NEURAL INFORMATION PROCESSING SYSTEMS

CSI-TextBench: A Dataset and Benchmark for Language-Grounded Ambient Perception with WiFi

Guozhen Zhu, Yuqian Hu, Sakila S Jayaweera, Wei-Hsiang Wang, Beibei Wang, K. J. Ray Liu

ADT Research

Presenter: Guozhen Zhu

Email: gzzhu@terpmail.umd.edu

Project page:

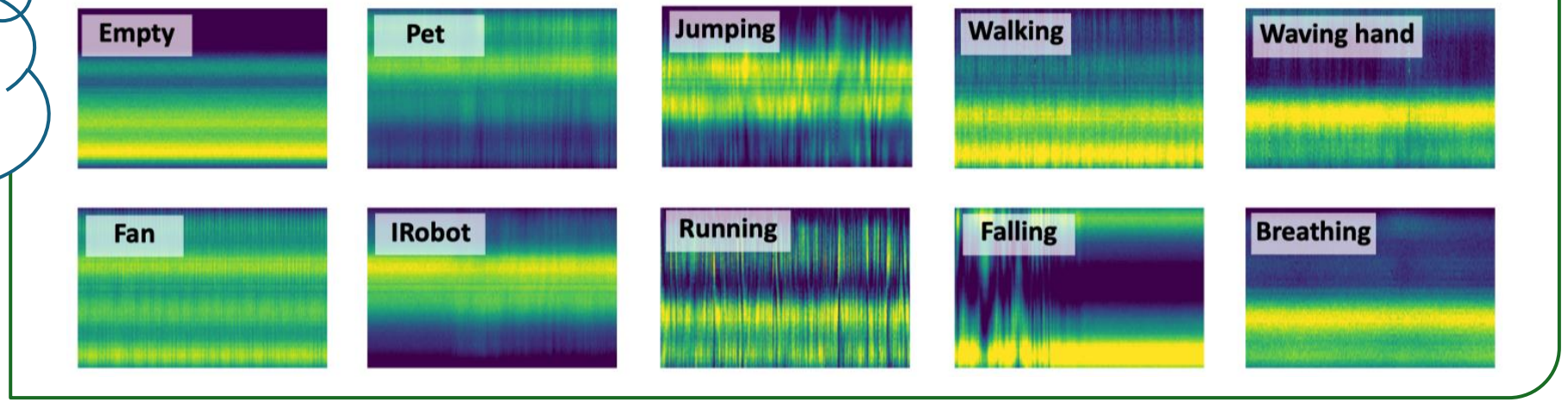
ORIGIN℠ | ADT

Ubiquitous WiFi Sensing

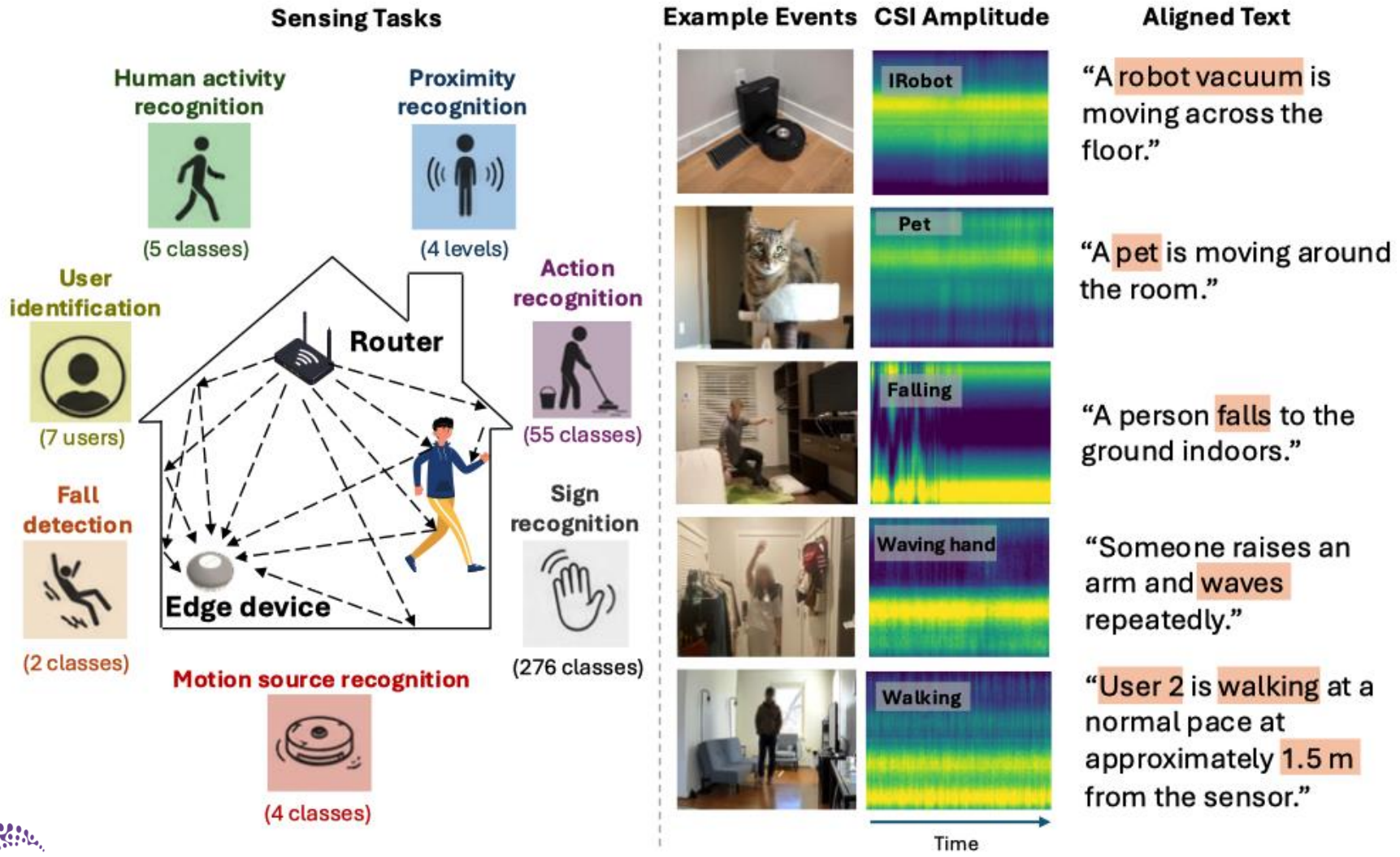
- Contactless
- Operates in the dark
- Works through walls
- Reuses existing infrastructure
- Privacy-preserving



Can we query WiFi with natural language?



Dataset

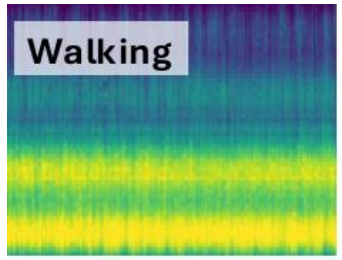


286k	task-level examples
7	sensing tasks
2.0M	CSI–text pairs
2,291	descriptions
4–9	descriptions per CSI

CSI-Bench XRF55 SignFi

Benchmark Design

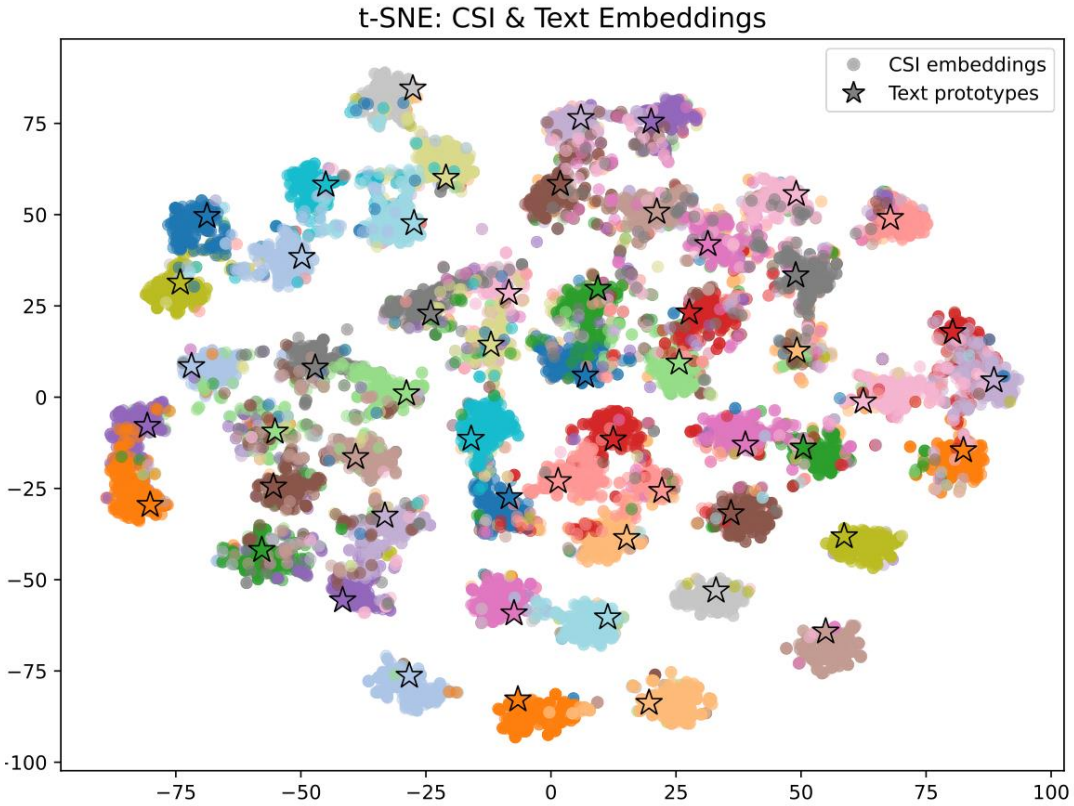
WiFi CSI



CSI encoder

"A person is walking"

Frozen text encoder



Aligned CSI and text embeddings

5 alignment methods

Match signals to candidate descriptions

Evaluation

Known classes

Method	HAR	Proximity	User ID	Fall	Motion Source	SignFi	XRF55
CLIP InfoNCE	**97.8**±0.6	**94.6**±0.4	**100.0**±0.0	94.2±0.0	**99.5**±0.2	**97.0**±1.0	77.5±0.4
SigLIP	95.7±2.8	84.4±3.8	100.0±0.0	**95.5**±0.5	99.5±0.3	71.8±32.2	74.5±1.2
LanguageBind	95.1±2.8	76.0±3.3	100.0±0.0	95.3±0.3	99.4±0.1	75.7±29.8	77.6±0.4
Barlow Twins	92.6±1.1	88.8±5.6	98.8±1.3	59.8±6.8	63.2±1.4	61.3±27.4	65.1±3.8
Masked Pred.	94.2±2.8	92.1±2.7	99.2±0.6	94.1±0.6	96.1±1.8	95.7±6.0	**78.3**±2.0
Supervised	97.9	96.3	100.0	95.5	99.8	99.4	82.9

CLIP is within 0.3 percentage points of supervised accuracy on three tasks with known classes.

Unseen classes

SignFi: train 200 classes, hold out 76

Zero-shot	1-shot	10-shot
11.3	54.0±3.0	**74.6**±1.1

A frozen encoder supports few-shot transfer. Ten labeled examples per class give 74.6% accuracy.

Evaluation

Open Vocabulary

HAR: same activities, new wording

Training template	Everyday	Expert (raw)	Sensor (raw)
95.9	**96.0**	59.1	35.4

The model handles unseen everyday wording. New prompts reach 96.0%, versus 95.9% for templates.

Complete Scene Understanding

Attribute-level QA

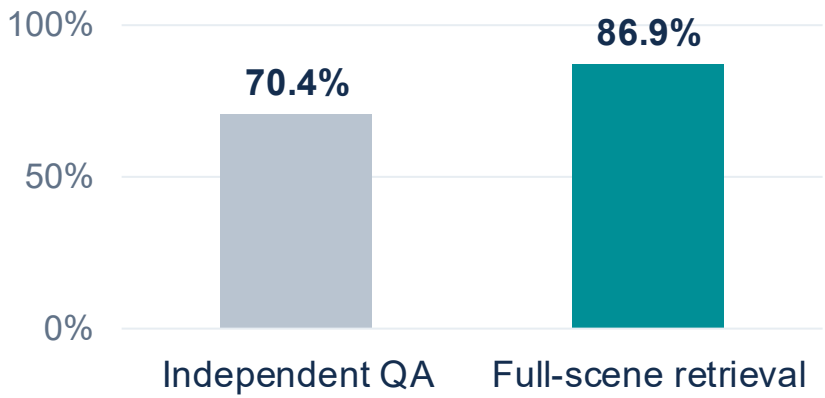
Q: What is the person doing?
A: Walking.

Q: How far is the person from the sensor?
A: Approximately 1.5 m

Q: Who is present?
A: User 2.

Composite Scene Understanding

Q: What is happening in the scene?
A: User 2 is walking at approximately 1.5 m from the sensor.



Joint descriptions recover the full scene more accurately than separate attribute predictions. The gain is 16.5 percentage points.

Access Dataset, Code, Paper, and Documentation



https://ai-csi-sensing.github.io/csi-text-bench/index.html