SCIENCE CHINA
Information Sciences

• NEWS & VIEWS •

September 2025, Vol. 68, Iss. 9, 197301:1–197301:2
https://doi.org/10.1007/s11432-024-4445-0

BUPTCMCC-6G-DataAI+: a generative channel dataset for 6G AI air-interface research

Li YU[1], Jianhua ZHANG[1*], Shuangfeng HAN[2] & Tao JIANG[2]

[1]*State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[2]*China Mobile Research Institute, Beijing 100032, China*

Received 20 December 2024/Revised 18 April 2025/Accepted 14 May 2025/Published online 22 July 2025

Citation Yu L, Zhang J H, Han S F, et al. BUPTCMCC-6G-DataAI+: a generative channel dataset for 6G AI air-interface research. Sci China Inf Sci, 2025, 68(9): 197301, https://doi.org/10.1007/s11432-024-4445-0

On September 24, 2024, Beijing University of Posts and Telecommunications and China Mobile Communications Corporation jointly released the BUPTCMCC-6G-DataAI+ dataset, which is the first proposed generative channel dataset for 6G artificial intelligence (AI) air-interface research. It is the upgraded version of BUPTCMCC-6G-DataAI, which was already published in June 2023, aiming at extending 6G new features and scenarios. BUPTCMCC-6G-DataAI+ is constructed with the basic channel and corresponding propagation environment dataset and a generative post-processing module to supply sufficient channel data for diverse scenarios. The accurate ray-tracing data function as the basic channel covering the spectrum from sub-6 GHz and new mid-bands to THz and supporting the potential 6G features, such as near-field effects, environment sensing, reconfigurable intelligent surfaces (RIS), and industrial internet of things (IIoT). Then, the generative post-processing module is designed to tailor the ray-tracing channel data to satisfy the requirements of user tasks. Configured with customized parameters, BUPTCMCC-6G-DataAI+ can adaptively generate scalable large-scale and small-scale channel parameters to train and test developed AI models, set benchmarks, and evaluate the algorithm performance.

6G is envisaged to supply various services as the next-generation communication infrastructure around 2030 [1]. Integrated AI and communication is defined by the International Telecommunication Union (ITU) as a typical new usage scenario for 6G [2]. Empowered by AI technologies, the AI radio access network (AI-RAN) is anticipated to effectively help reduce costs and enable smarter decision-making for network automation, such as in channel estimation, signal detection, channel state information (CSI) prediction and feedback, beam management, resource allocation, and network optimization [3]. To train and evaluate the AI algorithms for diverse RAN technologies, it is essential to have a sufficiently large dataset for researchers with the following key features.

- Support 6G new features. With the global development of 6G research, new technologies and features, higher frequency, and diverse scenarios have been proposed, such as integrated sensing and communication (ISAC), extremely large-scale multiple-input and multiple-output (XL-MIMO), near-field effects, RIS, and space-air-ground integrated network (SAGIN).
- Support specific environment. Most AI-RAN algorithms, such as beam and blockage prediction, rely on the channel characteristics of spatial consistency and time coherence, which can be derived from the environment features like scattering geometry or distribution of a specific environment setup. Some AI models directly take the environment features as essential training input; therefore, both the channel data and environment information are indispensable.
- Generative and configurable by user. Training AI models for different communication tasks requires diverse dataset configurations, including scenario type, antenna number and pattern, frequency, bandwidth, special features with RIS or ISAC, scatterer mobility, and receiver speed. Thus, there is an urgent need for a generative dataset that can be configured and adjusted by users to generate channel data according to the customized requirements of task-oriented AI.

Several datasets have recently been constructed for open access. The wireless-intelligence is a measurement-based channel dataset for AI tasks obtained from outdoor practical 5G networks[1)]. DeepMIMO is a ray-tracing based channel dataset, considering some indoor and outdoor scenarios, which is mainly targeted for MIMO research [4]. The wireless AI research dataset is also based on ray-tracing data for a large number of AI-generated environments [5]. However, some essential propagation mechanisms, e.g., diffraction, are not included in some of the existing datasets, whose features are vital for algorithm design, such as long-range blockage prediction [6]. In addition, 6G new features, such as the near-field spherical wave feature of XL-MIMO, RIS, and high-mobility scenarios, are not considered in the existing datasets. On June 4, 2023, BUPTCMCC-6G-DataAI was released to support the constantly evolving 6G features and

* Corresponding author (email: jhzhang@bupt.edu.cn)
1) https://wireless-intelligence.com/#/download.

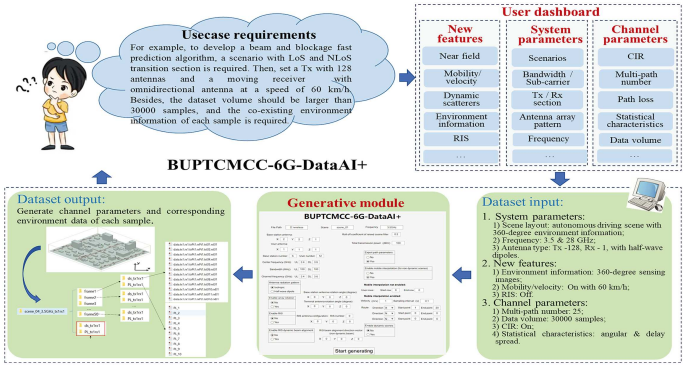

Figure 1 (Color online) Architecture of BUPTCMCC-6G-DataAI+.

scenarios and the new requirements brought by the users [7]. Since its release, BUPTCMCC-6G-DataAI has been downloaded over 1700 times by more than 90 domestic and international enterprises and universities, which shows urgent demand for the channel dataset.

BUPTCMCC-6G-DataAI+ is the upgraded version, which incorporates a generative module to transfer the requirements of user tasks into specific dataset configurations and further extends 6G new features by adding the scenarios and features of dynamic environment sensing, XL-MIMO, RIS, IIoT and new mid-bands. Specifically, BUPTCMCC-6G-DataAI+ has the following new features.

• Functions. As illustrated in Figure 1, BUPTCMCC-6G-DataAI+ is designed with a generative module to post-process and tailor the channel data according to the parameterized system/channel configurations set by the users. For example, a dataset user is considering designing a beam and blockage prediction algorithm. The demands for this task are as follows: the scenario should consist of a line of sight (LoS) and non-LoS (NLoS) transition section; MIMO antennas are needed to switch beams; the receiver (Rx) should be moving across the LoS/NLoS section with a certain speed, and the total samples should be large enough for the AI model training. Then, the aforementioned requirements are translated into specific configurations. As the user dashboard demonstrates, BUPTCMCC-6G-DataAI+ can provide a variety of configuration options, including scenarios, base stations, Rx section and trajectory, frequencies, mobile interpolation, velocity, and other possible features of RIS and dynamic scatterer settings. With the customized configurations, BUPTCMCC-6G-DataAI+ can post-process the basic channel data of ray-tracing multipath, to calculate and generate the required large-scale and small-scale channel parameters with specific features and volume.

• Scenarios. BUPTCMCC-6G-DataAI+ can provide multiple scenarios, including outdoor street and indoor office, and particularly extends 6G scenarios, such as outdoor sensing, indoor RIS, and IIoT. As a 6G new scenario, an outdoor street scenario is built for ISAC to obtain co-existing panoramic environment and channel data. An indoor RIS scenario consists of corridor corners and compartments with the walls between the Tx and Rx, and the RIS is set at the connection section to provide Tx-RIS-Rx cascade channels. Moreover, the beam alignment scheme can also be customized, which allows users to set the reflection beam direction of RIS as a specific coordinate or a moving trajectory. The IIoT scenario includes a large number of metal scatterers with various sizes representing the metallic machines, as well as two moving automated guided vehicles (AGVs), whose mobility can also be configured by the users.

• Applications. As a generative dataset for task-oriented AI, BUPTCMCC-6G-DataAI+ supports flexible section selection of the scenario layout and sets customized configurations for system, feature, and channel levels, according to the task requirements, including beam management in the physical layer, power allocation in the resource layer, and network planning and optimization in the network function layer. The accuracy of the generated channel data is validated through some AI task applications, such as CSI prediction [3] and beam prediction [7].

Conclusion. This study has presented the generative channel dataset BUPTCMCC-6G-DataAI+, which provides diversified channel data supporting the design, optimization, benchmark setting, and evaluation of 6G AI air-interface algorithms. The generative framework, new features, and scenarios of the dataset have been discussed. With the development of 6G, BUPTCMCC-6G-DataAI+ will be constantly optimized by adding more scenarios, features, and applications to support the ubiquitous AI for 6G. The dataset is available on the high-performance computing platform of Beijing University of Posts and Telecommunications[2)].

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 62401084), National Key R&D Program of China (Grant No. 2023YFB2904805), and Beijing University of Posts and Telecommunications-China Mobile Research Institute Joint Innovation Center.

References

1 Zhang J H, Lin J X, Tang P, et al. Channel measurement, modeling, and simulation for 6G: a survey and tutorial. 2023. ArXiv:2305.16616
2 Wang Z Q, Du Y, Wei K J, et al. Vision, application scenarios, and key technology trends for 6G mobile communications. Sci China Inf Sci, 2022, 65: 151301
3 Shi L Z, Zhang J H, Yu L, et al. Can wireless environmental information decrease pilot overhead: a CSI prediction example. 2024. ArXiv:2408.06558
4 Alkhateeb A. DeepMIMO: a generic deep learning dataset for millimeter wave and massive MIMO applications. In: Proceedings of Information Theory and Applications Workshop, 2019. 1–8
5 Huangfu Y R, Wang J, Dai S C, et al. WAIR-D: wireless AI research dataset. 2022. ArXiv:2212.02159
6 Yu L, Zhang J, Zhang Y, et al. Long-range blockage prediction based on diffraction fringe characteristics for mmWave communications. IEEE Commun Lett, 2022, 26: 1683–1687
7 Shen Z B, Yu L, Zhang Y X, et al. DataAI-6G: a system parameters configurable channel dataset for AI-6G research. In: Proceedings of IEEE Global Communications Conference Workshops, 2023. 1910–1915

2) https://hpc.bupt.edu.cn/dataset-public/datasets/.