CALIBRATION SLIDE

Who am I?

- Statistics and Machine Learning (StatML) PhD Student

- Previously, I studied Computer Engineering as an undergraduate degree at New York University Abu Dhabi

جامعة نيويورك أبوظبي

NYU | ABU DHABI

- <u>Previous research:</u>
  - High Density sEMG for Hand Gesture Detection
  - Multimodal Prosthetic Control Systems
  - Pretraining CXR and EHR encoders for clinical diagnosis

Clinical AI Lab

CENTER FOR ARTIFICIAL INTELLIGENCE AND ROBOTICS

Motion Capture is Not the Target Domain: Scaling Synthetic Data for Learning Motion Representations

Firas Darwish

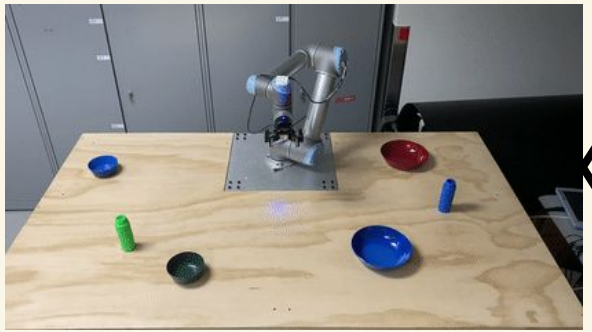
Synthetic Data

- Machine Learning runs on data (lots of data!)

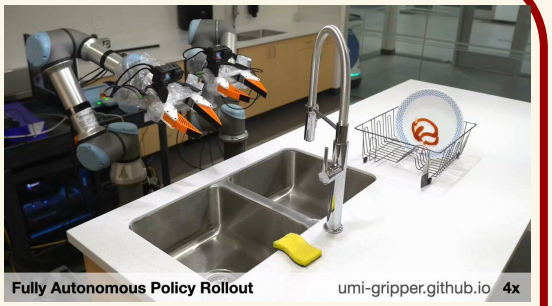
- In many regimes, collective **diverse**, **large** real-world datasets can be:
  - Expensive
  - Slow
  - Impractical

- What are our other options?
  - **Synthetic data** holds significant promise for addressing data scarcity.

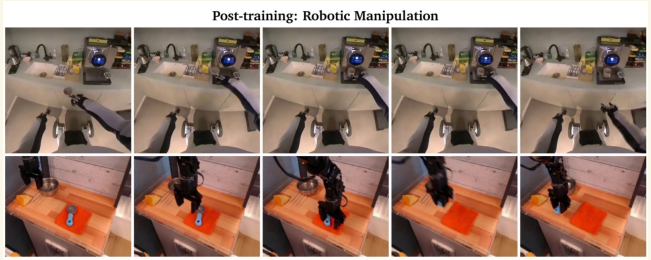
Synthetic Data (in Other Fields)*



Real-World Data Collection



Collect data manually → Train control policy on collected trajectories



"World models" that can generate synthetic trajectories in any environment → Generalist Robot policies

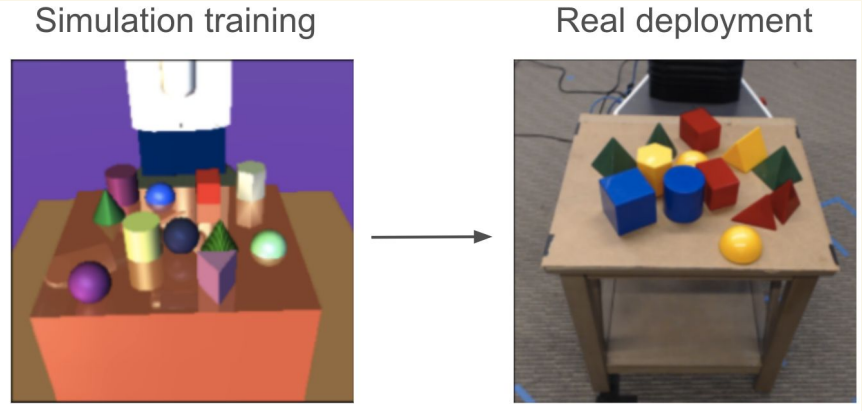
Synthetic Data

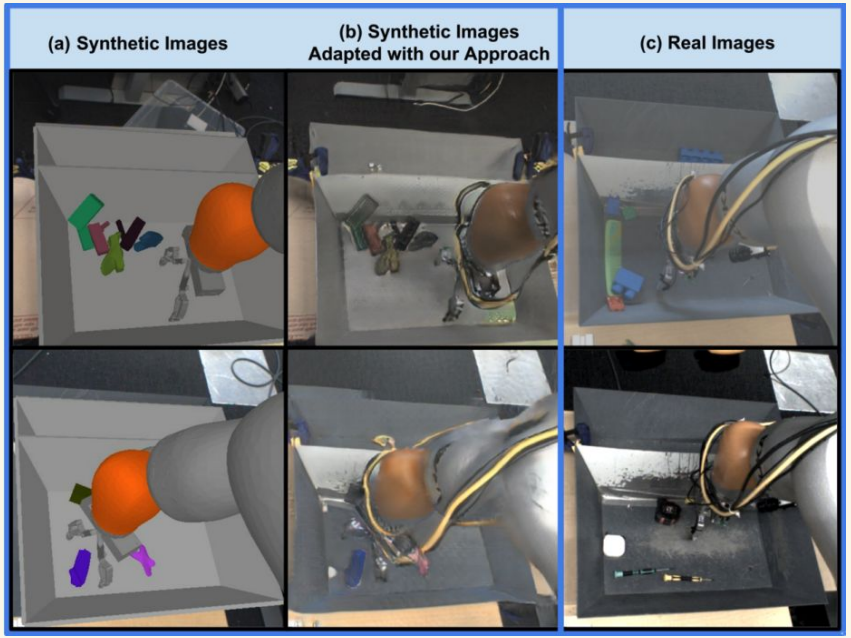
Synthetic Data (in Other Fields)*

Prompt: The video captures a highway scene with a white truck in the foreground, moving towards the camera. The truck has a large cargo area and is followed by a motorcyclist wearing a full-face helmet. The road is marked with white lines and has a metal guardrail on the right side. The sky is partly cloudy, and there are green trees and bushes visible on the roadside. The video is taken from a moving vehicle, as indicated by the motion blur and the changing perspective of the truck and motorcyclist.

Prompt: The footage shows a multi-car pile-up on a foggy highway. Visibility is severely reduced due to thick fog, with only the taillights of vehicles ahead visible. Suddenly, brake lights flash, and cars begin to swerve and stop abruptly. The highway is cluttered with stopped and crashed vehicles. The surroundings are obscured by fog, adding to the chaos and confusion of the scene.

Sim2Real/Syn2Real Gap

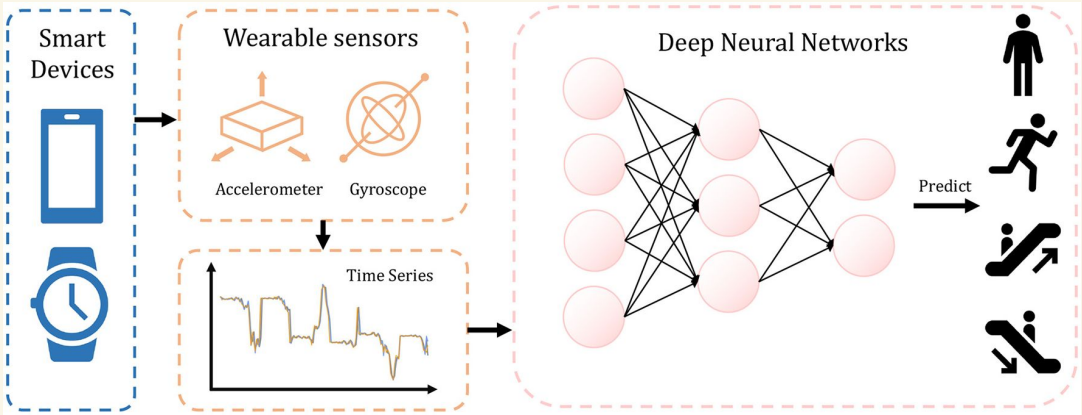




Mismatches in perceptual or dynamical details lead to degraded performance at deployment.

Wearables

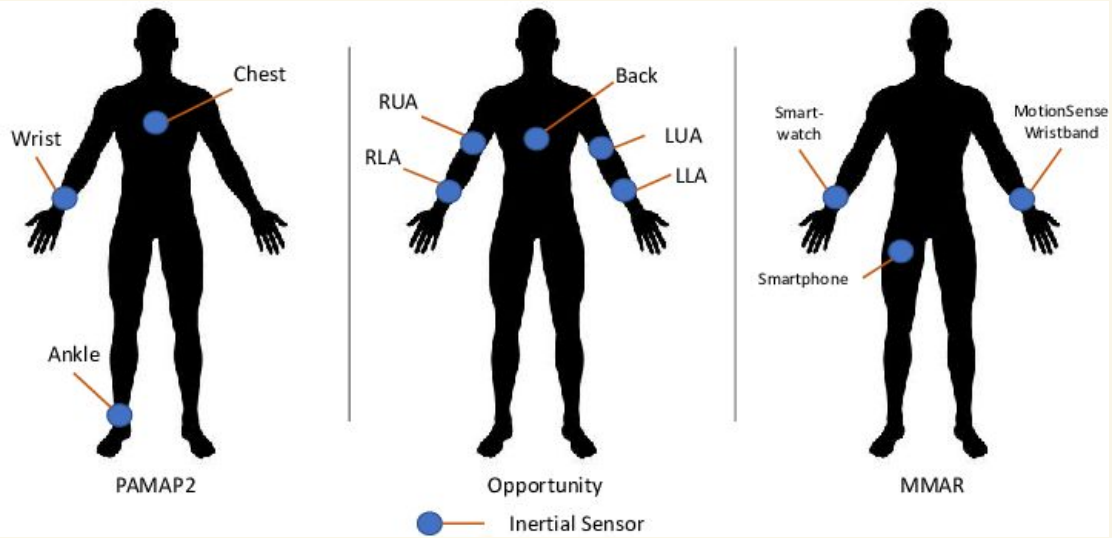
Human Activity Recognition:

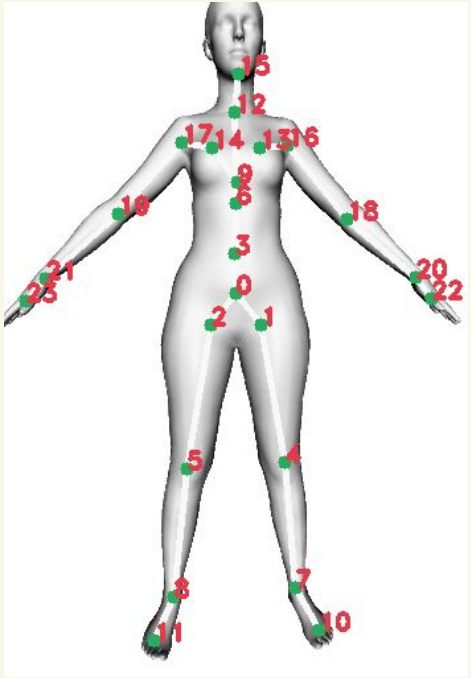


- HAR models trained and evaluated on the same datasets often generalise poorly:
  - Systematic variations in sensors, subjects, and environments

- **Pretraining** on large and diverse motion data is critical for learning representations that transfer effectively across HAR tasks

Wearables





- Pretraining should be done on **full-body** human motion data to allow effective transfer to diverse HAR settings.

Collecting Full-Body Human Motion

Collecting full-body human motion data at scale remains infeasible in practice.

Motion capture datasets are highly accurate but are limited in diversity, volume, and restricted to controlled laboratory environments.

The KIT Whole-Body Human Motion Database

Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, Tamim Asfour
Institute for Anthropomatics and Robotics
Karlsruhe Institute of Technology (KIT), Germany
{mandery, asfour}@kit.edu

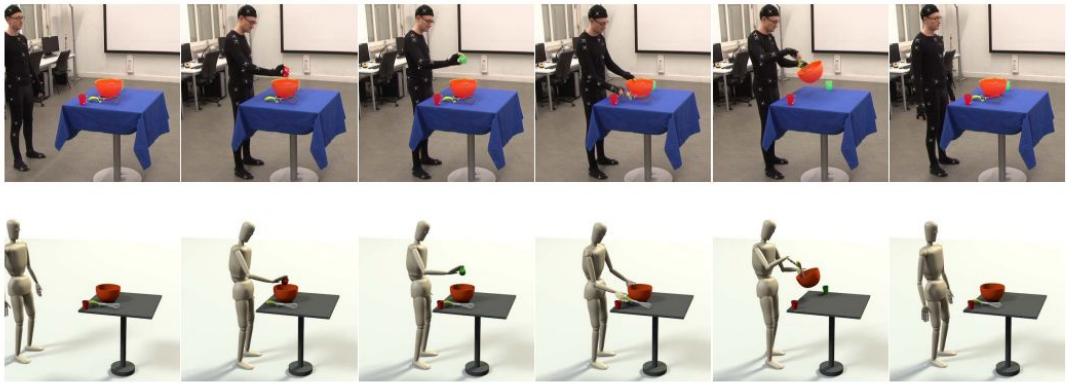
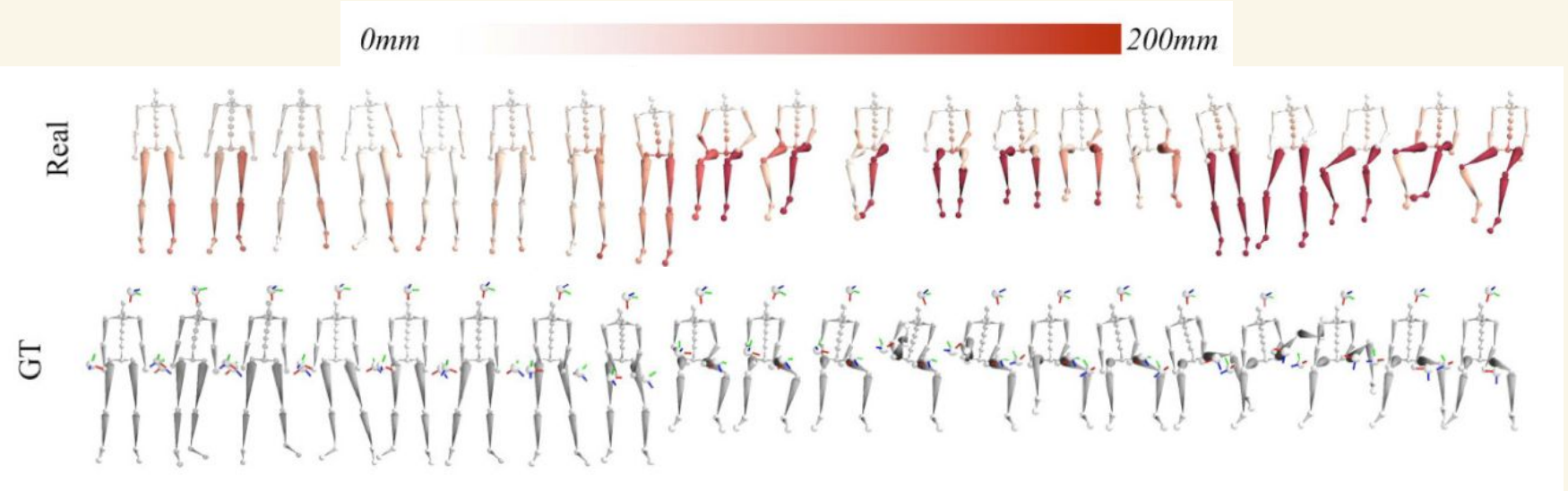


Fig. 1: Preparing the dough: Key frames of a manipulation task available in the motion database which includes four environmental objects.

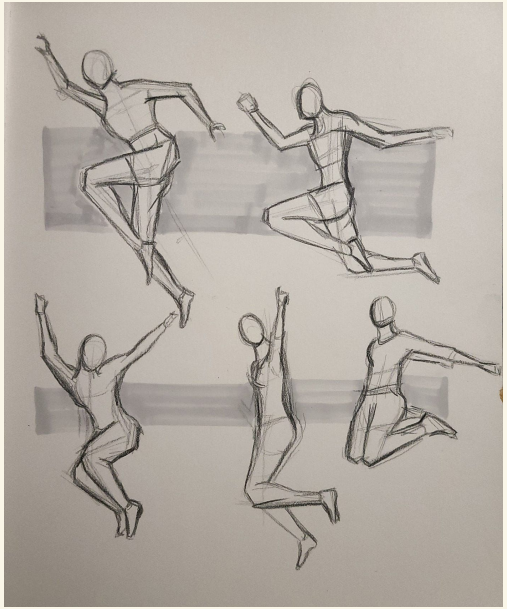
Collecting Full-Body Human Motion

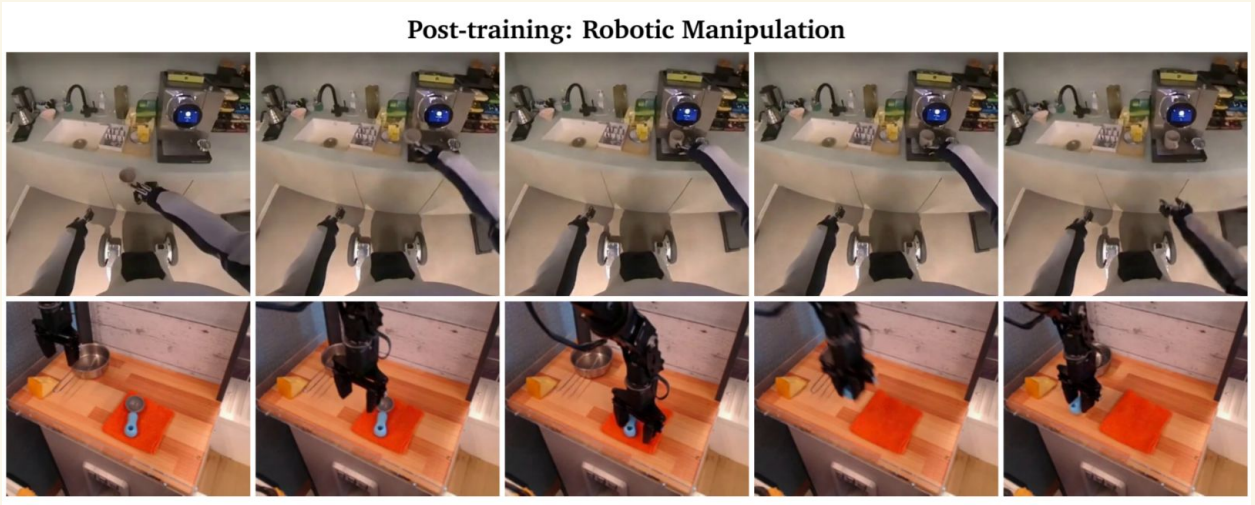
Collecting full-body human motion data at scale remains infeasible in practice.

Large-scale wearable data collection in the wild is limited to wrist motion.



Collecting Full-Body Human Motion





Human motion is governed by well-defined physical and kinematic that must be maintained over extended temporal horizons, such as balance, contact dynamics, and joint limits.

Objective 1 of 3

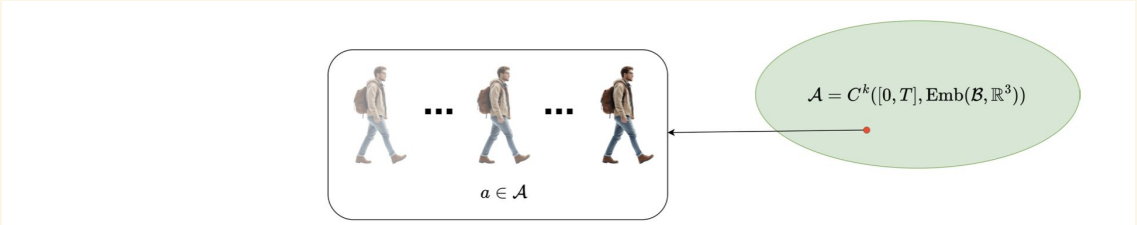
How does the sim2real paradigm apply in the context of wearables?

Motion Representation

- Human body is compact 3D material manifold

- At any time, a body configuration is given by a smooth embedding of this manifold into 3D space.

- The space of all human actions is an infinite-dimensional function space of k-times continuously differentiable functions where k ≥ 2.



Motion Representation

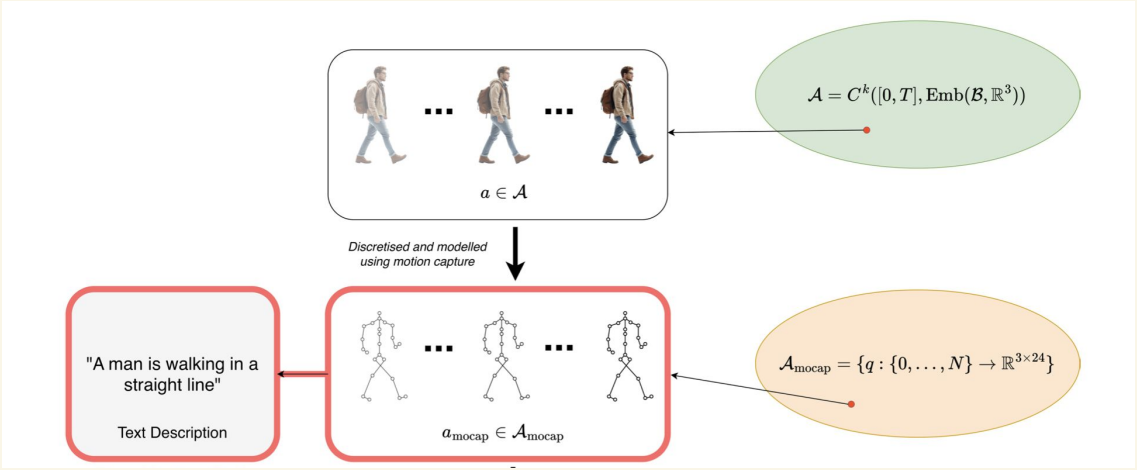
- Human body is compact 3D material manifold

- At any time, a body configuration is given by a smooth embedding of this manifold into 3D space.

- The space of all human actions is an infinite-dimensional function space of k-times continuously differentiable functions where k ≥ 2.

- Real-world sensing systems operate on finite, discretised observations that partially capture this action space.

- We **discretise** our time and **track global position** of a fixed set of 24 anatomical landmarks (joints) selected on the body



Motion Representation

- Human body is compact 3D material manifold

- At any time, a body configuration is given by a smooth embedding of this manifold into 3D space.

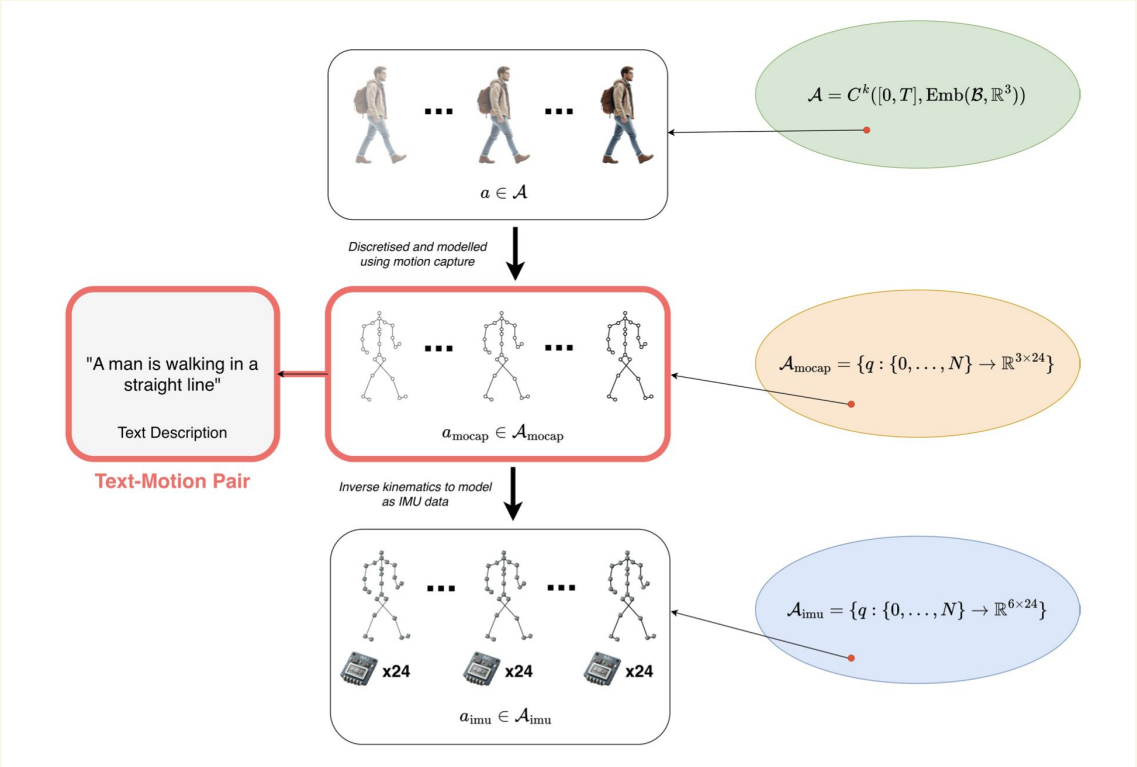
- The space of all human actions is an infinite-dimensional function space of k-times continuously differentiable functions where k ≥ 2.

- Real-world sensing systems operate on finite, discretised observations that partially capture this action space.

- We **discretise** our time and **track global position** of a fixed set of 24 anatomical landmarks (joints) selected on the body

- Because HAR systems operate on data collected from wearable inertial sensors, we transform this space into an IMU-based action space.

- This mapping is obtained via inverse kinematics and differentiation (<0.1% error rate) to produce joint-centric accelerometer and gyroscope signals



Motion Representation

- Human body is compact 3D material manifold

- At any time, a body configuration is given by a smooth embedding of this manifold into 3D space.

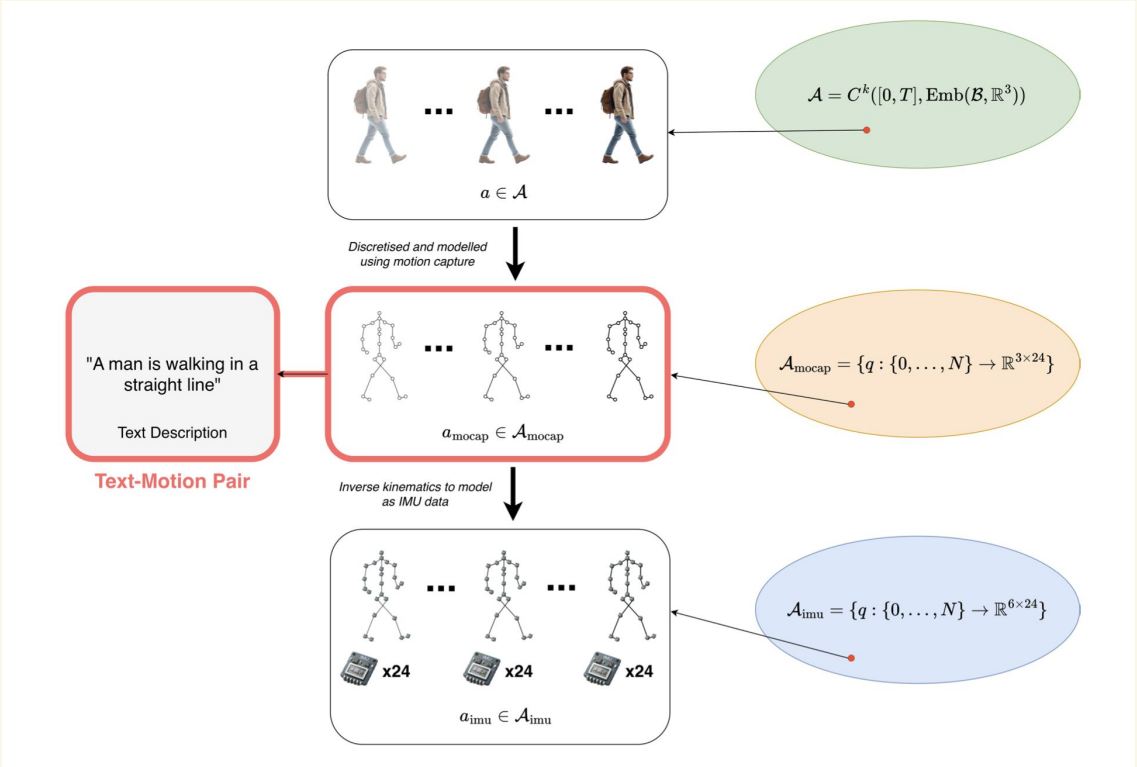
- The space of all human actions is an infinite-dimensional function space of k-times continuously differentiable functions where k ≥ 2.

- Real-world sensing systems operate on finite, discretised observations that partially capture this action space.

- We **discretise** our time and **track global position** of a fixed set of 24 anatomical landmarks (joints) selected on the body

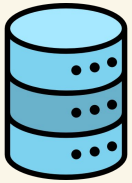
- Because HAR systems operate on data collected from wearable inertial sensors, we transform this space into an IMU-based action space.

- This mapping is obtained via inverse kinematics and differentiation (<0.1% error rate) to produce joint-centric accelerometer and gyroscope signals

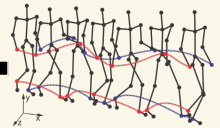


Great! We can get **full-body "wearable" motion** data by transforming **motion capture datasets**.

Generating Synthetic Motion



A person is walking forward



Dataset of real-world text-full body human motion pairs

Training

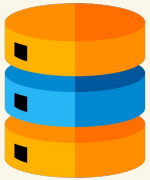
A person is walking forward

z

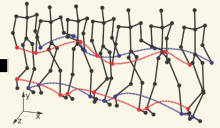
Text Input

Probabilistic latent

Text to Motion Generation Model

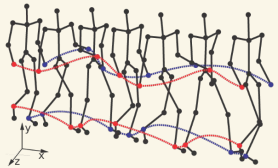


A person is walking forward



*Dataset of **synthetic** text-full body human motion pairs*

Generating



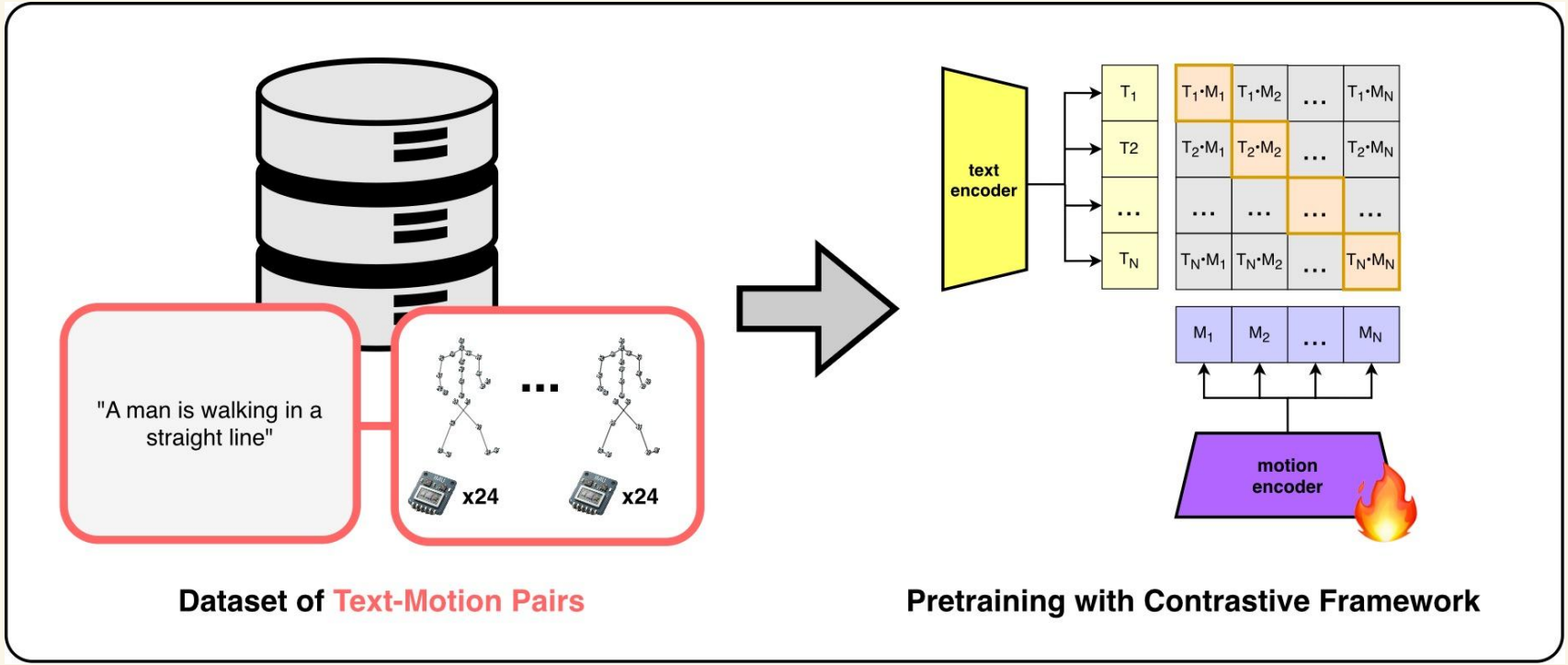
Motion Output

Generating Synthetic Motion

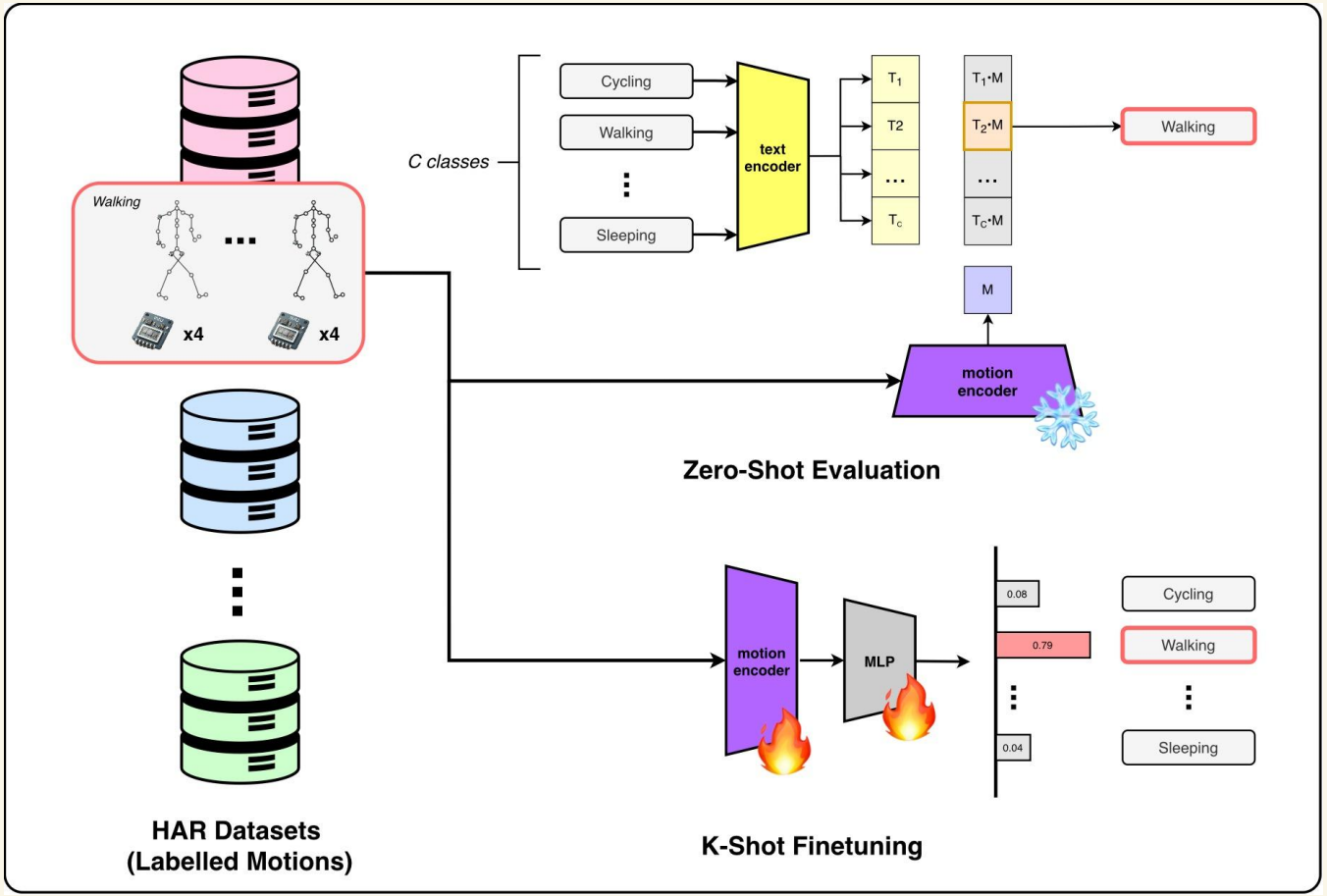
- Our real-world **motion capture** dataset is the **HumanML3D dataset**
  - 24,660 motion sequences
  - average motion sequence length of **7 seconds**
  - average **6 sentences** (text descriptions) per motion

- We use 3 different pre-trained text-to-motion models (all trained on HumanML3D) to generate our **synthetic data**
  - Motion Diffusion Model (MDM)
  - Text-to-Motion (T2M)
  - MotionDiffuse

Pretraining Motion Time-Series Model



Evaluating Motion Time-Series Model



Objective 2 of 3

What happens when we pretrain on synthetic human motion data?

Pretraining with Mixed Data

We first study **the effects of pretraining data composition** by comparing real-world data, synthetic data, and their mixtures while holding the total pretraining volume fixed.

Configurations:

- **Real-World Baseline**: 24,660 real-world text-motion pairs
- **Purely Synthetic Data:** 24,660 synthetic text-motion pairs generated by each text-to-motion model
- **Mixed Data**: A randomly sampled 50/50 mixture of real-world and synthetic text-motion pairs.

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**

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 **1x**, **2x**, **4x**, and **8x** the size of the real-world dataset.

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×	24,660
2×	49,320
4×	98,640
8×	197,280

Pretraining with Synthetic Data

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.

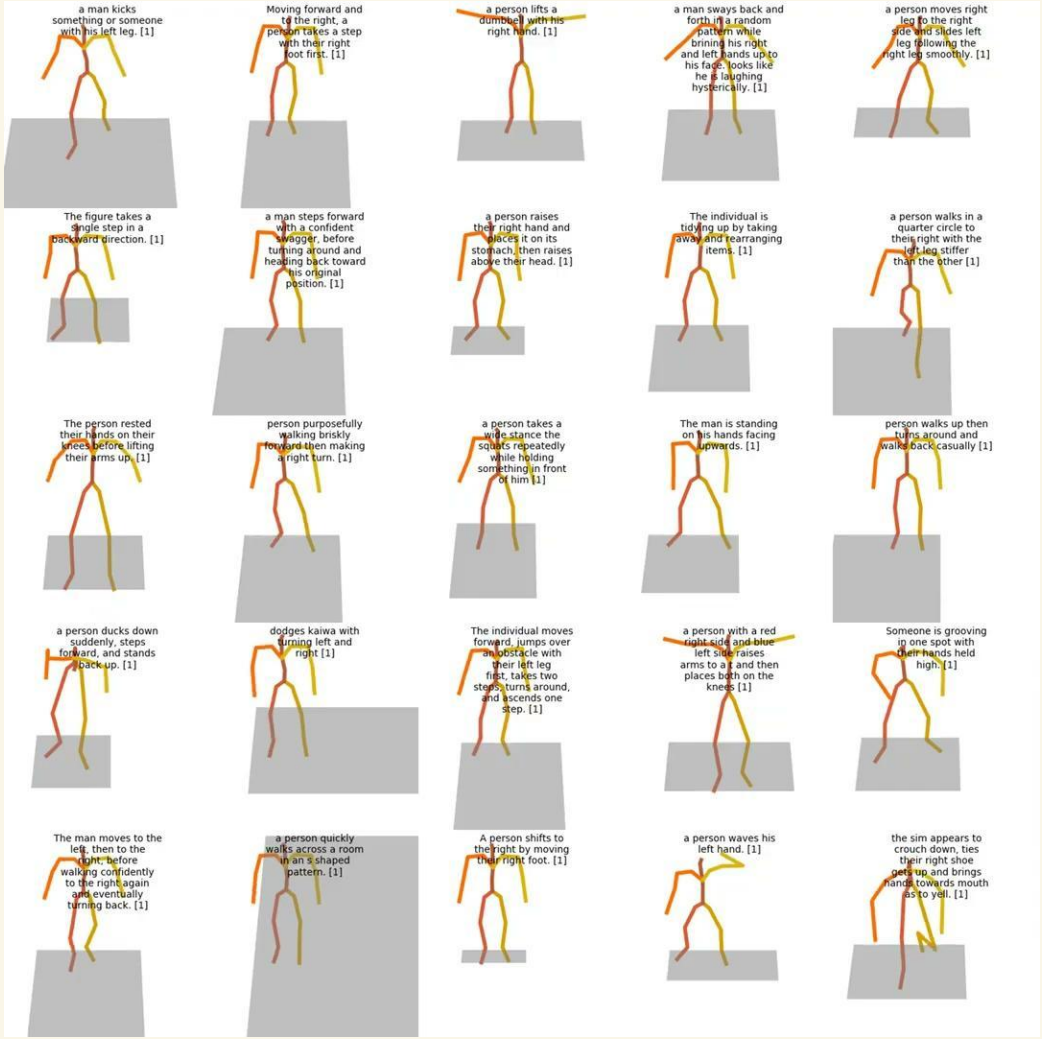
Method	0-shot	1-shot	2-shot	3-shot	5-shot	10-shot
Real-World Baseline	**0.3070**	0.5445	0.6243	0.6611	0.7041	0.7667
MDM (Synthetic)						
1×	0.2460	0.5120	0.6160	0.6559	0.6974	0.7653
2×	0.2928	0.5334	0.6089	0.6505	0.7008	0.7368
4×	0.2546	0.5299	0.6344	0.6749	0.7023	0.7663
8×	0.2654	**0.5517**	0.6266	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

- **Consistent gap** remains in the 0-shot setting!

- Suggests that real-world data induces representations that generalise more effectively **out-of-the-box**, whereas synthetic pretraining excels under **fine-tuning**.

- Performance improves as synthetic data scales but **trend is not strictly monotonic**

Possible challenges with the semantic expressiveness of sampled prompts or randomness in the sampled motion lengths



- Performance improves as synthetic data scales but **trend is not strictly monotonic**

Objective 3 of 3

Is there an issue with using motion capture data?

Scaling Real-World Data

We study the **effect of scaling** real-world (motion capture) data by:

1. randomly subsampling 10%, 20%, 40%, and 80% of the available text-motion pairs
2. pretraining models on each subset
3. Using three random seeds

Why?

1. Examine the stability of *k*-shot evaluations by quantifying variability across runs (why the 0-shot performance gap?)

2. (*commonly held expectation*) increased pretraining data volume **=** improved downstream performance

   Does this apply when the data is motion capture?

Scaling Real-World Data

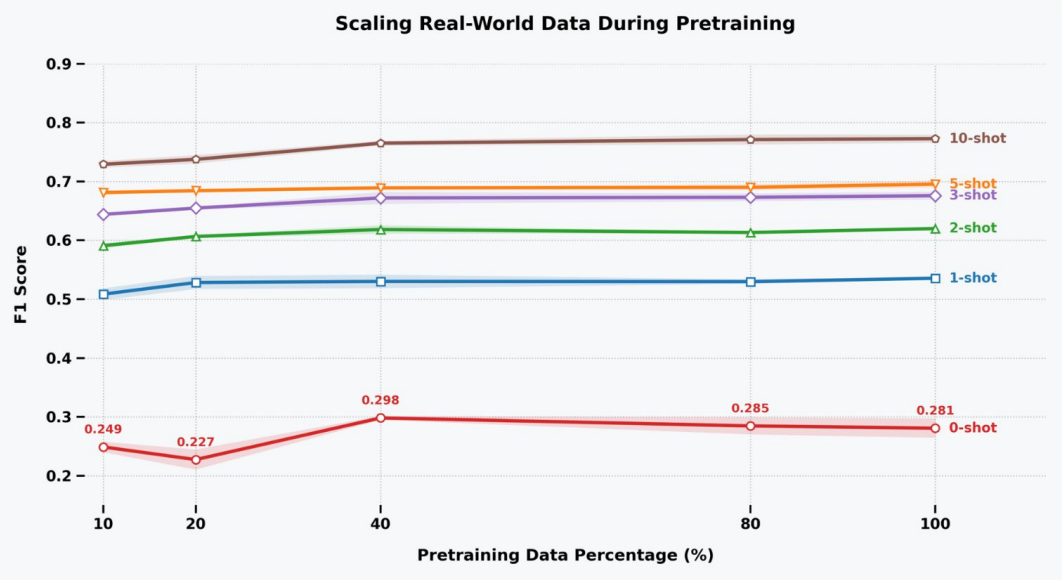


Figure 2: Fine-tuning stabilises gains from real-world pretraining. While increasing real-world pretraining data yields small but consistent improvements in fine-tuned (k-shot) performance, zero-shot performance remains unstable and does not scale reliably with data volume. Curves show results from three 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.

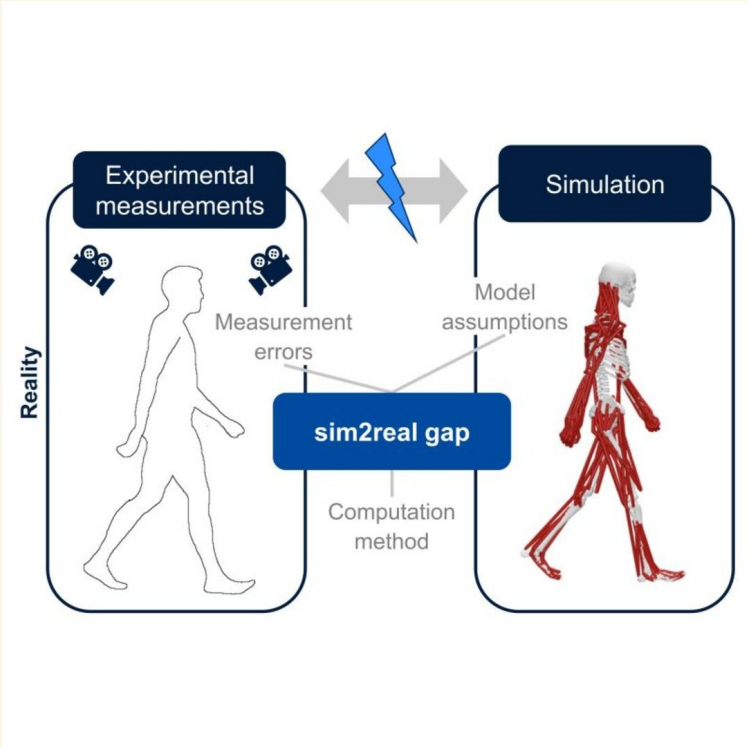
- **0-shot: high performance variability and inconsistent scaling!**

- **Fine-tuning setting:** pretraining on 40% of the data nearly matches performance obtained using the full dataset. ***Highly unexpected***

Domain Mismatch!

Pretraining on motion capture may emphasise factors that are *weakly aligned* with the signal characteristics required for wearable-based HAR

Motion capture may not be an appropriate domain for learning transferable motion representations here.



Important Notice on Generation

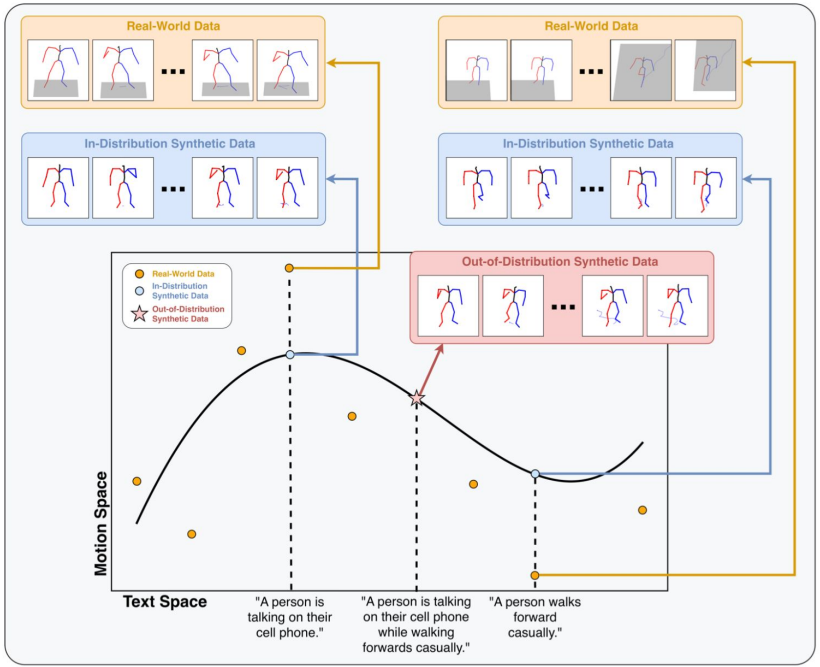


Figure 3: Using the Text-to-Motion Model. Assuming text and motion samples are embedded and projected into a shared low-dimensional space for visualisation, where the black curve represents the text-to-motion model (shown as deterministic mapping here). In-distribution synthetic data is obtained by sampling 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.

Thank you!

MOTION CAPTURE IS NOT THE TARGET DOMAIN: SCALING SYNTHETIC DATA FOR LEARNING MOTION REPRESENTATIONS

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.