CONVERGE Challenge: Multimodal Learning for 6G Wireless Communications

ICASSP 2026 SP Grand Challenge

General Guidelines for the Challenge

v1.0 (22/12/2025)

Jichao Chen
Communication Systems Department, EURECOM, France
jichao.chen@eurecom.fr

Filipe B. Teixeira
INESC TEC, Faculdade de Engenharia, Universidade do Porto, Portugal
filipe.b.teixeira@inesctec.pt

Francisco M. Ribeiro
INESC TEC, Faculdade de Engenharia, Universidade do Porto, Portugal
francisco.m.ribeiro@inesctec.pt

Ahmed Alkhateeb
Arizona State University, Tempe, USA
alkhateeb@asu.edu

Luis M. Pessoa
INESC TEC, Faculdade de Engenharia, Universidade do Porto, Portugal
luis.m.pessoa@inesctec.pt

Dirk Slock
Communication Systems Department, EURECOM, France
dirk.slock@eurecom.fr

December 22, 2025

Abstract

High-frequency mmWave communication enables ultra-high data rates and low latency but faces considerable challenges due to severe path loss, especially in non-line-of-sight (NLoS) scenarios. Augmenting radios with visual sensing has recently proven effective, as cameras provide rich environmental context that helps predict obstructions and guide proactive network actions. In this CONVERGE Challenge, we invite participants to develop machine learning models that integrate visual and radio data to address key communication tasks in high-frequency wireless systems.

The challenge consists of three independent tracks: 1) blockage prediction, 2) UE localization and position prediction, and 3) channel prediction. They are based on a rich, real-world multimodal dataset collected in a controlled indoor mmWave testbed. This challenge offers an opportunity to benchmark cross-modal learning approaches and promotes interdisciplinary collaboration among the wireless communications, signal processing, computer vision, and AI communities.

Contents

List of Figures

List of Tables

Chapter 1

Call for Participation

With the rapid evolution of wireless communication, the operational frequencies continue to increase, now extending into the millimeter-wave (mmWave) and sub-terahertz (sub-THz) bands. This progression unlocks significant potential, offering advantages such as increased bandwidth, substantially higher data rates, and remarkably reduced latency, thereby supporting emerging applications including augmented reality (AR), virtual reality (VR), ultra-high-definition streaming, and ultra-reliable low-latency communications [1]. These advancements promise significant improvements in user experience and enable new, bandwidth-intensive, and latency-sensitive services critical for future communication systems.

Despite these benefits, higher frequency communications face considerable challenges due to signal sensitivity, particularly in line-of-sight (LoS) and non-line-of-sight (NLoS) conditions. At mmWave and sub-THz frequencies, signals experience severe attenuation and are highly susceptible to blockage from common environmental objects and human bodies, leading to frequent disruptions [2]. Moreover, beamforming techniques, essential for maintaining high antenna gains and increasing data rates, introduce additional complexities.

The narrow beams required for high gains demand precise alignment between the transmitter and the receiver, which can be hard to maintain in dynamic environments, increasing the overhead associated with beam training and adaptation [3]. To overcome these limitations, integrating additional sensing modalities, such as visual sensors like cameras, has emerged as a promising approach [3][4]. By capturing detailed visual information, the system gains a richer environmental context, which significantly improves its capability to discern spatial relationships and anticipate obstructions in the signal path.

Through advanced computer vision and machine learning algorithms, future blockage events and UE position can be effectively predicted, enabling proactive network responses such as beam switching, handovers, or adaptive

resource allocation. The obtained environmental context is also valuable for channel prediction, reducing the overhead associated with channel estimation.

In this challenge, participants are invited to develop innovative machine learning solutions using visual and radio data to tackle the unique demands of fast-changing, high-frequency communication environments. Participants will work on one or more tasks using visual data captured from cameras and radio-frequency data collected by wireless nodes:

- **Blockage Prediction:** Accurately forecast future blockage conditions to enable proactive network adjustments.
- **User Equipment (UE) Localization and Position Prediction:** Determine and predict UE positions to support wireless communication.
- **Channel Prediction:** Predict the Sounding Reference Signal (SRS) to reduce the channel estimation overhead.

This challenge is grounded in and supported by the experimental infrastructure developed in the CONVERGE project [5], which provides a unique platform for multimodal experimentation. The CONVERGE chamber integrates a mobile FR2-capable gNB and UE, a programmable obstacle, stereo RGB-D cameras, Reconfigurable Intelligent Surfaces (RIS), and a programmable control architecture. This environment enables the generation and collection of synchronized radio and visual datasets under realistic, controlled indoor conditions with configurable occlusions and mobility patterns, which reflects the complexity of near-field beam management in dynamic environments.

We warmly invite researchers and industry practitioners to participate in this CONVERGE Challenge: Multimodal Learning for 6G Wireless Communications. With such a diverse and realistic dataset, as well as a wide range of well-designed tasks to choose from, we believe this challenge will attract broad interest and participation from a wider community. We hope that through this challenge, participants will gain a deeper understanding of how visual data integration can enhance and revolutionize wireless communications, driving forward research and innovation in this emerging interdisciplinary field.

The top five teams will be invited to submit a 2-page paper describing their approaches, which will be presented at ICASSP 2026, and accepted papers will be published in the ICASSP proceedings. In addition, teams presenting in person at ICASSP will be encouraged to submit a full-length paper to the IEEE Open Journal of Signal Processing (OJ-SP).

NOTE: All intellectual property (IP) rights for the data and baseline code provided in this challenge remain with the organizers. Participants

retain ownership of any methods or models they develop. By submitting results, participants grant the organizers to use the submitted solutions for evaluation and publication purposes related to the challenge. Participants are responsible for ensuring that their submissions do not infringe upon the rights of third parties.

Chapter 2

Challenge Description

The CONVERGE Challenge: Multimodal Learning for 6G Wireless Communications challenge aims to leverage both visual sensor data and wireless signal data to develop ML solutions for supporting critical communication tasks, such as proactive blockage prediction and beam prediction. Designing an effective ML model for these multimodal communication tasks demands a huge amount of real-world data. To support the development and evaluation of the ML models, we will build a comprehensive visual-radio dataset within a controlled chamber environment, providing synchronized image and radio signal data, coupled with clearly labeled ground truth information for the corresponding tasks.

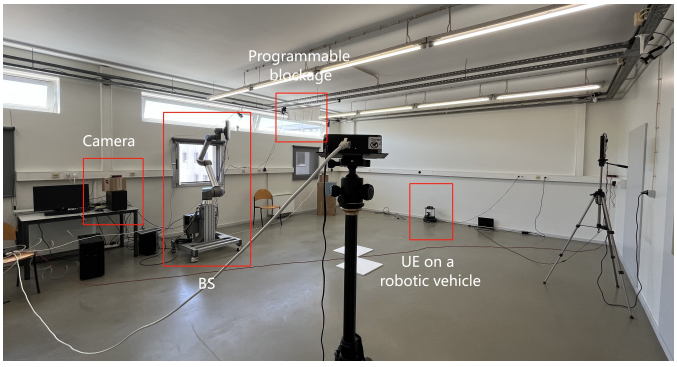

Figure 2.1: CONVERGE chamber with mobile BS, UE, and camera in Porto, Portugal.

In our environment, as illustrated in Fig. 2.1, we deploy a mobile Base Station (BS) as a transmitter with mmWave Radio Unit that points to the receiver. The BS is implemented with a LiteOn FR2 Radio Unit as the mmWave front-end, integrated with the OpenAirInterface (OAI) software stack to implement the 5G NR gNB and mounted on a Universal Robotics UR-10e. The receiver comprises a RM530F user equipment (UE) from

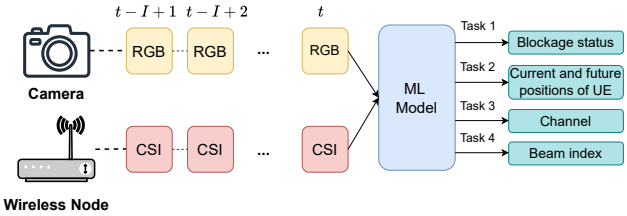

Figure 2.2: Overview of input data sequence utilized for three tasks.

Quectel, mounted on a Turtlebot 4 robotic vehicle, fitted with mmWave phased arrays oriented in multiple directions. Additionally, the setup includes a custom-made programmable RF shield curtain capable of dynamically blocking the communication path between the transmitter and receiver. To gather visual context, a Precision Time Protocol (PTP) synchronized Nerian Ruby 3D RGB-D camera will be positioned at the BS to capture the surrounding environment and UE interactions.

The challenge will focus on three primary tasks:

- **Blockage Prediction:** Predict future blockage events, enabling proactive adjustments in communication strategies.
- **UE Localization and Position Prediction:** Determine the current position of the UE and predict its future position. The UE position information is crucial for communication tasks such as beam management and channel estimation.
- **Channel Prediction:** Using past SRS that includes only pilot symbols plus synchronized images, participants must infer the complete SRS matrix for the next time slot, thereby cutting channel-estimation overhead.

The three tracks are independent and scored separately, so teams may enter any subset, i.e., from one task to all three. Furthermore, we encourage exploiting cross-task synergies, for example, using blockage or UE-position cues to boost channel or beam predictions.

Chapter 3

Dataset Description

3.1 Overview

The CONVERGE Challenge dataset is a real-world, multimodal visual–radio dataset collected in a controlled indoor mmWave/FR2 testbed, referred to as the *CONVERGE chamber*. The dataset is designed to support learning-based solutions for three independent but complementary tasks: *blockage prediction*, *UE localization and position prediction*, and *channel prediction*.

All sensing modalities are precisely time-synchronized using Precision Time Protocol (PTP), enabling accurate cross-modal learning between visual observations and wireless measurements. The dataset captures realistic dynamics such as mobility, blockage events, beam misalignment, and channel variations, while maintaining controlled experimental conditions to ensure reproducibility.

3.2 Experimental Environment

Data collection is performed using a mobile FR2-capable gNB mounted on a robotic arm and a UE mounted either on a static tripod or on a mobile Turtlebot platform, depending on the task scenario. A programmable RF-shield curtain is used to induce controlled blockage and non-line-of-sight (NLoS) conditions.

Visual context is captured by a PTP-synchronized Nerian Ruby RGB-D camera mounted on top of the gNB and oriented towards the UE. For Task 2, a Qualisys motion-capture system provides sub-millimeter-accurate ground-truth UE positioning.

Multiple experimental scenarios are conducted, covering static and mobile UE configurations, aligned and non-aligned beams, and diverse blockage transitions, including no blockage, partial blockage, and full blockage.

3.3 Modalities

Each dataset sample consists of synchronized visual data, radio measurements, and task-specific ground-truth annotations.

Visual Data

- RGB images captured at 7 frames per second (one frame every 142 ms).
- Corresponding disparity frames enabling depth estimation and 3D reconstruction.
- Camera calibration parameters are provided to allow metric-scale visual processing.

Radio Data Two categories of radio measurements are provided:

- **E2 interface measurements** collected every 50 ms, including PHY/MAC indicators such as SINR, RSRP, RSSI, CQI, MCS, BLER, rank indicator, and throughput counters.
- **Sounding Reference Signal (SRS) measurements** captured every 10 ms for each antenna. Channel estimates are obtained via least-squares estimation followed by interpolation across subcarriers. Each SRS snapshot contains complex-valued CSI across all beams and subcarriers.

3.4 Dataset Organization

The dataset follows a unified directory structure across all tasks:

```
dataset/
 |-- calibration/
 |   `-- nerian_gnb_1_calib.yaml
 |-- task1/exp1..exp5/
 |   |-- annotations/
 |   |   `-- exp*.csv
 |   |-- radio/
 |   |   |-- E2-*.csv
 |   |   `-- SRS-*.json
 |   `-- video/
 |       |-- frames.csv
 |       |-- color/
 |       `-- disparity/
 |-- task2/exp6..exp8/
 |   |-- annotations/
```

```
|   |   `-- exp*_translation.csv
|   |-- radio/
|   |   |-- E2-*.csv
|   |   `-- SRS-*.json
|   `-- video/
|       |-- frames.csv
|       |-- color/
|       `-- disparity/
`-- index.csv
```

A global `index.csv` file provides one entry per scenario, specifying the task type, scenario identifier, paths to visual data, radio measurements, annotation files, and a short textual description.

3.5 Temporal Structure

Each learning sample consists of a temporal context window of I consecutive timestamps, spanning from $t - I + 1$ to t. This design enables models to exploit motion cues, temporal radio dynamics, and evolving blockage conditions.

Timestamps across video frames, radio measurements, and annotations are expressed in seconds and aligned to a common absolute time reference.

Ground-Truth Annotations

Task-specific ground-truth labels are provided as follows:

- **Task 1 – Blockage Prediction**: Frame-level blockage annotations with three states (*no*, *partial*, *full*), indicating UE visibility and expected signal degradation. The prediction target is a future blockage indicator within a predefined horizon.
- **Task 2 – UE Localization and Position Prediction**: Ground-truth UE positions expressed as 3D translations (x, y, z) in millimeters relative to the gNB coordinate frame, obtained from motion-capture measurements. Labels include current and future UE positions.
- **Task 3 – Channel Prediction**: Ground-truth CSI matrices for the next time step, including all beams and OFDM symbols, represented as complex-valued tensors. Additional labels include per-beam received power and the optimal beam index.

Training and Evaluation Splits

Participants are provided with a fully labeled training dataset. Final evaluation is conducted on a held-out test set, for which inputs and labels remain

private to the organizers. The same data format and temporal structure are preserved across training and testing to ensure fair and reproducible benchmarking.

Summary

By combining synchronized RGB-D imagery, high-resolution channel measurements, and precise ground-truth annotations, the CONVERGE dataset enables rigorous evaluation of multimodal learning techniques for future 6G systems. Its unified structure across tasks facilitates cross-task knowledge transfer and encourages the development of holistic vision–radio models.

Chapter 4

Evaluation Criteria

The primary evaluation metrics for each of the three tasks are summarized in Table 4.1. Ranking of submissions is based on a composite score that combines each task's primary metric with measured inference latency.

Table 4.1: Primary evaluation metrics for each task.

Task	**Metric**	**Definition**	**Justification**
Blockage Prediction	**F1-Score**	Harmonic mean of precision and recall	Balances false alarms and missed detections, well-suited for imbalanced blockage / no-blockage classes
UE Localization and Position Prediction	**RMSE [m]**	Square root of the mean squared distance between predicted and true UE positions	Standard in positioning literature, directly reflects the spatial error in meters
Channel Prediction	**NMSE [dB]**	Mean squared error between predicted and true complex CSI values, normalized by the power of the true CSI	Standard in channel estimation, enables fair comparison across power levels

All submissions will be containerized and executed under the same standardized inference environment (identical hardware and software stack). After 100 warm-up inferences, we record per-sample latency over the full test set and report mean latency L (ms).

Let P be the task's primary performance metric. To bring P and L onto

the same $[0, 1]$ scale, we normalize each as

$$\hat{P} = \begin{cases} \dfrac{P - P_{\min}}{P_{\max} - P_{\min}}, & \text{if } P \text{ is “higher-is-better” (F1, APL),} \\ \dfrac{P_{\max} - P}{P_{\max} - P_{\min}}, & \text{if } P \text{ is “lower-is-better” (RMSE, NMSE),} \end{cases} \qquad \hat{L} = \frac{L_{\max} - L}{L_{\max} - L_{\min}}.$$

We then compute a single score

$$S = \alpha\, \hat{P} \; + \; (1 - \alpha)\, \hat{L}, \quad \alpha = 0.7.$$

Submissions are ranked by descending S.

Chapter 5

Guidelines for Participants

Teams must register on the CONVERGE website before the registration deadline. The registration must state the team name, each member's full name and affiliation, and a contact e-mail address. Each person may belong to *only one* team, but teams may enter any subset of the three tasks and can submit results for one, several, or all of them.

After evaluation, the winners of the three individual tasks will be invited to submit a two-page paper for presentation at ICASSP 2026. A fifth invitation is reserved for a "wildcard" entry that blends strong performance with breadth. For every team that tackles at least *two* tasks, we compute

$$\text{WildcardScore} = \underbrace{\frac{1}{T}\sum_{i=1}^{T}\pi_i}_{\text{mean percentile}} + \underbrace{\frac{T-1}{3}}_{\text{breadth bonus}},$$

where T is the number of tasks entered and π_i the team's percentile rank in task i. The highest WildcardScore earns the wildcard invitation.

By the submission deadline, teams must provide a self-contained inference package—such as a Docker image or Python module—with the trained model, all inference code, and a requirements file. After the challenge, participants are encouraged to open-source their code to promote reproducibility. The inference interface must adhere to the organizers' specification so it can automatically load the held-out inputs and produce outputs in the correct format.

The prediction outputs should be bundled in a CSV file with the following columns (and additional files for CSI predictions):

- `id`: The absolute ID of the sample in the testing dataset.
- `y_blockage`: The predicted future blockage status if the team participates in task 1.

- `y_ue_position_current_and_future`: The 2D positions of the UE at timestamps t. It contains two columns: `y_ue_position_current_p1`, `y_ue_position_current_p2`- This field applies to task 2 participants.

- `y_srs`: File name of the predicted SRS data if the team participates in task 3. In this file, the SRS should have the same format as the ground truth CSI data in the training dataset.

Chapter 6

Challenge Timeline

- Competition Launch and Data Release: December 22, 2025
- Registration Deadline: October 31, 2025
- Submission Deadline: January 16, 2026
- Results and Rankings Notification: January 23, 2026
- 2-page Papers Due (by invitation): January 31, 2026
- 2-page Paper Acceptance Notification: February 7, 2026
- Camera-ready Submission: February 14, 2026 (FIRM)

Chapter 7

Potential Participants

This challenge is expected to attract researchers, engineers, and students from academia and industry with expertise in wireless communications, signal processing, computer vision, and machine learning. In particular, the challenge will appeal to research groups working on mmWave and FR2 beamforming, channel modeling, and mobility-aware communication; computer-vision and machine-learning specialists focused on multimodal perception, sensor fusion, and real-time inference; teams exploring joint communication-and-sensing and digital-twin system modeling; the broader 6G community, including contributors from CONVERGE, 6G-XR, 6G-SANDBOX, and SLICES-RI, and finally, early-stage researchers and graduate students seeking hands-on experience with realistic vision-RF datasets.

By providing high-quality, labeled multimodal datasets collected in a real-world FR2 testbed (the CONVERGE chamber), the challenge lowers the barrier for participation while offering an opportunity to benchmark and publish competitive solutions in a cutting-edge domain.

Bibliography

[1] C.-X. Wang, X. You, X. Gao, X. Zhu, Z. Li, C. Zhang, H. Wang, Y. Huang, Y. Chen, H. Haas *et al.*, "On the road to 6g: Visions, requirements, key technologies, and testbeds," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 2, pp. 905–974, 2023.

[2] S. Wu, M. Alrabeiah, C. Chakrabarti, and A. Alkhateeb, "Blockage prediction using wireless signatures: Deep learning enables real-world demonstration," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 776–796, 2022.

[3] A. Alkhateeb, G. Charan, T. Osman, A. Hredzak, J. Morais, U. Demirhan, and N. Srinivas, "Deepsense 6g: A large-scale real-world multi-modal sensing and communication dataset," *IEEE Communications Magazine*, vol. 61, no. 9, pp. 122–128, 2023.

[4] T. Nishio, Y. Koda, J. Park, M. Bennis, and K. Doppler, "When wireless communications meet computer vision in beyond 5g," *IEEE Communications Standards Magazine*, vol. 5, no. 2, pp. 76–83, 2021.

[5] "CONVERGE Project: View-to-communicate and communicate-to-view," https://www.converge-project.eu, accessed: 2025-07-09.