relative to the real-world dataset. The multipliers indicate the ratio of generated motion sequences compared to the original motion capture set size.

Volume Ratio	No. of Generated Sequences
$1\times$	$24,660$
$2\times$	$49,320$
$4\times$	$98,640$
$8\times$	$197,280$