

IF-D: A HIGH-FREQUENCY, GENERAL-PURPOSE INERTIAL FOUNDATION DATASET FOR SELF-SUPERVISED LEARNING

Patrick Ferreira, Paula Costa

Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (UNICAMP)
Campinas, Brazil

ABSTRACT

We present IF-D, a large-scale inertial dataset designed to enable self-supervised and foundational learning for IMU time series. IF-D comprises continuous, long-duration multichannel recordings (accelerometer, gyroscope, magnetometer) sampled at 200Hz using a UM7 IMU mounted inside a 3D-printed spherical enclosure that promotes diverse, free rotations during vehicle traversal. The collection spans approximately 135 minutes of recording, yielding around 1.6 million samples across nine sensor channels. We describe the data acquisition setup, preprocessing, and calibration procedures (six-orientation accelerometer calibration, stationary gyroscope bias estimation, and ellipsoid fitting for magnetometer hard-/soft-iron correction), and provide quantitative calibration results. IF-D is designed to mitigate platform-specific motion bias and expose models to both physical dynamics and typical measurement noise, thereby facilitating robust representation learning and downstream tasks such as event detection, motion mode recognition, and inertial navigation.

Index Terms— inertial sensing, dataset, self-supervised learning, foundation models, IMU

1. INTRODUCTION

Inertial sensors (accelerometers, gyroscopes, magnetometers) are central to robotics, navigation, wearable sensing and industrial systems due to their capacity to provide motion information independent of external references [1, 2]. However, raw IMU time series pose specific challenges: they typically lack clear periodicity, present high-dimensional coupled channels, and are affected by significant noise and bias that degrade naive integration and linear filtering. These characteristics limit the applicability of classical time-series assumptions (stationarity, separability) and motivate methods that capture subtle temporal and cross-channel dependencies.

Motivated by the training paradigm of large language models—which learn both real-world structure and recurring artifacts from massive, minimally preprocessed corpora—we propose to collect a large, diverse inertial corpus to enable

analogous foundational learning for IMU data. To this end we introduce the Inertial Foundation Dataset (IF-D): long, high-frequency (200Hz) multichannel sequences captured with a UM7 IMU mounted inside a 3D-printed sphere to maximize motion diversity and reduce platform-specific bias. We expect IF-D to support large-scale self-supervised pre-training and the development of robust models that learn both physical patterns and characteristic measurement errors, as suggested by recent foundational-model efforts in other temporal domains [3, 4].

2. RELATED WORK

Transformer-based architectures and large-scale pre-training have advanced time-series forecasting by capturing long-range dependencies. Representative models include Informer, Autoformer and FEDformer for efficient attention and decomposition [5, 6, 7]; PatchTST and SparseTSF for computationally efficient local/ sparse modeling [8, 9]; and foundational/time-scale efforts such as TimesFM and structural approaches like TimesNet [10, 11]. These methods have been validated mainly on series with clear trends or seasonality (weather, traffic, finance), where stable temporal regularities favor attention-based extrapolation.

Conversely, publicly available inertial datasets are specialized and limited in scope (sampling rate, motion regimes or duration), e.g., UCI HAR [12], RoNIN [13], EuRoC MAV [14], TUM VI [15], and KITTI [16]. Each covers a subspace of motions (walking, drone flights, vehicle trajectories) and thus embeds application-specific bias that impedes generalization. This fragmented landscape motivates IF-D, which aims to provide long, high-fidelity, diverse multichannel inertial data suitable for self-supervised and Transformer-based pre-training and for bridging the gap between existing time-series foundations and inertial sensing applications.

3. MATERIALS AND METHODS

The dataset was collected using the **IMU UM7** model from CH Robotics, which features a sampling rate of 200 Hz and

provides three degrees of freedom in the measurement of linear acceleration, angular velocity, and magnetic field. The platform used for data acquisition was the NVIDIA **Jetson Nano**, chosen for its portability and robust embedded hardware.

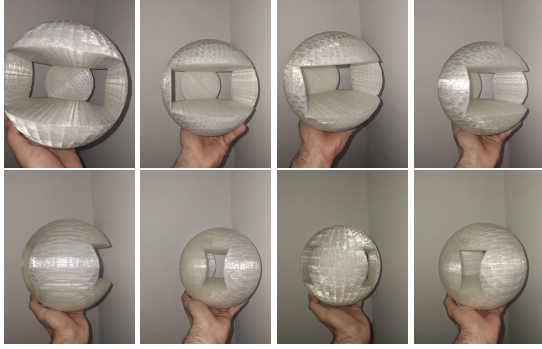


Fig. 1: Multiple views of the 3D-printed spherical model enabling free movement during dataset collection.

Inertial datasets are subject to biases in data collection due to the motion patterns of the vehicle or platform used for acquisition [17]. To maximize the range of possible trajectories, a spherical structure for data collection was devised, 3D-printed, and is shown in Figure 1. The concept behind using this sphere is that sensors placed inside it experience a wide range of accelerations and movements as it moves. By allowing free movement, the aim is to decouple the recorded signals from the typical motion biases of vehicles. The sphere has a diameter of **25 cm** and was carved from a parallelepiped measuring **25 cm x 18 cm x 10 cm**, with its center shifted by **5 cm** relative to the solid sphere, resulting in the model described here.



Fig. 2: The sphere containing the entire collection set used for the dataset, within the vehicular platform. The structure moves freely during the journey.

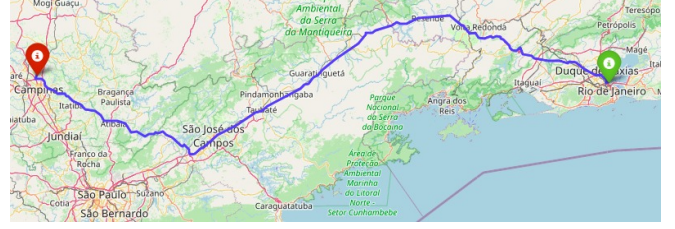


Fig. 3: Trajectory followed for dataset collection. Source: OpenRouteService (HeiGIT) — Map data © OpenStreetMap contributors.

A moving platform is still necessary to set the sphere in motion and record different types of accelerations along its trajectory. For this purpose, a vehicle with an expanded trunk was used (Figure 2), travelling approximately 135 minutes (Figure 3), between the Universidade Estadual de Campinas (UNICAMP, Zeferino Vaz campus) and the Federal University of Rio de Janeiro (UFRJ, Cidade Universitária campus). The final dataset is published as a CSV file with 10 columns: the first column corresponds to timestamps (in seconds), and the following nine columns represent the three axes of the accelerometer (in meters per second squared), gyroscope (in radians per second), and magnetometer (unit-normalized, pointing along the magnetic field direction). This dataset includes over 1.6 million consecutive samples (both raw and calibrated), a volume previously found only in datasets characterized by highly repetitive movement patterns (e.g., pedestrian locomotion) and/or low sampling rates.

3.1. IMU Reference Frames and Calibration

The IMU provides measurements from an accelerometer, a gyroscope and a magnetometer, each expressed in the sensor's body frame. Interpreting these measurements in an inertial (world) frame requires knowledge of the sensor orientation and the removal of systematic sensor imperfections via calibration. The accelerometer measurement model is

$$\mathbf{a}_m = \mathbf{S}_a(\mathbf{a}_{true} + \mathbf{g}) + \mathbf{b}_a + \mathbf{n}_a,$$

where $\mathbf{S}_a$ is the accelerometer scale matrix, $\mathbf{b}_a$ the bias, $\mathbf{n}_a$ measurement noise, and $\mathbf{g}$ the gravity vector expressed in the sensor frame. The gyroscope is modeled as

$$\boldsymbol{\omega}_m = \mathbf{S}_g\boldsymbol{\omega}_{true} + \mathbf{b}_g + \mathbf{n}_g,$$

with $\mathbf{S}_g$, $\mathbf{b}_g$ and $\mathbf{n}_g$ denoting scale, bias and noise respectively. The magnetometer, which is affected by hard-iron (additive) and soft-iron (multiplicative/shape) distortions, is represented by

$$\mathbf{m}_m = \mathbf{S}_s(\mathbf{m}_{true} + \mathbf{b}_h) + \mathbf{n}_m,$$

where $\mathbf{b}_h$ denotes the hard-iron bias and $\mathbf{S}_s$ encodes soft-iron effects.

To correct these systematic effects we applied standard procedures: the six-orientation method for the accelerometer, stationary bias estimation for the gyroscope, and ellipsoid fitting for the magnetometer. In the six-orientation procedure the IMU is placed in six known orthogonal poses to estimate $\mathbf{b}_a$ and $\mathbf{S}_a$ by enforcing

$$\|\mathbf{a}_m - \mathbf{b}_a\| \approx g,$$

with g the known gravitational acceleration. Gyroscope bias $\mathbf{b}_g$ is estimated under static conditions ($\boldsymbol{\omega}_{true} = \mathbf{0}$) as the sample mean:

$$\mathbf{b}_g = \frac{1}{N} \sum_{i=1}^N \boldsymbol{\omega}_m[i].$$

For the magnetometer, we collected measurements across many orientations and performed ellipsoid fitting to estimate $\mathbf{b}_h$ and $\mathbf{S}_s$. The objective is to transform the measured ellipsoid into a unit sphere, i.e.

$$(\mathbf{S}_s(\mathbf{m}_m - \mathbf{b}_h))^\top (\mathbf{S}_s(\mathbf{m}_m - \mathbf{b}_h)) = 1,$$

thereby centering the cloud (hard-iron correction) and correcting its shape (soft-iron correction).

These calibrations remove offsets, scale factors and inter-axis misalignments, enabling consistent interpretation of sensor outputs in a common reference and preparing the data for sensor fusion and machine-learning pipelines. The full calibration sequence and scripts are provided in the dataset repository (section 8).

4. RESULTS

Table 1: Accelerometer calibration parameters

Parameter	X-axis	Y-axis	Z-axis
Bias (m/s^2)	0.26784	-0.21032	0.21283
Scale	1.00522	0.99968	1.01633
Std Bias (m/s^2)	0.05613	0.05261	0.06113

Table 2: Gyroscope calibration parameters

Parameter	X-axis	Y-axis	Z-axis
Bias (rad/s)	0.000978	-0.004603	0.004887
Std Bias (rad/s)	0.001206	0.001148	0.001159

Table 3: Magnetometer calibration parameters (unit-normalized)

Parameter	X-axis	Y-axis	Z-axis
Hard Iron Bias	-0.51625	0.32675	1.17227
Scale	1.00392	1.00564	1.00459

Table 4: Magnetometer Soft Iron matrix

Axis	X	Y	Z
X	0.94163	-0.07392	0.16855
Y	-0.073923	0.88483	0.12147
Z	0.16855	0.12147	1.21098

Table 1 shows the accelerometer calibration parameters. The bias values remain below $0.3 m/s^2$, which is considered adequate for MEMS (*Micro-Electro-Mechanical Systems*) sensors. The scale factors are very close to 1.0 (with variations below 2%), confirming the good linearity of the sensor. The standard deviation of the bias (*Std Bias*) ranges between $0.056 m/s^2$ and $0.062 m/s^2$, within the expected range for this type of device.

In Table 2, corresponding to the gyroscope, the obtained biases are below $0.005 rad/s$, which is approximately $0.25^\circ/s$, consistent with low-cost embedded inertial sensors. The standard deviations are also low (around $0.001 rad/s$), indicating good stability.

Table 3 shows the magnetometer parameters. The hard iron bias values range from -0.52 to $+1.17$, which is acceptable and correctable during calibration, although the Z-axis is more affected. The scale factor is almost unity on all axes, consistent with the range of values provided by the hardware and predicted in its manual. The term *hard iron* refers to magnetic distortions caused by permanent field sources near the sensor, such as magnets or DC currents, which shift the center of the measured ellipsoid.

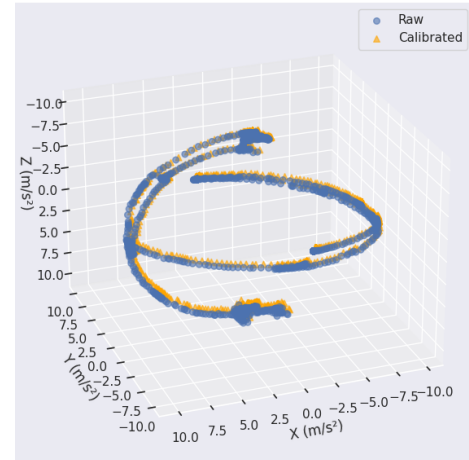


Fig. 4: 3D scatter plot of accelerometer measurements. Raw data (blue circles) and calibrated data (orange triangles) are shown.

Table 4 presents the soft iron correction matrix. The diagonal elements differ from 1.0, ranging approximately from 0.88 to 1.21, indicating moderate distortions. Off-diagonal terms remain relatively small (< 0.2), suggesting limited cou-

pling between axes. The term *soft iron* refers to distortions due to nearby ferromagnetic materials, which deform magnetic field lines, resulting in a tilted ellipsoid relative to the ideal axes.

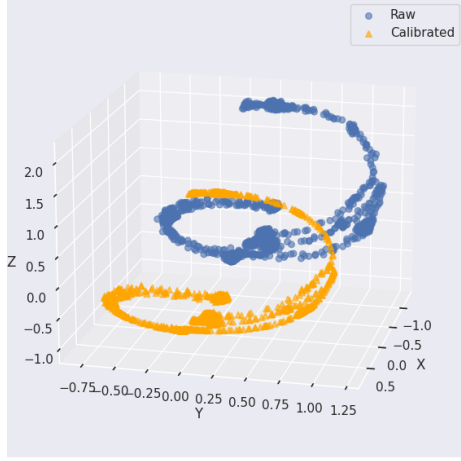


Fig. 5: 3D scatter plot of magnetometer measurements. Raw data (blue circles) and calibrated data (orange triangles) are shown.

The estimated calibration parameters adjust the raw measurements, as can be visually verified. Scatter plots for sensors referenced to an inertial frame (accelerometer and magnetometer) were generated over a 2-minute segment of the dataset, comparing pre- and post-calibration measurements. The accelerometer primarily relies on the inertial frame defined by Earth’s gravitational field, represented as a constant vector pointing toward the center of the planet. Therefore, after calibration, the plot in Figure 4 shows trajectories centered around the origin with a radius of approximately 9.81 m/s . The magnetometer relies on another inertial reference defined by Earth’s magnetic field, pointing toward magnetic north. Prior to calibration, a significant bias is evident (as expected from Table 3 and typical of uncalibrated magnetometers), but after calibration, the data dispersion (Figure 5) shows that measurements are consistently unit-length and centered around the origin. Corrections for scale and ellipsoid adjustment are minor, as expected from the calibration parameters, and visually imperceptible in the plots of Tables 1 and 4.

Overall, the results demonstrate that all three sensors exhibit biases and non-linearities within typical limits for low-cost MEMS-based inertial units. These microfabricated sensors detect acceleration, angular velocity, and magnetic field intensity with small form factor, low power consumption, and affordability, albeit subject to noise and drift. The estimated corrections allow reliable use of the data in sensor fusion and machine learning applications, compatible with the type of signals encountered in such tasks. With the diversity of motions and high sampling rate enabled by the innovative dataset

acquisition, the training of large-scale AI models in the inertial sensor domain is facilitated.

5. CONCLUSION

We presented a novel high-frequency (200Hz) inertial dataset obtained with a UM7 IMU placed inside a 3D-printed spherical structure, yielding long, diverse multichannel sequences with reduced motion bias. Calibration procedures for accelerometer, gyroscope, and magnetometer were applied and shown to place sensor errors within expected ranges for low-cost MEMS devices. Crucially, the use of a low-cost IMU and the natural noise introduced by recording inside a moving vehicle across many motion patterns is intended to enable models to learn and mitigate not only the underlying physical dynamics but also the characteristic measurement errors—analogueous to how large language models learn phenomena and systematic artifacts from massive, unstandardized text corpora. By providing long-duration, high-resolution, and largely unlabeled inertial data, this resource facilitates self-supervised and Transformer-based approaches in the inertial domain. As immediate future work, we plan to develop and train a foundational model for inertial time series on the dataset released here, aiming to demonstrate improved generalization in tasks such as event detection, motion-mode recognition, and inertial navigation.

6. ACKNOWLEDGEMENTS

This work was partially funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 and by the Ministry of Science, Technology, and Innovation of Brazil, with resources granted by the Federal Law 8.248 of October 23, 1991, under the PPI-Softex, DOU 01245.003479/2024-10. The authors are also affiliated with the Artificial Intelligence Lab, Recod.ai.

7. COMPLIANCE WITH ETHICAL STANDARDS STATEMENT

The data collection did not involve human subjects or require ethical approval.

8. DATA AVAILABILITY

The IF-D dataset [18] is available in the link: <https://zenodo.org/records/17148626>.

9. REFERENCES

- [1] Naser El-Sheimy and Ahmed Youssef, “Inertial sensors technologies for navigation applications: state of the art

- and future trends,” *Satellite Navigation*, vol. 1, no. 1, pp. 1–21, dec 2020.
- [2] Yi Gao, Fei Li, and Jia Chen, “Random Weighting Adaptive Estimation of Model Errors on Attitude Measurement for Rotary Steerable System,” *IEEE Access*, vol. 10, pp. 80794–80803, 2022.
 - [3] Shibo Zhao, Sifan Zhou, Raphael Blanchard, Yuheng Qiu, Wenshan Wang, and Sebastian Scherer, “Tartan IMU: A light foundation model for inertial positioning in robotics,” in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2025.
 - [4] Christoph Wieland and Victor Pankratius, “Inertial signal forecasting with foundation model techniques,” *IEEE Sensors Journal*, 2025.
 - [5] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang, “Informer: Beyond efficient transformer for long sequence time-series forecasting,” *May 2021*, vol. 35, pp. 11106–11115.
 - [6] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long, “Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting,” in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, Eds. 2021, vol. 34, pp. 22419–22430, Curran Associates, Inc.
 - [7] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin, “FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting,” in *International Conference on Machine Learning (ICML)*, 2022, pp. 27268–27286.
 - [8] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam, “A time series is worth 64 words: Long-term forecasting with transformers,” in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2023.
 - [9] Shengsheng Lin, Weiwei Lin, Wentai Wu, Haojun Chen, and Junjie Yang, “Sparsesf: modeling long-term time series forecasting with 1k parameters,” in *Proceedings of the 41st International Conference on Machine Learning*, 2024, ICML’24, JMLR.org.
 - [10] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou, “A decoder-only foundation model for time-series forecasting,” 2024.
 - [11] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long, “TimesNet: Temporal 2d-variation modeling for general time series analysis,” in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2023.
 - [12] Jorge Reyes-Ortiz, Davide Anguita, Alessandro Ghio, Luca Oneto, and Xavier Parra, “Human activity recognition using smartphones,” *UCI Machine Learning Repository*, 2013, DOI: 10.24432/C54S4K.
 - [13] Sachini Herath, Hang Yan, and Yasutaka Furukawa, “RoNIN: Robust Neural Inertial Navigation in the Wild: Benchmark, Evaluations, New Methods,” *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 3146–3152, may 2020.
 - [14] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W Achtelik, and Roland Siegwart, “The euroc micro aerial vehicle datasets,” *Int. J. Rob. Res.*, vol. 35, no. 10, pp. 1157–1163, Sept. 2016.
 - [15] David Schubert, Thore Goll, Nikolaus Demmel, Vladyslav Usenko, Jörg Stückler, and Daniel Cremers, “The tum vi benchmark for evaluating visual-inertial odometry,” in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 1680–1687.
 - [16] A Geiger, P Lenz, C Stiller, and R Urtasun, “Vision meets robotics: The kitti dataset,” *Int. J. Rob. Res.*, vol. 32, no. 11, pp. 1231–1237, Sept. 2013.
 - [17] Artur Shurin, Alex Saraev, Mor Yona, Yevgeni Gutnik, Sharon Faber, Aviad Etzion, and Itzik Klein, “The autonomous platforms inertial dataset,” *IEEE Access*, vol. 10, pp. 10191–10201, 2022.
 - [18] Patrick Ferreira and Paula Costa, “If-d: A high-frequency, general-purpose inertial foundation dataset for self-supervised learning,” Sept. 2025.