Generating Virtual On-body Accelerometer Data from Virtual Textual Descriptions for Human Activity Recognition

Zikang Leng
zleng7@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Hyeokhyen Kwon
hyeokhyen.kwon@emory.edu
Emory University
Atlanta, Georgia, USA

Thomas Plötz
thomas.ploetz@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

ABSTRACT

The development of robust, generalized models for human activity recognition (HAR) has been hindered by the scarcity of large-scale, labeled data sets. Recent work has shown that virtual IMU data extracted from videos using computer vision techniques can lead to substantial performance improvements when training HAR models combined with small portions of real IMU data. Inspired by recent advances in motion synthesis from textual descriptions and connecting Large Language Models (LLMs) to various AI models, we introduce an automated pipeline that first uses ChatGPT to generate diverse textual descriptions of activities. These textual descriptions are then used to generate 3D human motion sequences via a motion synthesis model, T2M-GPT, and later converted to streams of virtual IMU data. We benchmarked our approach on three HAR datasets (RealWorld, PAMAP2, and USC-HAD) and demonstrate that the use of virtual IMU training data generated using our new approach leads to significantly improved HAR model performance compared to only using real IMU data. Our approach contributes to the growing field of cross-modality transfer methods and illustrate how HAR models can be improved through the generation of virtual training data that do not require any manual effort.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing**; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Virtual IMU Data; Activity recognition; Motion Synthesis; Large Language Models; Wearable Sensors

ACM Reference Format:
Zikang Leng, Hyeokhyen Kwon, and Thomas Plötz. 2023. Generating Virtual On-body Accelerometer Data from Virtual Textual Descriptions for Human Activity Recognition. In *Proceedings of the 2023 International Symposium on Wearable Computers (ISWC '23), October 08–12, 2023, Cancun, Quintana Roo, Mexico.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3594738.3611361

1 INTRODUCTION

The development of accurate and robust predictive models for human activity recognition (HAR) is essential for, e.g., monitoring fitness, analyzing health-related behavior, and improving industrial processes [2, 4, 14, 23]. However, one of the major challenges in HAR research is the scarcity of labeled activity data, which hinders the effectiveness of supervised learning methods [5].

To address this challenge, researchers have explored methods for acquiring labeled data that are more flexible and cost-effective. One such method is the automated generation of virtual IMU data. In recent years, effective *cross-modality transfer* approaches [10–12] have been developed for extracting virtual IMU data from 2D RGB videos of human activities. Virtual IMU data can expand training datasets for motion exercise recognition and can be used to build personalized HAR systems that meet the diverse needs of individual users [28]. By leveraging the advantages of virtual IMU data, researchers can improve the accuracy and robustness of HAR models and facilitate the widespread adoption of sensor-based HAR in a variety of domains.

In this work, we present a method that can *generate diverse textual descriptions of activities that can then be converted to streams of virtual IMU data.* In our automated pipeline, the name of an activity is first passed to ChatGPT to automatically generate textual prompts that describe–in plain language–a person doing the activity.

The generated textual prompts are then used to generate 3D human motion using a motion synthesis model, which can then be converted to streams of virtual IMU data. By using ChatGPT to generate the diverse textual descriptions of activities, we can generate virtual IMU data that capture the different variations of how activities can be performed. With ChatGPT, no prompt engineering is needed and essentially unlimited amounts of virtual IMU data can be generated.

The contributions of this paper are two-fold:

- We leverage ChatGPT's natural language generation capabilities to automatically generate textual descriptions of activities, which are then used in conjunction with motion synthesis and signal processing techniques to generate virtual IMU data streams. By using this approach, we can significantly reduce the time and cost required for data collection, while covering a wide range of activity variations.
- We evaluate our approach on three standard HAR datasets–Realworld, Pamap2, and USC-HAD–and demonstrate the overall effectiveness through improved activity recognition results across the board for models that utilize virtual IMU data generated through our approach.

The results of our approach are significant – they contribute to the growing field of cross-modality transfer that promises to alleviate the much lamented lack of annotated training data in HAR, thereby requiring virtually no manual effort at all.

ISWC '23, October 08–12, 2023, Cancun, Quintana Roo, Mexico

ACM ISBN 979-8-4007-0199-3/23/10.
https://doi.org/10.1145/3594738.3611361

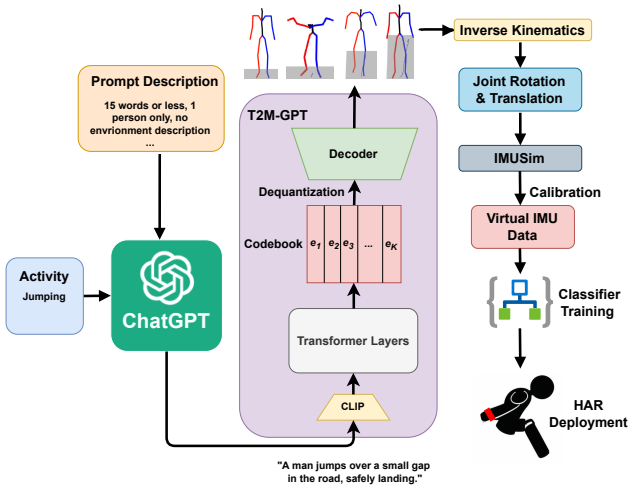


Figure 1: Overview of the proposed approach. The activity name and general description of prompts are provided to ChatGPT for prompts generation. Using the generated prompts, T2M-GPT generates 3D human motion sequences. Using the motion sequences, virtual IMU data are generated using inverse kinematics and IMUSim. After calibrating the virtual IMU data with a small amount of real IMU data, the virtual IMU data can be used to train a deployable classifier.

2 RELATED WORK

Virtual IMU Data Generation: Recently, IMUTube [11] was introduced to extract virtual IMU from 2D RGB videos. IMUTube uses computer vision methods such as 2D/3D pose tracking to extract the 3D human motion in the given video. The extracted 3D human motion information is used to estimate 3D joint rotations and global motion, which is then used generate the virtual IMU data. Previous studies [10, 12] have shown that the extracted virtual IMU data lead to improved model performance when mixed with the real IMU data and allowed for effective training of more complex models.

To improve the quality of the extracted virtual IMU data, Xia *et al.* [28] proposed a spring-joint model to augment the extracted virtual acceleration signal and trained a classifier on the augmented virtual IMU data to recognize reverse lunge, warm up, and high knee tap. Vision-based systems such as IMUTube are limited by the quality of the input videos. In order for the extracted virtual IMU data to be of suitable quality, the videos should exhibit little to no camera egomotion and only include people performing the desired activity. Hence, selecting videos of good quality can be time-consuming. Since our system is text-based, the time-consuming process of selecting videos is eliminated.

Text-driven Human Motion Synthesis: The goal of text-driven Human Motion Synthesis is to generate 3D human motion using textual descriptions. With the recently released HumanML3D [8], the current largest 3D human motion dataset with textual descriptions, numerous models have been introduced that can produce significantly more realistic human motion sequences than previous models. MDM [25], MLD [29], and MotionDiffuse [34] are three recently introduced diffusion-based models. In this work, we use T2M-GPT [33] as the motion synthesis model for our system.

Large Language Models: Large Language Models (LLMs) such as PaLM [6], LLaMA [26], GPT-3 [3], and ChatGPT (built upon InstructGPT [17]) have attracted enormous attention for their superior performances in many natural language processing (NLP) tasks. However, LLMs alone cannot solve complex AI tasks that require processing information from multiple modalities. Recently, Visual ChatGPT [27] and HuggingGPT [22] were introduced to tackle complex multi-modal tasks. Both use ChatGPT as a controller that can divide user input into sub-tasks and select the relevent AI model from a pool of models to solve the complex task. Inspired by this idea, we use ChatGPT as a prompt generator to generate diverse textual descriptions for activities that are then used as input for the motion synthesis model in our system.

Table 1: Real and virtual IMU datasets size for the three HAR datasets we used for evaluation.

Dataset	Real Size	Virtual Size
RealWorld	1,107 min	41 min
PAMAP2	322 min	68 min
USC-HAD	469 min	69 min

3 GENERATING VIRTUAL IMU DATA FROM VIRTUAL TEXTUAL DESCRIPTIONS

The key idea of our approach lies in generating a wide range of diverse textual descriptions for a given activity, and to then feed those textual descriptions into a motion synthesis model that is connected to a virtual IMU data generation pipeline. Fig. 1 provides an overview of the developed approach. Human activities are inherently variable; a person can walk happily, confidently, quickly, or in many other ways. This variability is reflected in the IMU data collected by wearable sensors, which must be accurately represented in the training data to ensure HAR models generalize well. We address this challenge by employing ChatGPT to–automatically–create detailed and varied textual descriptions of activities, which then serve as prompts for 3D human motion synthesis.

During prompt generation, the activity name, few example textual descriptions (not activity specific), and general description of the desired prompts are provided to ChatGPT. The example textual descriptions serves as few-shot examples that ChatGPT can learn from. The prompt description is provided to help align ChatGPT's output with our desired prompts. Some descriptions that we used include: *prompts should be 15 words or less; prompts should only include a single person performing the activity; prompts should not contain extensive description of the environment.* Example generated prompts are shown in Table 2. [1]

The generated prompts are then fed into the motion synthesis model, T2M-GPT [33], to generate 3D human motion sequences.[2] To do so, CLIP [18], a pre-trained text encoder, first extracts the text embedding from the prompt. Using this, a learned transformer generates code indices autoregressively until an end token is generated. The sequence of code indices is de-quantized into latent vectors by looking up the corresponding vector in the codebook for each index. Lastly, a learned decoder maps the sequence of latent vectors to 3D human motion sequence, represented as a sequence of 22 joints' positions.

We estimate each joint's rotation with respect to the parent joint and the root joint's (pelvis) translation using inverse kinematics [30]

[1]All generated prompts can be found at this link
[2]We used the same T2M-GPT model found in (link).

Table 2: Example generated prompts from ChatGPT

Climb up stairs	A person struggles to climb the stairs with a heavy load on their back.
Climb down stairs	Someone holds onto the handrail while walking down a set of stairs.
Jumping	A person jumps up to touch a basketball rim, feeling victorious.
Lying	A man stretches out on a blanket, feeling the grass beneath him.
Running	A girl tries to catch up with her siblings as they race around the house.
Sitting	A man sits confidently while conducting a business meeting.
Standing	A woman stands with her arms in front of her, crossed at the wrists.
Walking	A retiree takes a long, therapeutic walk in the park.

with the joints' positions and the skeleton's hierarchical structure as input.[3] IMUSim [32] is then used to calculate the joint's acceleration movement and angular velocity using the estimated local joints' rotations and root translation. This allows us to extract virtual IMU data from 22 on-body sensor locations. Additionally, IMUSim introduces noises to the generated virtual IMU data to simulate the noises that real IMU data typically exhibit.

Inevitably there will be a domain gap between the virtual IMU data (source) and the real IMU data (target) due to potential differences in coordinate systems, sensor orientations and placements, and the size of real human and virtual skeleton. We employ domain adaptation to bridge the gap between the two domains. Following Kwon *et al.* [11], we perform a distribution mapping between the virtual IMU data and the real IMU data using the rank transformation approach [7]. To calibrate the virtual IMU data, only a small amount (few minutes) of real IMU data is needed.

After calibration, the process of virtual IMU data generation is complete. The extracted virtual IMU data can then be used to train a HAR model either alone or in combination with some real IMU data. Lastly, the trained model is deployed in the real world.

4 EXPERIMENTAL EVALUATION

We evaluated the effectiveness of our approach in a set of experiments where we train activity recognizers for benchmark recognition tasks and analyze the performance (F1 scores) for scenarios where only real, only virtual, and mixtures of real and virtual training data are used, respectively (similar to previous work, e.g., [10–12]).

4.1 Datasets

Real IMU Dataset: To evaluate the value of the virtual IMU data generated by our proposed approach, we use the RealWorld [24], PAMAP2 [19], and USC-HAD [35] datasets (details in Table 1). All real IMU datasets were downsampled to 20 Hz to match the virtual IMU datasets.

Virtual IMU Dataset: To generate the virtual IMU dataset, we used our system to generate 50 clips of virtual IMU data for each activity. Each clip corresponds to a different–automatically generated–prompt from ChatGPT, The length of the clips ranges from five to ten seconds, and the exact length of the clip depends on when the transformer generates the end token, which in turn depends on the textual prompt. The virtual IMU data was extracted from joint locations of the virtual skeleton that were selected to be physically closest to the sensor locations on the subjects.

[3]We converted 3D pose sequences to estimated joint rotations in BVH file format using the BasicInverseKinematics method in this link

Table 3: Model performances (Macro F1) for the experimental evaluation of our approach for the three HAR datasets.

(a) Random Forest

Dataset	PAMAP2	RealWorld	USC-HAD
Real	0.659 ± 0.003	0.715 ± 0.011	0.478 ± 0.002
Virtual	0.628 ± 0.003	0.746 ± 0.003	0.448 ± 0.003
Real+Virtual	**0.699** ± 0.004	**0.770** ± 0.004	**0.486** ± 0.003

(b) DeepConvLSTM

Dataset	PAMAP2	RealWorld	USC-HAD
Real	0.687 ± 0.008	0.796 ± 0.015	**0.646** ± 0.008
Virtual	0.626 ± 0.015	0.681 ± 0.005	0.453 ± 0.008
Real+Virtual	**0.723** ± 0.007	**0.820** ± 0.002	0.640 ± 0.002

4.2 Classifier Training

We perform our evaluation on a Random Forest classifier and on DeepConvLSTM [16]. Sliding windows of two seconds duration and with 50% overlap are used to segment the real and virtual IMU data. For the Random Forest classifier, ECDF features [9] (15 components) are extracted from the windows for training. We train a classifier only on the real IMU data to establish a baseline. Additionally, we trained a classifier on only virtual IMU data and another classifier on both real and virtual IMU data. Only the accelerometry signal is used following Kwon *et al.* [11]. For evaluation, we performed leave-one-subject-out cross-validation on the test real IMU dataset. The training real IMU set is not used when training a classifier only on virtual IMU data. We report macro F1 scores averaged across all folds over three runs and their normal approximation interval.

To evaluate the benefit of the virtual IMU data when different amounts of real IMU data is available, we varied the amount of real IMU data used for training. Starting with 2% of the available to real IMU data, we gradually increased the size of the real IMU dataset for training. The virtual IMU dataset and the testing dataset is left unchanged.

4.3 Results

Results are listed in Table 3. The classifier trained on both real and virtual IMU data shows significant improvement in F1 score compared to a classifier trained only on real IMU data for the PAMAP2 and RealWorld dataset. Furthermore, on the RealWorld dataset, we observe that the Random Forest classifier trained on only virtual IMU data outperforms the classifier trained on real IMU data. We find this surprising because the size of the virtual IMU dataset is less than 4% of the size of the real IMU dataset. We attribute this performance improvement to the diverse textual prompts that ChatGPT generated, which led to a diverse set of virtual IMU clips. Using such a diverse training data, the model learns to recognize the many variations of each activity. Yet, when only virtual IMU dataset was used for DeepConvLSTM, the performance significantly drop due to too small training dataset.

Figure 2 shows the model performances when varying amount real IMU data is used for training. We observe that the classifier trained on both real and virtual IMU data consistently outperform the classifier trained only on real IMU data for varying amount of real IMU data. The performance improvement is especially apparent when the size of the real IMU dataset is greatly reduced. This shows the use of virtual IMU data for training is exceptionally beneficial when the amount of available real IMU data is limited.

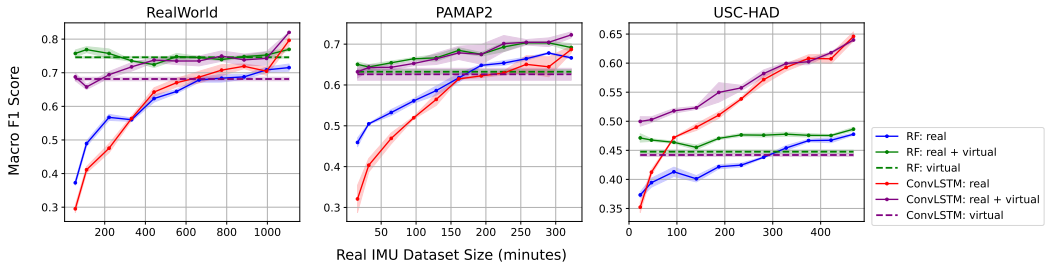
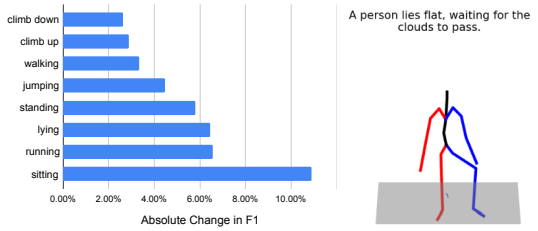


Figure 2: Model performance on RealWorld [24], PAMAP2 [19], and USC-HAD [35] datasets when different amount of real IMU data are used for training. The amount of virtual IMU data used remains the same.



(a) **(b)**

Figure 3: (a) Differences in F1 score for each activity between the classifier trained on only real IMU data and the classifier trained on both real and virtual IMU data evaluated on the RealWorld [24] dataset. (b) Example where the motion synthesis model confused waiting with lying.

5 DISCUSSION

The experimental evaluation demonstrates the effectiveness of our proposed approach. In this section we explore current limitations and outline future directio that could further enhance the utility of our method.

First, the pipeline will only be able to generate virtual IMU data for activities that are described in the HumanML3D dataset. If the prompt contains activities that are not captured by the HumanML3D dataset, our pipeline will fail to generate realistic virtual IMU data for the activity. One potential solution would be to extend the HumanML3D dataset with new activities. A cost-effective method for extension would be to use computer vision techniques such as 3D human pose estimation [31] on existing videos to extract the human motion sequence for the new activities.

Second, the motion synthesis model sometimes confuses closely related activities or two verbs in the same prompt. For instance, T2M-GPT sometimes generates a motion sequence for climbing up the stairs when the input prompt is for climbing down the stairs and vice versa. As per Fig. 3a, climbing up and down stairs gained the least increase in per class F1 score from the addition of virtual IMU data. Also, T2M-GPT confused frequently between walking forward, walking counter-clockwise, and walking clockwise. Since the USC-HAD dataset contains these activities, the classifier trained on the USC-HAD dataset received the least performance improvement from the virtual IMU data. Additionally, T2M-GPT sometimes confuses another verb in the prompt for the activity. As shown in Fig. 3b, T2M-GPT confuses "waiting" with "lies", which causes the generated motion sequence to be more similar to sitting than lying. In our study, we did not manually filter out those failure cases of T2M-GPT, as our goal was to study the feasibility of using text-generated virtual IMU data with minimum manual input to the system.

Our experiment results successfully show that, although with the presence of noisy virtual IMU data from T2M-GPT, the generated virtual IMU data can significantly improve the model performance overall, thus, suggesting the scalability of this approach. In future work, we plan to study the effect of manual cleaning of those failure cases on model performance and also explore other motion synthesis methods based on diffusion-based models [25, 29, 34] regarding potential biases associated between text prompts and generated motions. This includes exploring motion style transfer [1] to apply different motion styles to the generated motion sequences. Also, we will also study the effect of using prompt weighting (often used in text-to-image generation [20]), giving more weights to the activity-related parts of the prompt, which allows the motion synthesis model to focus more on the activity.

Our motivation to use ChatGPT was to generate diverse textual descriptions of activities and eliminate the manual effort needed for prompt engineering. Yet, further study is needed to understand how much the ChatGPT-generated text descriptions help compared to manually generated prompts for generating diverse movements within each activity category. Moreover, there exists multiple LLMs other than Chat-GPT. We used ChatGPT in this study for its user-friendly API encouraging practitioners to use the proposed system. But, we plan to explore other state-of-the-art LLMs [6, 26] to understand their capability of generating realistic, yet variable, text prompts related to activity keywords provided.

Finally, our studies were mainly conducted for locomotion activities. Although detecting locomotion is important due to its relevance to individuals' health, daily activities also involve sporadic and complex motions, such as washing dishes, bed making, etc [13, 15], and even rare activities, for example in wet labs [21]. Thus, it still leaves a question for ChatGPT and T2M-GPT for its capability to generate reasonable virtual IMU data for such complex or sporadic activities.

6 CONCLUSION

We have introduced a method that uses ChatGPT to generate virtual textual descriptions, which are subsequently used to generate 3D human motion sequence and later streams of virtual IMU data. We have demonstrated the effectiveness of our approach to generate virtual IMU data through HAR experiments on three benchmark datasets: RealWorld, PAMAP2, and USC-HAD. Virtual IMU data generated through our approach can be used for significantly improving the recognition performance of HAR models, thereby not requiring any additional manual effort.

REFERENCES

[1] Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. 2020. Unpaired Motion Style Transfer from Video to Animation. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 64.

[2] Matthias Bächlin, Meir Plotnik, Daniel Roggen, Nir Giladi, Jeffrey M Hausdorff, and Gerhard Tröster. 2010. Wearable assistant for Parkinson's disease patients with the freezing of gait symptom. *IEEE Transactions on Information Technology in Biomedicine* 14, 2 (2010), 436–446. https://doi.org/10.1109/TITB.2009.2036165

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 1877–1901.

[4] Ricardo Chavarriaga, Hesam Sagha, Alberto Calatroni, Sundara Tejaswi Digumarti, Gerhard Tröster, José del R Millán, and Daniel Roggen. 2013. The opportunity challenge: a benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters* 34, 15 (2013), 2033–2042. https://doi.org/10.1016/j.patrec.2012.12.014

[5] Wenqiang Chen, Shupei Lin, Elizabeth Thompson, and John Stankovic. 2021. SenseCollect: We Need Efficient Ways to Collect On-body Sensor-based Human Activity Data! *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–27.

[6] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. arXiv:2204.02311 [cs.CL]

[7] W. J. Conover and Ronald L. Iman. 1981. Rank Transformations as a Bridge Between Parametric and Nonparametric Statistics. *The American Statistician* 35, 3 (1981), 124–129.

[8] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022. Generating Diverse and Natural 3D Human Motions From Text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5152–5161.

[9] N. Y. Hammerla, R. Kirkham, P. Andras, and T. Ploetz. 2013. On preserving statistical characteristics of accelerometry data using their empirical cumulative distribution. In *Proceedings of the 2013 international symposium on wearable computers*. 65–68.

[10] Hyeokhyen Kwon, Gregory D Abowd, and Thomas Plötz. 2021. Complex Deep Neural Networks from Large Scale Virtual IMU Data for Effective Human Activity Recognition Using Wearables. *Sensors* 21, 24 (2021), 8337.

[11] Hyeokhyen Kwon, Catherine Tong, Harish Haresamudram, Yan Gao, Gregory D Abowd, Nicholas D Lane, and Thomas Ploetz. 2020. Imutube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–29.

[12] Hyeokhyen Kwon, Bingyao Wang, Gregory D Abowd, and Thomas Plötz. 2021. Approaching the Real-World: Supporting Activity Recognition Training with Virtual IMU Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–32.

[13] Hong Li, Gregory D. Abowd, and Thomas Plötz. 2018. On Specialized Window Lengths and Detector Based Human Activity Recognition. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*. Association for Computing Machinery.

[14] Daniyal Liaqat, Mohamed Abdalla, Pegah Abed-Esfahani, Moshe Gabel, Tatiana Son, Robert Wu, Andrea Gershon, Frank Rudzicz, and Eyal De Lara. 2019. WearBreathing: Real World Respiratory Rate Monitoring Using Smartwatches. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 2, Article 56 (jun 2019), 22 pages. https://doi.org/10.1145/3328927

[15] Sara Mohammed, Reda Elbasiony, and Walid Gomaa. 2018. An LSTM-based Descriptor for Human Activities Recognition using IMU Sensors. 504–511. https://doi.org/10.5220/0006902405040511

[16] Francisco Javier Ordóñez and Daniel Roggen. 2016. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors* (2016).

[17] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, Vol. 35. Curran Associates, Inc., 27730–27744.

[18] 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. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139. PMLR, 8748–8763.

[19] Attila Reiss and Didier Stricker. 2012. Introducing a New Benchmarked Dataset for Activity Monitoring *(ISWC '12)*. IEEE Computer Society. https://doi.org/10.1109/ISWC.2012.13

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]

[21] Philipp M Scholl, Matthias Wille, and Kristof Van Laerhoven. 2015. Wearables in the wet lab: a laboratory system for capturing and guiding experiments. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 589–599.

[22] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in HuggingFace. arXiv:2303.17580 [cs.CL]

[23] Thomas Stiefmeier, Daniel Roggen, Georg Ogris, Paul Lukowicz, and Gerhard Tröster. 2008. Wearable Activity Tracking in Car Manufacturing. *IEEE Pervasive Computing* 7, 2 (2008), 42–50. https://doi.org/10.1109/MPRV.2008.40

[24] Timo Sztyler and Heiner Stuckenschmidt. 2016. On-body localization of wearable devices: An investigation of position-aware activity recognition. In *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. 1–9. https://doi.org/10.1109/PERCOM.2016.7456521

[25] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Amit H Bermano, and Daniel Cohen-Or. 2022. Human Motion Diffusion Model. *arXiv preprint arXiv:2209.14916* (2022).

[26] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL]

[27] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models. arXiv:2303.04671 [cs.CV]

[28] Chengshuo Xia and Yuta Sugiura. 2022. Virtual IMU Data Augmentation by Spring-Joint Model for Motion Exercises Recognition without Using Real Data. In *Proceedings of the 2022 ACM International Symposium on Wearable Computers (ISWC '22)*. Association for Computing Machinery, 79–83. https://doi.org/10.1145/3544794.3558460

[29] Chen Xin, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. 2023. Executing your Commands via Motion Diffusion in Latent Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[30] K. Yamane and Y. Nakamura. 2003. Natural motion animation through constraining and deconstraining at will. *IEEE Transactions on Visualization and Computer Graphics* 9, 3 (2003), 352–360. https://doi.org/10.1109/TVCG.2003.1207443

[31] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. 2023. Decoupling Human and Camera Motion from Videos in the Wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[32] A. D. Young, M. J. Ling, and D. K. Arvind. 2011. IMUSim: A simulation environment for inertial sensing algorithm design and evaluation. In *Proceedings of the 10th ACM/IEEE International Conference on Information Processing in Sensor Networks*. 199–210.

[33] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. 2023. T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[34] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2022. MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model. *arXiv preprint arXiv:2208.15001* (2022).

[35] Mi Zhang and Alexander A. Sawchuk. 2012. USC-HAD: A Daily Activity Dataset for Ubiquitous Activity Recognition Using Wearable Sensors. Association for Computing Machinery.