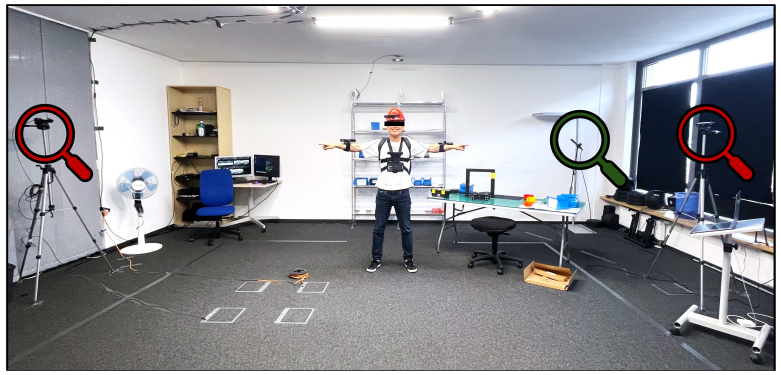
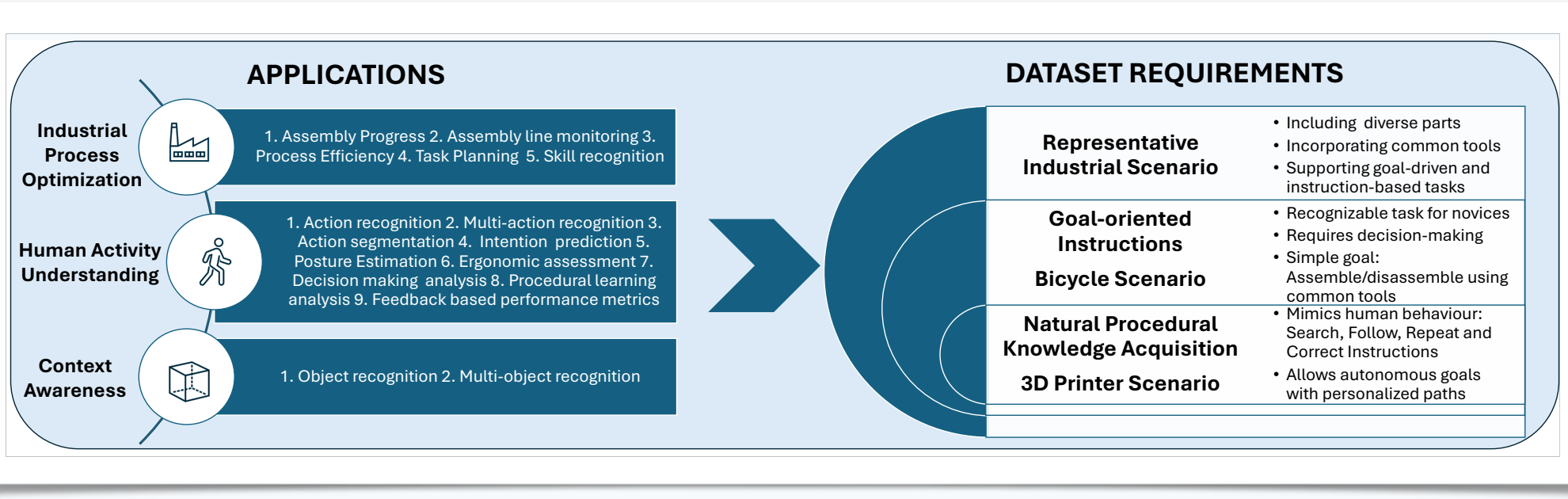
OpenMarcie

CVPR JUNE 3-7, 2026 DENVER COLORADO

Dataset for Multimodal Action Recognition in Industrial Environments

Hymalai Bello, Lala Ray, Joanna Sorysz, Sungho Suh and Paul Lukowicz





Why another industrial HAR dataset?

Existing datasets leave three gaps:

- Limited synchronized multimodality across wearables, vision, and audio.
- Highly constrained tasks that miss open-ended industrial workflows.
- Few datasets model overlapping actions such as walking while carrying a tool.

OpenMarcie response

- Industrial assembly tasks with natural variation.
- Wearable + ego + exo sensing under a shared timeline.
- Multi-action, verb–object–tool labels for fine-grained analysis.

Capability	Ours	Typical
Wearable IMU multi-position	✓	✗
Egocentric + exocentric views	✓	✓ / partial
Open-vocabulary captioning	✓	✗
Concurrent multi-action labels	✓	✗
Full industrial coverage	✓	✗

Dataset Comparison

OpenMarcie at a glance

37+ h	36	8	200+	3
synchronized multimodal recordings	participants across two settings	distinct data types	independent channels	benchmark tasks

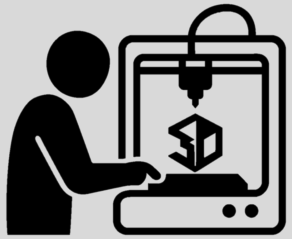
Two complementary assembly regimes

AD-HOC BICYCLE ASSEMBLY : Scenario (a)

- Standard tools (hex key, screwdriver, hammer, wrench, scissors, pump)
- Goal oriented scenario as assembly/disassembly a bicycle without instructions.
- Represent a simpler and more familiar task for the volunteers.

PROCEDURAL 3D PRINTER ASSEMBLY : Scenario (b)

- Sequential collaborative assembly between participants.
- Assessment of progress and correction of prior errors if needed.
- ˜15 large, ˜40 medium and ˜150 small size components.
- Frame and Structural Elements
- Motion Components
- Electronics
- Plastic Parts
- Cooling Fan
- Fasteners
- Miscellaneous (e.g. zip ties)
- Hobbyist level instructions with different task precedence.



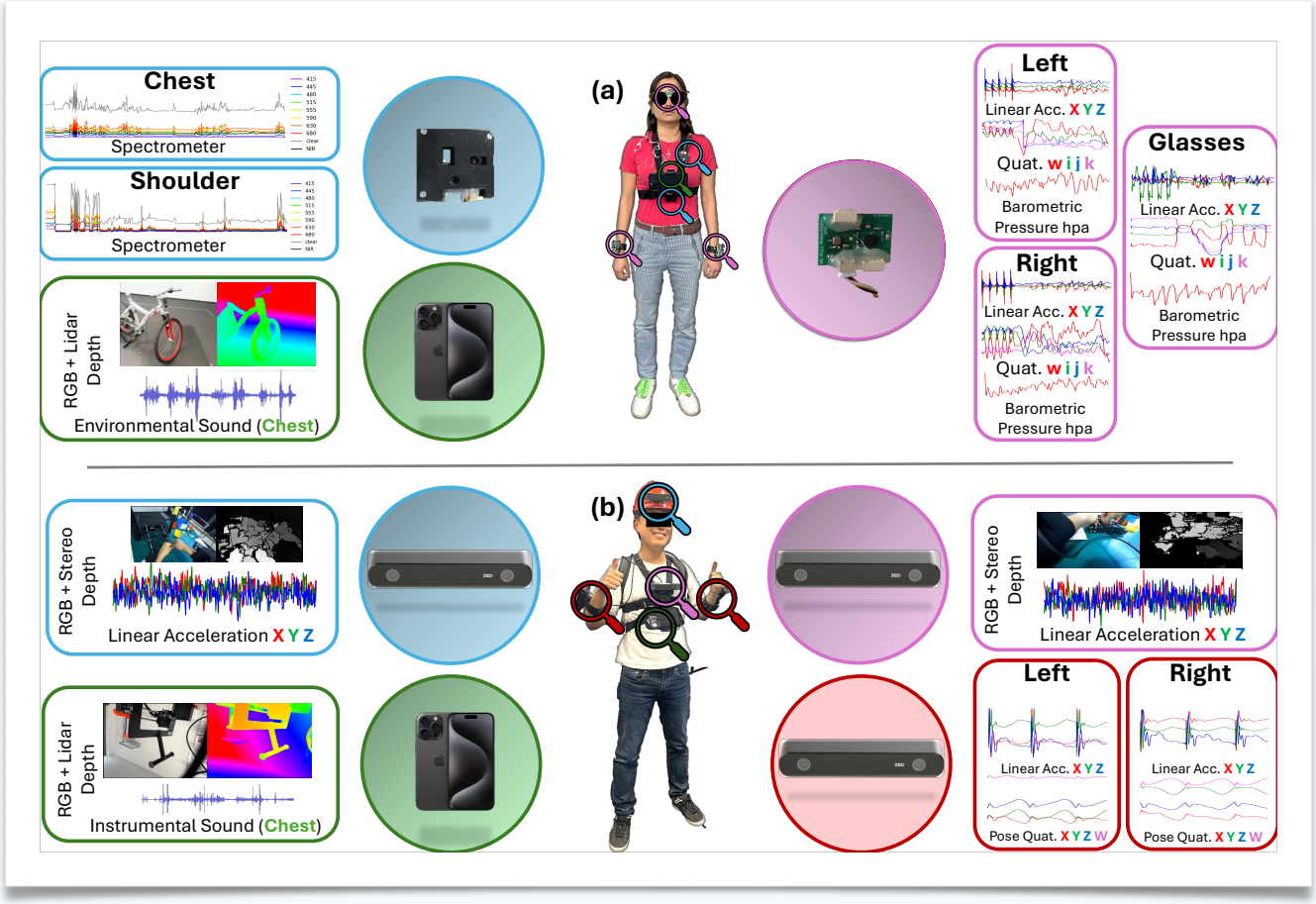
Sensing and annotations are aligned across time

Modalities

- Egocentric RGB-D / LiDAR
- Exocentric RGB-D cameras
- Wrist IMUs + barometers
- Thermal and spectrometer signals
- Stereo audio and narration

Annotation strategy

- Intent-aware verb–object–tool labels.
- Multi-label segments for concurrent actions.
- Soft natural-language captions plus hard labels for model training.



Three validation benchmarks

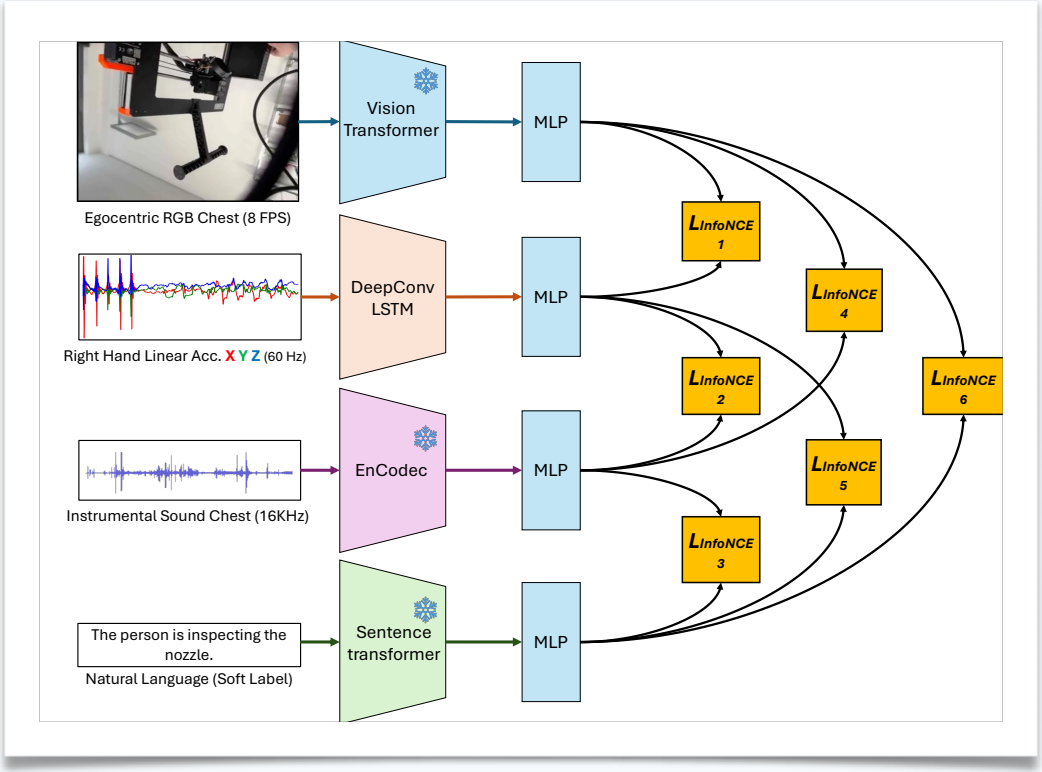
1
Human Activity Recognition

2
Open-vocabulary Captioning

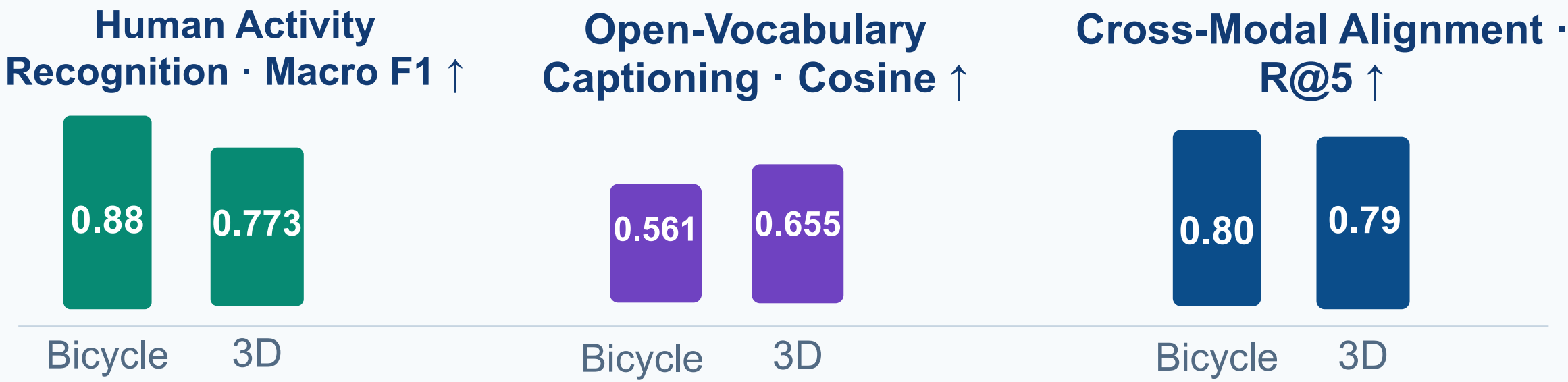
3
Cross-modal Alignment

Evaluation design

- Unimodal, pairwise, and three-way fusion settings.
- Vision: ViT; IMU: DeepConvLSTM; audio: EnCodec + temporal classifier.
- Text alignment uses contrastive learning over synchronized positive pairs.
- Subject-disjoint splits test generalization across participants.



Main result: inertial + vision is consistently strongest



Interpretation

- Vision captures object and scene context; IMU captures hand dynamics.
- Audio alone is weaker in the controlled test bench, but can still contribute in fusion.
- Full fusion gains taper, suggesting redundancy and modality-specific limits.

Takeaway for CVPR

OpenMarcie is a controlled, industrially inspired benchmark for multimodal perception in assembly workflows.

Contribution

wearables + ego/exo video + multimodal labels

Use cases

fine-grained HAR, captioning, alignment, HRC

FUTURE WORK

- richer long-term procedural annotations
- object / action / pose grounding
- domain shift evaluation in new environments
- embodied AI and human-robot team benchmarks