CONVERGE CHALLENGE: MULTIMODAL LEARNING FOR 6G WIRELESS COMMUNICATIONS

Jichao Chen[1], *Filipe B. Teixeira*[2], *Francisco M. Ribeiro*[2], *Ahmed Alkhateeb*[3], *Manuel Ricardo*[2], *Luis M. Pessoa*[2], *Dirk Slock*[1]

[1] Communication Systems Department, EURECOM, France
[2] INESC TEC, Faculdade de Engenharia, Universidade do Porto, Porto, Portugal
[3] School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, USA

{jichao.chen, dirk.slock}@eurecom.fr, alkhateeb@asu.edu,
{filipe.b.teixeira, francisco.m.ribeiro, manuel.ricardo, luis.m.pessoa}@inesctec.pt

ABSTRACT

High-frequency mmWave and sub-THz systems enable ultra-high data rates but suffer from severe path loss and blockage sensitivity. Visual sensing can enhance reliability by providing environmental awareness for proactive beam management, yet progress has been limited by the lack of synchronized real-world multimodal datasets. This CONVERGE challenge addresses this gap with a novel indoor mmWave dataset and three tasks: Blockage Prediction, UE Localization, and Channel Prediction. These tasks are designed to benchmark cross-modal learning and promote collaboration between the wireless and computer vision communities.

***Index Terms*—** 6G, multimodal learning, mmWave, computer vision, wireless communications, ISAC

1. INTRODUCTION

The evolution toward 6G networks relies on exploiting millimeter-wave (mmWave) and sub-terahertz (sub-THz) bands to support high-data-rate and low-latency services such as extended reality (XR) [1]. However, these high-frequency signals are highly vulnerable to blockage and severe path loss [2], requiring narrow beamforming with frequent and costly beam training and alignment [3].

Vision-aided wireless communication has emerged as a promising solution by leveraging visual sensors to capture environmental dynamics that are not directly observable through radio frequency (RF) signals [4]. Fusing visual and RF data enables proactive blockage prediction, user localization, and channel state inference with reduced overhead. Nevertheless, progress is limited by the lack of open, real-world multimodal datasets, as most studies rely on simulations or loosely synchronized measurements. This CONVERGE challenge, based on CONVERGE research infrastructure [5, 6], addresses this gap by providing a large-scale, tightly synchronized visual–RF dataset collected in a realistic experimental environment. Participants are invited to develop machine learning solutions that jointly exploit visual and radio data for high-frequency communications, focusing on blockage prediction, user equipment (UE) localization, and channel prediction.

This work was supported by the CONVERGE project which has received funding under the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101094831. Project and challenge website: https://converge-project.eu.

2. EXPERIMENTAL SETUP AND DATASET

This section outlines the experimental setup and the synchronized multimodal dataset used in the challenge.

2.1. Experimental Setup

The challenge dataset is collected using the CONVERGE experimental infrastructure, designed for research on Integrated Sensing and Communication (ISAC). Measurements are conducted in the CONVERGE chamber at INESC TEC / FEUP (Portugal) with a LiteOn FR2 OpenAirInterface gNB mounted on a controllable robotic arm and a Quectel FR2 UE placed on a tripod or a mobile Turtlebot 4 robot. Controlled blockage and non-line-of-sight (NLoS) conditions are created using a programmable RF-shield curtain, while visual context is captured by a Nerian Ruby RGB-D camera mounted on the gNB. The dataset includes diverse scenarios with static and mobile UEs, different beam alignments, and varying blockage conditions.

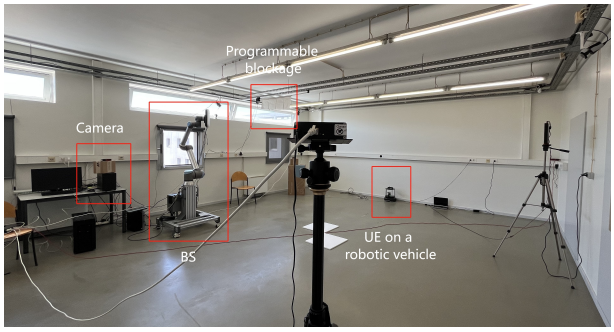

Fig. 1. CONVERGE chamber with mobile BS, UE, camera and programmable obstacle at INESC TEC / FEUP, Portugal.

2.2. Data Modalities

All modalities in the dataset are time-synchronized using Precision Time Protocol (PTP), enabling accurate cross-modal learning. The dataset includes synchronized visual, radio, and ground-truth data.

Visual Data: A RGB-D camera mounted on the gNB captures stereo RGB images and disparity maps at 7 fps, providing a gNB-centric view of the UE and surrounding obstacles. Also, camera calibration parameters are provided for metric-scale processing.

Radio Data: RF measurements include E2 interface metrics collected every 50 ms (e.g., SINR, RSRP, CQI, MCS, BLER, and throughput), and Sounding Reference Signal (SRS) measurements captured every 10 ms, consisting of complex-valued measurements over 2 antennas, 12 subcarriers, and 128 physical resource blocks.

Ground Truth: High-precision UE positions are obtained using a Qualisys motion capture system. Blockage states are manually annotated, and channel estimates are derived from pilot-based least-squares estimation with subcarrier interpolation.

3. CHALLENGE TASKS

The challenge consists of three independent tasks. For all tasks, participants may use visual and/or radio data available up to time t, but not future information, to ensure causality.

Task 1: Blockage Prediction. High-frequency wireless links are vulnerable to environmental blockages, causing abrupt failures. This task focuses on predicting the link blockage status for the next video frame ($t + 142$ ms) to enable proactive adaptation, with three classes: no blockage, partial blockage, and full blockage.

Task 2: UE Localization. Knowing the spatial relationship between the transmitter and receiver can reduce the beam search space and aids in channel estimation. This task focuses on estimating the current 3D translation (x, y, z) of the UE relative to the gNB within the gNB's local coordinate system. Estimates should be generated for each visual frame, corresponding to a sampling interval of 142 ms.

Task 3: Channel Prediction. Acquiring full channel state information (CSI) typically requires significant pilot overhead. While historical channel data captures temporal correlations, visual data provides specific environmental context, such as spatial geometry and potential obstructions, that is valuable for inferring channel dynamics. In this task, participants need to predict the complete, complex-valued SRS channel matrices for the next E2 frame ($t + 50$ ms).

4. BASELINES AND EVALUATION

4.1. Baseline Solutions

We provide a modular baseline architecture for all tasks, comprising:

1) Visual Branch: Processes RGB-D inputs using either a lightweight 3-layer CNN or a pre-trained ResNet18 backbone. The first layer is adapted to handle varying channel dimensions, supporting RGB images, disparity/depth maps, or 4-channel RGB-D inputs;

2) Radio Branch: Encodes wireless data (E2 or SRS data) into low-dimensional embeddings using a multi-layer perceptron (MLP);

3) Fusion Head: Concatenates the features from both branches and passes them through a task-specific MLP to produce the final output: categorical blockage state for Task 1, 3D coordinates for Task 2, or SRS channel matrices for Task 3. This baseline framework supports flexible configurations, allowing training on single modality (radio-only or vision-only) or fused data.

4.2. Evaluation

The preliminary evaluation metrics are summarized in Table 1. Submissions are evaluated using a composite score $S = \alpha\hat{P} + (1 - \alpha)\hat{L}$, balancing normalized performance $\hat{P}$, and total execution time $\hat{L}$. We set $\alpha = 0.7$ to prioritize accuracy while penalizing high computational overhead. Participants are provided with a training dataset to develop their models. To ensure a fair comparison, all evaluations are conducted on a separate, unseen testing dataset using a standardized hardware environment.

Table 1. Summary of evaluation metrics for each task.

Task	Primary Metric (P)	Definition
1. Blockage Prediction	**F1-Score**	Harmonic mean of precision/recall
2. UE Localization	**RMSE (m)**	Root mean squared error
3. Channel Prediction	**NMSE (dB)**	Normalized mean squared error

Table 2. Results for the three challenge tasks. Teams are ranked by the composite score (S). The top-ranked team and the best results are highlighted in bold. The baseline solution is included for comparison.

Rank	Team Name	Primary Metric	Total Exec. Time (ms)	Score (S)
	Task 1: Blockage Prediction *(F1-Score ↑)*			
1	**BeamSync**	**0.84**	16784	**0.92**
2	Signal Verse	0.76	20206	0.81
–	*Baseline*	0.63	**6066**	0.75
3	RayWise	0.57	8056	0.67
4	ISL-AAST	0.58	12354	0.64
5	CYO Wins	0.46	8982	0.53
6	BabyBus	0.26	9467	0.28
7	BUET_PixelWave	0.25	47089	0.00
	Task 2: UE Localization *(RMSE [m] ↓)*			
1	**ChillyByte**	**0.78**	**10364**	**1.00**
–	*Baseline*	0.93	19540	0.91
2	Brasil6G-UFRJ-CEFET	0.99	33057	0.87
3	BUET_PixelWave	1.03	221526	0.58
4	BeamSync	1.66	42502	0.53
5	EMCAS	1.74	83040	0.44
6	RayWise	2.25	15592	0.29
	Task 3: Channel Prediction *(NMSE [dB] ↓)*			
1	**BabyBus**	**0.02**	37674	**0.89**
2	2check1mate	0.81	70681	0.48
–	*Baseline*	2.01	**17981**	0.45
3	RayWise	2.53	57813	0.07

5. CHALLENGE RESULTS AND CONCLUSION

The challenge attracted diverse participation, with 26 registrations from global teams yielding 11 valid submissions across the three tasks. Table 2 summarizes the performance of these submissions compared to the official baseline.

In **Task 1**, BeamSync secured the top rank with an F1-score of 0.84, outperforming the baseline (0.63). SignalVerse also exceeded the baseline, achieving an F1-score of 0.76, while the baseline remained the most time-efficient solution. In **Task 2**, ChillyByte delivered exceptional performance, being the only team to surpass the baseline in both localization accuracy and inference execution time. In **Task 3**, BabyBus achieved a outstanding NMSE of 0.02 dB, far exceeding the baseline (2.01 dB). 2check1mate team also exceeded the baseline solution with a NMSE of 0.81 dB. Overall, despite the composite score (S) designed to balance performance and efficiency, we observed that many models suffered from overfitting, resulting in poor generalization on the held-out test set. Furthermore, several computationally intensive submissions incurred high execution time penalties without delivering commensurate performance gains.

This CONVERGE challenge represents a significant step toward realizing robust 6G networks through multimodal sensing. The results demonstrate that utilizing visual data can effectively support high-frequency wireless communications. Notably, the superior performance of the lightweight baseline suggests that future research should seek a better balance between model complexity, generalization, and real-time inference efficiency. We anticipate that the public release of this dataset will further catalyze collaboration across the wireless, computer vision, and artificial intelligence domains, paving the way for more reliable communication systems.

6. REFERENCES

[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] F. B. Teixeira, M. Ricardo, A. Coelho, H. P. Oliveira, P. Viana, N. Paulino, H. Fontes, P. Marques, R. Campos, and L. Pessoa, "Converge: towards an efficient multi-modal sensing research infrastructure for next-generation 6 g networks," *EURASIP Journal on Wireless Communications and Networking*, 2025.

[6] F. B. Teixeira, C. Simões, P. Fidalgo, W. Pedrosa, A. Coelho, M. Ricardo, and L. M. Pessoa, "Converge: A multi-agent vision-radio architecture for xapps," in *2024 IEEE Globecom Workshops (GC Wkshps)*, 2024, pp. 1–7.