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

Firas Darwish [1] George Nicholson [1] Aiden Doherty [1] Hang Yuan [1]

[1]University of Oxford

Introduction

Synthetic Data helps address data scarcity in regimes where collecting large-scale, diverse real-world datasets is impractical (e.g. robotics, autonomous driving).

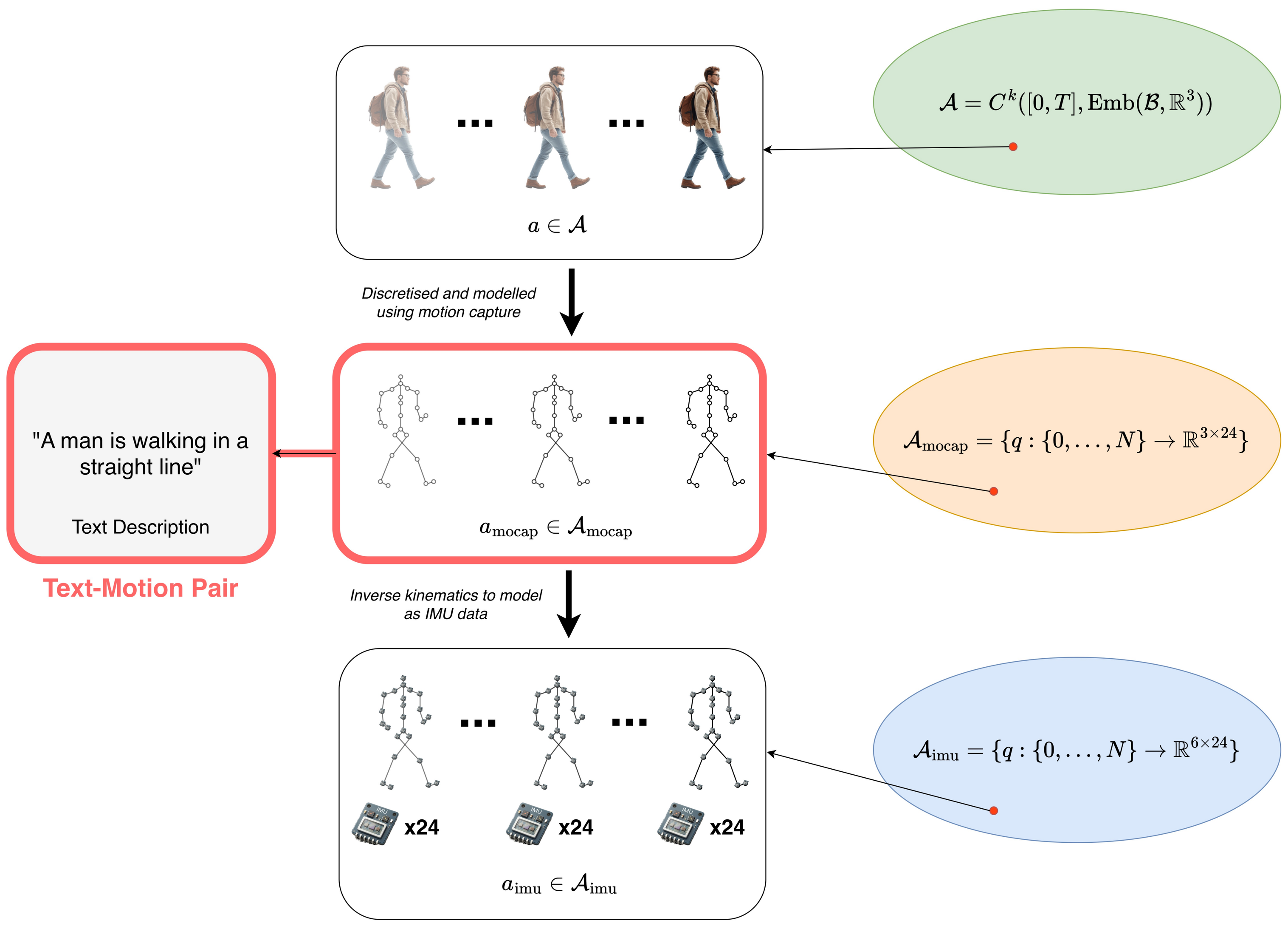
Human Activity Recognition (HAR) models use human motion data from wearable devices to classify activities over time.

Pretraining on large and diverse full-body human motion data is critical for learning representations that **generalise** effectively across HAR tasks.

Full-Body Human Motion

What is the **limitation** with current full-body human motion data?

- **Motion capture** datasets are accurate but are limited in both diversity and volume and restricted to controlled laboratory environments.
- Large-scale **wearable data** collection in the wild is limited to wrist motion.



Method

1. We **generate** synthetic motion datasets by sampling motion trajectories from trained text-conditioned motion models (MDM, T2M, MotionDiffuse).

2. We **pretrain** a motion encoder jointly with a pretrained text encoder using **contrastive learning** on paired text-motion data.

3. We **evaluate** downstream performance in $\{0, 1, 2, 3, 5, 10\}$-shot settings on 18 HAR datasets (reporting macro-averaged F1) that vary in number and placement of sensors, activity classes, and number of subjects.

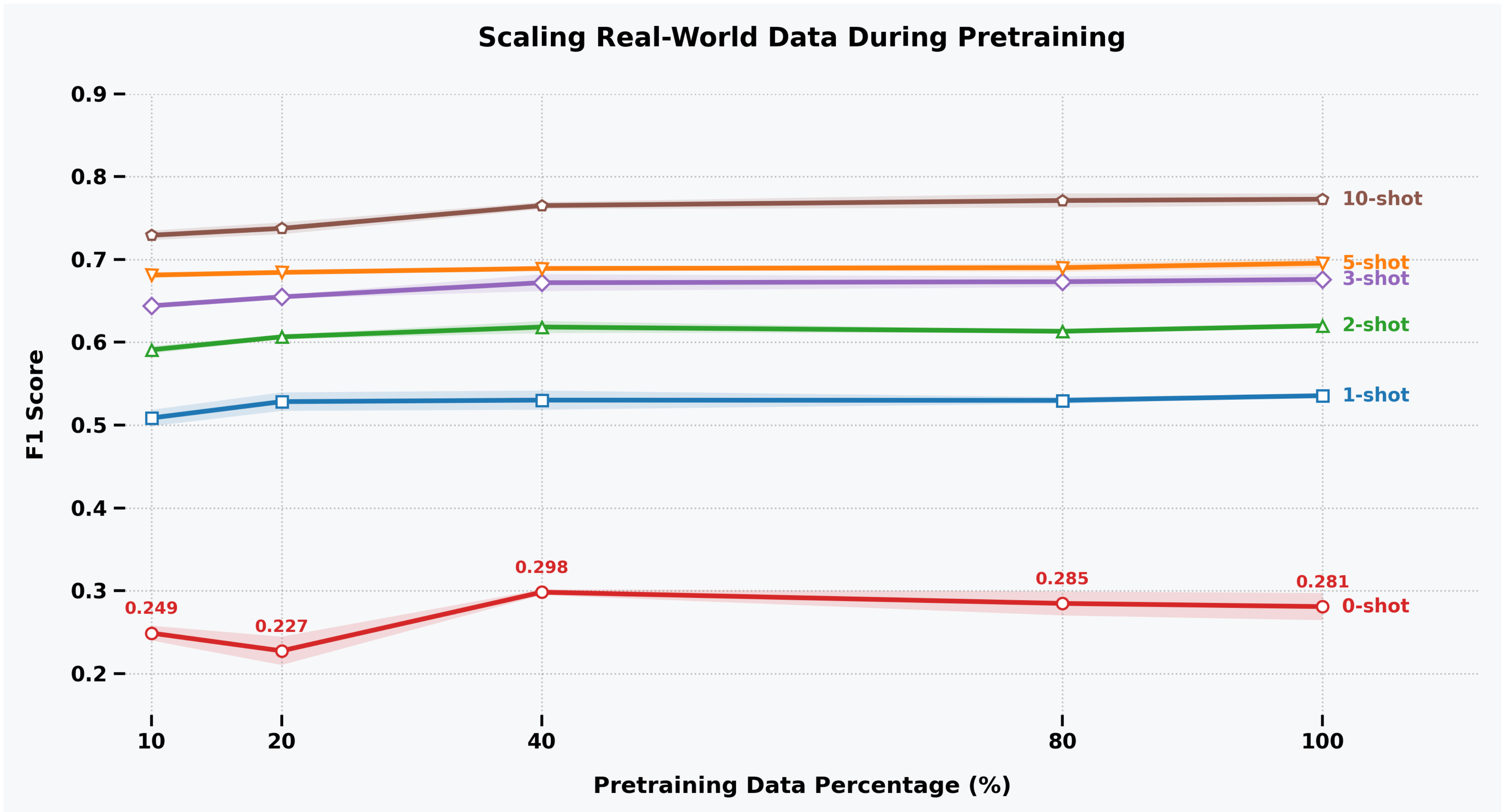
Scaling Synthetic Data

At sufficient scale (8x), synthetic pretraining outperforms real-world pretraining in all downstream HAR settings, with the exception of the 0-shot regime.

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

Scaling Real Data

Increasing real-world pretraining data yields small but consistent improvements in fine-tuned performance but 0-shot performance remains unstable and does not scale reliably with data volume.



Conclusions

Generalization 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.

Domain mismatch! Pretraining on motion capture may emphasize factors that are weakly aligned with the signal characteristics required for wearable-based HAR.

ICLR International Conference On Learning Representations

BIG DATA INSTITUTE

UNIVERSITY OF OXFORD