Google Research
Google DeepMind

SensorLM: Learning the language of wearable sensors

Yuwei Zhang*, Kumar Ayush*, Siyuan Qiao, A. Ali Heydari, Girish Narayanswamy, Maxwell A. Xu, Ahmed Metwally, Shawn Xu, Jake Garrison, Xuhai Xu, Tim Althoff, Yun Liu, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Cecilia Mascolo, Xin Liu, Daniel McDuff, and Yuzhe Yang*
*corresponding authors: yz798@cam.ac.uk, {kmrayush, yuzheyang}@google.com

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

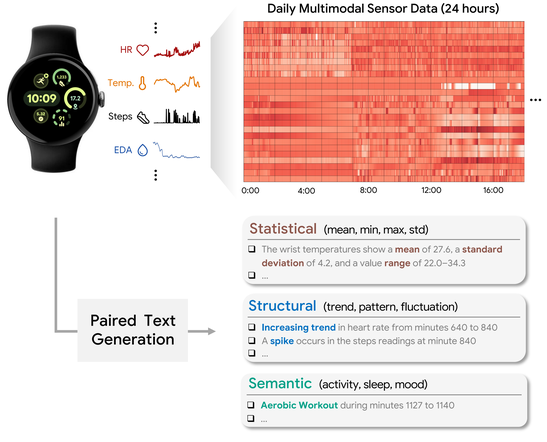
Generalist Models for Wearable Data

Contributions

- **Hierarchical Captioning Pipeline**: generates statistical, structural, and semantic captions for raw sensor data.
- **Largest Sensor-Language Dataset:** 59M+ hours of paired data enabling large-scale training.
- **Unified Foundation Model:** extends CLIP/CoCa to sensor data, excelling in zero-/few-shot tasks, retrieval, and captioning.

Dataset Details

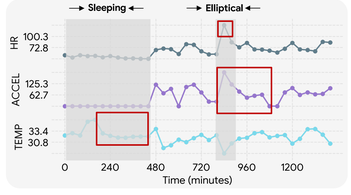
- 59.7 million data hours from 103,000+ subjects.
- 5 sensor modalities: PPG, accelerometer, skin conductance, skin temperature, altimeter.
- Input of 26 minute-aggregated sensor features x 24 hour window.



Method

Captioning

Hierarchical Captioning Pipeline

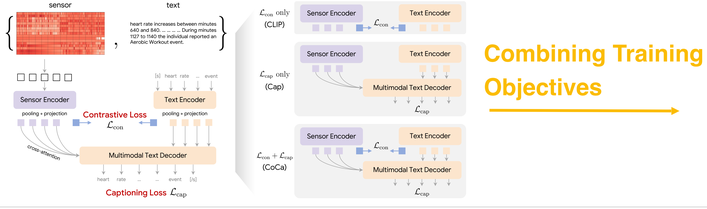


Generation of the largest paired sensor-language dataset to date!

Enabling Large-Scale Sensor-Language Alignment

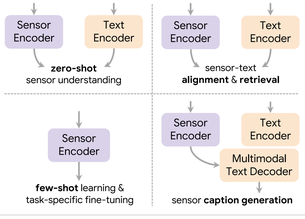
Training

Modular Multimodal Pretraining Architecture



Combining Training Objectives

New Capabilities



Zero-Shot AR

Zero-shot activity and concept classification

(a) **Activity recognition (20-class)**

Metrics	AUROC↑	F1↑	BAcc↑
Gemma-3-27B [30]	0.50	0.01	0.05
Gemini 2.0 [29]	0.51	0.03	0.07
Gemini 2.0 (SFT) [29]	–	0.06	0.10
SensorLM	**0.84** (+33%)	**0.29** (+23%)	**0.31** (+21%)

(b) **Activity by environmental context**

Metrics	AUROC↑	F1↑	BAcc↑
Gemma-3-27B [30]	0.51	0.22	0.25
Gemini 2.0 [29]	0.50	0.19	0.25
Gemini 2.0 (SFT) [29]	–	0.25	0.28
SensorLM	**0.64** (+13%)	**0.33** (+08%)	**0.38** (+10%)

(c) **Cardio *vs.* strength training**

Metrics	AUROC↑	F1↑	BAcc↑
Gemma-3-27B [30]	0.53	0.42	0.49
Gemini 2.0 [29]	0.50	0.39	0.50
Gemini 2.0 (SFT) [29]	–	0.44	0.53
SensorLM	**0.71** (+18%)	**0.63** (+19%)	**0.66** (+13%)

(d) **Locomotion *vs.* stationary**

Metrics	AUROC↑	F1↑	BAcc↑
Gemma-3-27B [30]	0.51	0.40	0.51
Gemini 2.0 [29]	0.55	0.52	0.55
Gemini 2.0 (SFT) [29]	–	0.53	0.55
SensorLM	**0.61** (+06%)	**0.58** (+05%)	**0.58** (+03%)

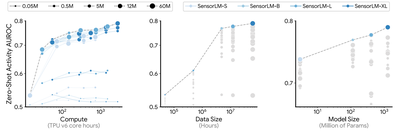
(e) **Fine-grained recognition (gym cardio)**

Metrics	AUROC↑	F1↑	BAcc↑
Gemma-3-27B [30]	0.49	0.16	0.25
Gemini 2.0 [29]	0.52	0.20	0.26
Gemini 2.0 (SFT) [29]	–	0.28	0.35
SensorLM	**0.76** (+24%)	**0.50** (+22%)	**0.51** (+16%)

(f) **Fine-grained recognition (outdoor sports)**

Metrics	AUROC↑	F1↑	BAcc↑
Gemma-3-27B [30]	0.50	0.18	0.25
Gemini 2.0 [29]	0.54	0.22	0.29
Gemini 2.0 (SFT) [29]	–	0.26	0.32
SensorLM	**0.83** (+29%)	**0.52** (+26%)	**0.53** (+21%)

Resource Scaling (Compute, Data, Model Size)



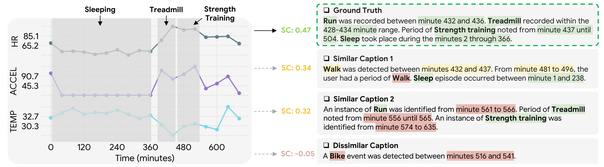
Generalizable to unseen activities!

New Capabilities

Retrieval

Sensor-to-Text Retrieval: Given a sensor data segment, find the best text description.

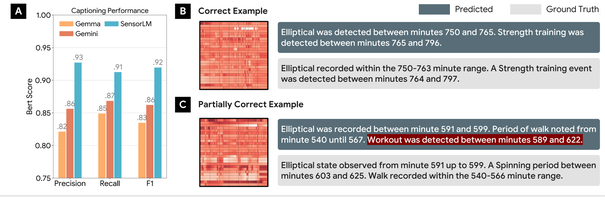
- **SensorLM achieves near-perfect retrieval** (0.96 Recall@1 on 50k samples)
- **LLM baselines struggle** (<0.1 Recall@1 on 100 samples).



Captioning

Caption Generation: Generates descriptive natural language from sensor data.

- **SensorLM generates much more meaningful captions compared with LLMs.**



Ablations

Comparing Caption Variants

Caption Variant			Zero-Shot	Linear Probing	
statistical	structural	semantic	Activity	Activity	Anxiety
✓	✗	✗	0.51	0.76	0.67
✗	✓	✗	0.50	0.78	0.63
✗	✗	✓	0.71	**0.95**	0.65
✓	✗	✓	0.66	0.84	**0.68**
✗	✓	✓	**0.84**	0.94	0.65
✓	✓	✗	0.49	0.79	0.67
✓	✓	✓	0.66	0.86	**0.68**

Comparing Model Variants

Arch Variant	Zero-Shot		Linear Probing	
	AUROC↑	F1↑	AUROC↑	F1↑
SensorLM (CLIP)	0.83	**0.29**	0.93	0.53
SensorLM (SigLIP)	0.78	0.17	0.87	0.38
SensorLM (Cap)	0.55	0.01	0.90	0.32
SensorLM (CoCa)	**0.84**	**0.29**	**0.94**	**0.57**