

OpenMarcie: Dataset for Multimodal Action Recognition in Industrial Environments

Hymalai Bello^{1†} Lala Ray^{1†} Joanna Sorysz^{1,2} Sungho Suh^{3*} Paul Lukowicz^{1,2}

¹DFKI Kaiserslautern ²RPTU Kaiserslautern-Landau ³Korea University

{Hymalai.Bello, Lala.Shakti.Swarup.Ray, Joanna.Sorysz, Paul.Lukowicz}@dfki.de
 sungho.suh@korea.ac.kr

[†]Equal contribution *Corresponding author

Abstract

Smart factories use advanced technologies to optimize production and increase efficiency. To this end, the recognition of worker activity allows for accurate quantification of performance metrics, improving efficiency holistically while contributing to worker safety. OpenMarcie is, to the best of our knowledge, the biggest multimodal dataset designed for human action monitoring in manufacturing environments. It includes data from wearables sensing modalities and cameras distributed in the surroundings. The dataset is structured around two experimental settings, involving a total of 36 participants. In the first setting, twelve participants perform a bicycle assembly and disassembly task under semi-realistic conditions without a fixed protocol, promoting divergent and goal-oriented problem-solving. The second experiment involves twenty-five volunteers (24 valid data) engaged in a 3D printer assembly task, with the 3D printer manufacturer’s instructions provided to guide the volunteers in acquiring procedural knowledge. This setting also includes sequential collaborative assembly, where participants assess and correct each other’s progress, reflecting real-world manufacturing dynamics. OpenMarcie includes over 37 hours of egocentric and exocentric, multimodal, and multipositional data, featuring eight distinct data types and more than 200 independent information channels. The dataset is benchmarked across three human activity recognition tasks: activity classification, open vocabulary captioning, and cross-modal alignment. The dataset and code are available at [OpenMarcie](#).

1. Introduction

The advancement of smart factories depends on the integration of human intelligence into automated systems, promoting a more efficient, adaptive, and human-centered industrial paradigm. A fundamental component of this integra-

tion is human activity recognition (HAR), which enables systems to understand, support, and optimize human motion within complex industrial workflows. Typically, video data has served as a rich source of valuable information for HAR. As a result, extensive research efforts are dedicated to developing innovative vision-based methods and video-based datasets to support the community [6, 8], including event cameras [27]. Tab. 1 presents a comparison with state-of-the-art dataset for HAR in industrial environments. Despite recent advances, existing HAR datasets in industrial contexts suffer from three major limitations: (1) a lack of truly multimodal data combining wearable sensors, vision, and audio in a synchronized manner, (2) a reliance on highly constrained, protocol-driven tasks, which do not reflect the open-ended, procedural nature of real-world industrial work, and (3) limited demographic diversity or task complexity. In general, most datasets focus on short, isolated actions, failing to capture the extended, multi-step activities typical of human workflows in manufacturing. Furthermore, human action is inherently multimodal, integrating sensory, cognitive, and motor processes. Actions depend on visual, auditory, tactile, while cognitive and emotional states shape movement, speech, and expression [23]. Variability in motion, context dependence, and environmental interactions add further complexity, making multimodal analysis essential for accurate interpretation [2, 3]. To effectively understand and replicate human activity, AI and robotics must process diverse data sources, including video, audio, and motion sensors. Wearable and multipositional data sources and synchronized visual information (egocentric and exocentric) have proven to be highly valuable [9, 28, 29]. This is especially true in industrial environments, where vision-only-based methods may raise privacy concerns and increase the risk of technology leakage [4]. Accordingly, we present OpenMarcie (see Figure 1). Our contributions are threefold. First, we introduce OpenMarcie, a multimodal dataset with eight sens-

Table 1. Comparative table of human action recognition datasets in the industrial domain. Multi-actions indicates whether the dataset supports concurrent action labeling (e.g., walking while carrying), based on overlapping verb–object–tool annotations. Only OpenMarcie unifies **wearables**, **egocentric + exocentric multiview video**, **multi-action labels**, and **full industrial coverage**.

Dataset	Industrial	Wearables	Ego	Exo	Multiview	Multi-Act.	Scale / Focus
InHARD [9]	✓	✓	x	✓	✓	x	HRC assembly, RGB+IMU
LARa [18]	✓	✓	x	✓	x	x	Logistics, IMU+mocap
OpenPack [28]	✓	✓	x	✓	x	x	Logistics, 50h+ IMU/IoT
Assembly101 [25]	✓	x	✓	✓	✓	x	Vision-only, procedural assembly
IKEA-ASM [5]	✓	x	x	✓	✓	x	Exo RGB-D, furniture assembly
HA4M [7]	✓	x	x	✓	x	x	Exo RGB-D+IR+pose
HA-ViD [29]	✓	x	x	✓	✓	x	Rich semantic labels, multi-route
IndustReal [24]	✓	x	✓	x	x	x	PSR, ego-only, error modeling
Ego-Exo4D [14]	(~6% ind.)	✓	✓	✓	✓	x	1,200h skilled activities
OpenMarcie (ours)	(100%)✓	✓	✓	✓	✓	✓	Industrial multitasking, 8 modalities

ing modalities, including LiDAR depth, multi-view pose estimation, object detection, and precise positioning. Second, the dataset captures realistic industrial assembly and disassembly tasks where participant decisions shape alternative routes, enabling the study of procedural knowledge acquisition. Third, we provide rich metadata from 36 volunteers, synchronized wearable sensing streams, speech-to-text narrations, and detailed manual annotations that capture overlapping, multi-primitive actions—supporting fine-grained analysis of complex workflows.

2. OpenMarcie dataset

Two primary experimental settings are defined: Ad-hoc and Procedural scenarios, featuring bicycle and 3D printer assembly/disassembly, respectively, as shown in Figure 1. The scenarios involved a diverse set of skills, including: Mechanical competencies, such as tool usage, component alignment, and fastening procedures; Cognitive processes, including interpretation of visual instructions or schematics and problem-solving when inconsistencies arise; Fine motor skills, essential for manipulating small or intricate components with precision. We select two contrasting tasks: bicycle assembly, a familiar, goal-oriented scenario with open-ended actions, and 3D printer assembly, a highly procedural task requiring interpretation of detailed instructions and unfamiliar components. Together, they capture unscripted repair and structured production-line assembly while incorporating broader activities such as wiring, cable routing, pulley installation, unpacking, and part organization. This diversity, enables benchmarking across real-world industrial workflows. Current benchmarks focus on activity classification, open-vocabulary captioning, and cross-modal alignment, establishing strong baselines for the dataset’s scope and utility. For both scenarios, 3 ZED X AI stereo cameras without polarizers are distributed as best as possible to cover the entire view of the experiment’s room. Figure 2 depicts the experimental room setting with example views of the exocentric cameras for the scenario (a) ad-hoc bicycle assembly and (b) procedural scenario 3D printer experiment. ZED stereo cameras come with a

rich SDK and allow customized AI methods to be included in the pipeline. In Tab. 2 the recorded sensing modalities for both scenarios are presented. In total, the number of raw channels available is around 282. This includes various sensor placements. Furthermore, each participant reviewed and signed an ethical agreement for each scenario, ensuring compliance with the Declaration of Helsinki. The Ethical Board of the German Research Center for Artificial Intelligence approved the study under agreement numbers HRW-35/24 and SMD-30/24.

2.1. Ad-hoc Scenario (a): Bicycle

The assembly and disassembly of a bicycle is proposed as a structured, goal-oriented task that allows participants to autonomously determine their approach. It reflects the nature of many practical, outcome-driven scenarios encountered in both industrial and experimental settings. Common errors, such as improperly tightened bolts or misaligned brakes, are generally straightforward to detect and rectify, contributing to the task’s safety and pedagogical value. Due to the widespread familiarity and accessibility of bicycles, this task is especially relevant for use in training environments, educational contexts, and research applications, including studies on human-robot collaboration and task planning. The volunteers were equipped with the wearable sensor suite illustrated in Figure 3. The setup includes inertial measurement units (IMUs) and barometric pressure sensors mounted on both hands and on the head (integrated into a glasses frame). Additionally, thermal and spectrometer sensors are positioned on the chest and shoulder. The system is complemented by an egocentric RGB-D camera, which includes LiDAR-based depth sensing. Figure 3 also presents sample visualizations of the multimodal sensor data collected during the task. For ground truth, we selected the most informative exocentric view (typically the door-side camera in bicycle assembly; Figure 2) and we (humans) manually annotated it using an intent-aware verb–object–tool scheme (Figure 4). This yields semantically meaningful segments aligned with task intent, including multi-label cases (e.g., walking while carrying). Each annotation (Verb, Tools, Object, Remarks) is

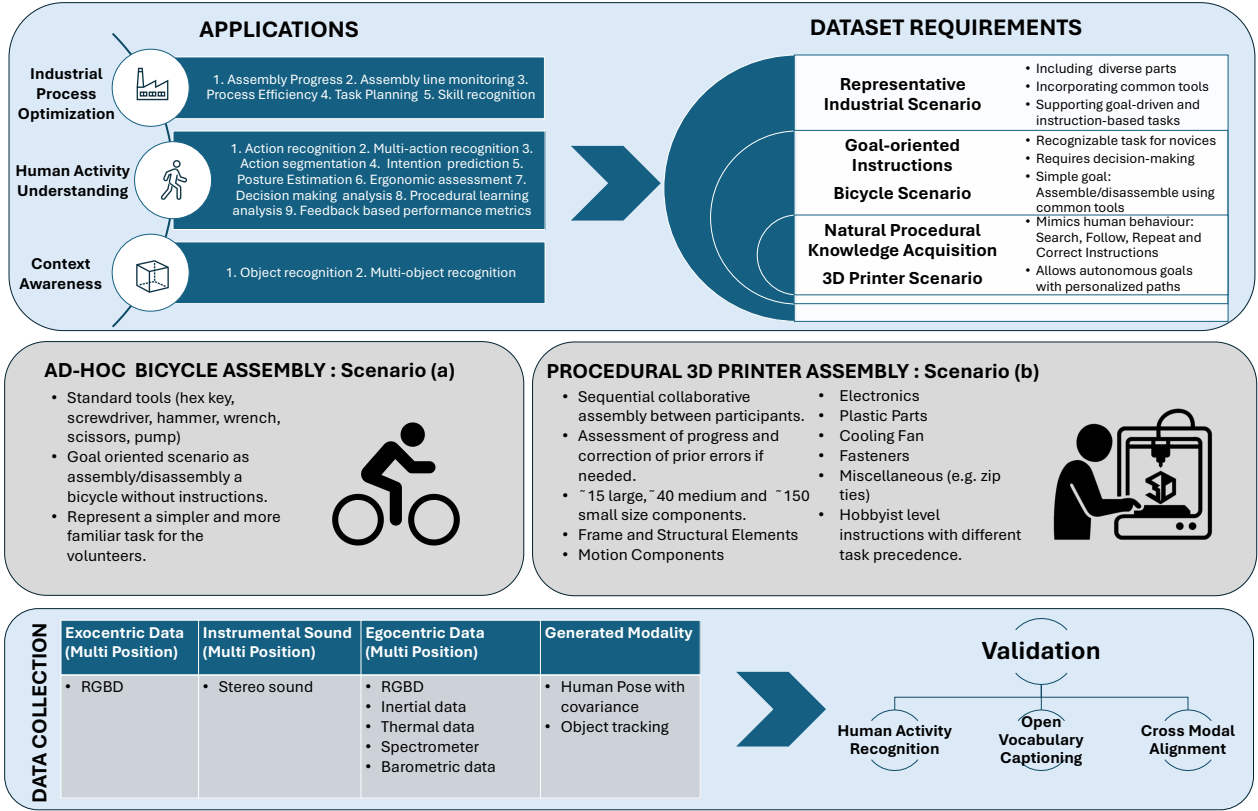


Figure 1. OpenMarcie is an open dataset for multimodal action recognition in industrial environments. **Top** Its applications include industrial process optimization, human activity understanding, and context awareness. **Middle** two assembly scenarios represent ad-hoc goal-oriented action and a procedural scenario to promote natural knowledge acquisition. **Bottom** multimodal data is collected with more than 200 independent channels, and a validation with three different benchmarks is done.

temporally aligned across all modalities (egocentric video, inertial, thermal, audio). We additionally generate concise natural-language descriptions with GPT-4o [1](see ??), providing soft labels that improve consistency and interpretability while preserving human-created ground truth.

2.2. Procedural Scenario (b): 3D Printer

Building a 3D printer requires significant cognitive effort to interpret written and video instructions. Additionally, it involves handling both small components, such as screws, and larger hardware parts like the metal main frame. Therefore, this scenario provides an ideal setting for monitoring human action recognition, closely resembling tasks commonly seen in industrial assembly lines. The Assemble Yourself 3D printer original Prusa i3 MK3S+ kit is used for this scenario. The creators offer a detailed set of assembly instructions, with each plastic part accompanied by its 3D model STL file for easy reprinting. This also means that the 3D models, combined with OpenMarcie’s egocentric videos, can enhance object detection through a predefined dictionary of parts. This approach intuitively monitors the printer’s assembly status and guides the builder toward the correct next step, accelerating the construction

Table 2. Sensing modalities information for each scenario

Scenario	Modality	Position	Ch
Ad-Hoc (a)	IMU	Wrists, Forehead	21
	Magnetometer	Wrists, Forehead	9
	Barometer	Wrists, Head	3
	Temperature	Wrists, Forehead	3
	Spectrometer	Chest, R. Shoulder	24
	Thermal Camera	Chest, R. Shoulder	(8x8)x2
	RGB-Lidar	Chest	4
	Stereo Sound	Chest	2
	RGBD	2 Exocentric cameras	8
Procedural (b)	IMU	Wrists	20
	Magnetometer	Wrists	6
	Barometer	Wrists	2
	Temperature	Wrists, Forehead,	4
	RGBD-IMU	Chest	28
	RGB-LiDAR	Forehead, Chest	8
	Stereo Sound	Chest, Table	4
	RGBD	Chest, Table	4
	RGBD	2 Exocentric cameras	8
Both Scenarios	17 modalities	9 positions	282 Ch

process. Before the experiment, each volunteer is required to watch a 30-minute video of the complete Prusa assembly guide, available online at <https://www.youtube.com/watch?v=uToqS1h64R4>. This ensures they understand the key aspects of the task without needing further clarification or instructions from the observer, allowing them to assemble the printer independently. The idea

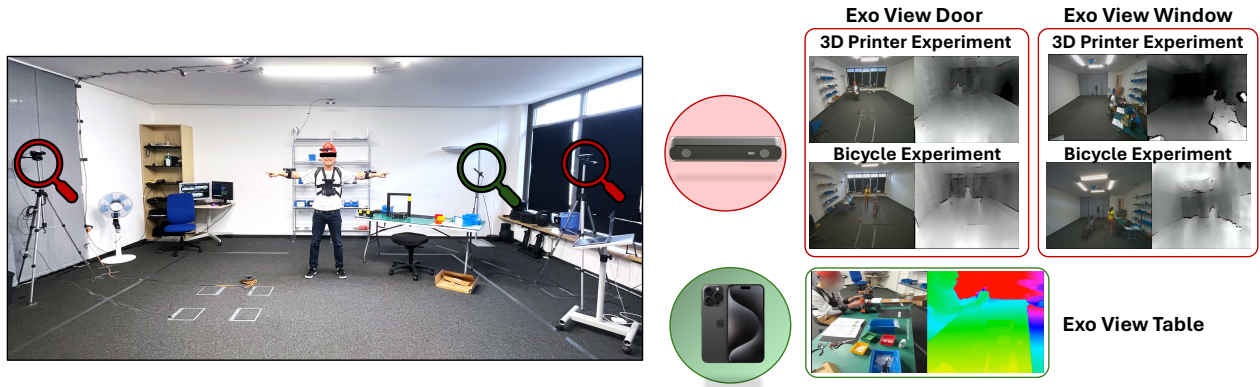


Figure 2. Experiment room setting with example views of the exocentric RGBD cameras.

is for one volunteer to start the assembly using only out-of-the-box instructions and settings, with approximately one hour to build as much as possible. The next volunteer would then continue from where the previous participant left off, and this process would repeat sequentially. This setup requires each subsequent participant to assess the current assembly status, understand the previous volunteer’s progress, and determine how best to proceed, increasing the cognitive challenge to the task. The participants wore 4 ZED 2 - AI Stereo Cameras. Two ZEDs were mounted on the wrists (used only for IMU and barometer signals, with stereo cameras disabled), one on the forehead via a helmet, and one on the chest via a strap to capture visual streams. On the person’s chest, an iPhone 15 Pro is set up with a cellphone cover with the camera and the Lidar sensor facing the front. The data is transferred by USB cables to a Jetson Orin AGX, which is placed inside a backpack together with a Green Cell Powerbank 26800mAh 128W to seamlessly record the data for one hour by a volunteer. During the experiment, an external human observer provided real-time narration of the participants’ actions and contextual details of the scene. The audio recordings of these narrations were subsequently transcribed using faster-whisper (v1.1.1) and ctranslate2 (v4.4.0), employing the “large-v3” model configuration [21]. Examples of the soft labels generated using the methods are depicted in Figure 4 **Scenario (b)**. The hard labels were then obtained via a two-stage pipeline. The process involves iteratively identifying and refining discrete action classes using Deepseek-r1 [15], then using GPT-4o [1] with prompt engineering to convert soft-label sentences into persistent hard labels for model training. This process—consisting of human annotation followed by a two-stage LLM-based translation—yields our final hard labels for downstream model training. In Scenario (b), we adopt a two-stage pipeline leveraging transcribed narration, unlike Scenario (a). Despite this methodological difference, both yield a consistent multimodal annotation format. To validate LLM-generated labels, we perform bidirectional con-

sistency checks (structured→caption→structured), achieving strong alignment (Macro F1 = 0.715 in Scenario a; ME-TEOR = 0.531 in Scenario b; see ??). These results support the reliability of LLMs as structured translators in our partially supervised pipeline. Moreover, sensor placements were adapted to suit the specific ergonomics and task demands of each scenario. While this design results in some variation—particularly between Scenario (a) (bicycle assembly) and Scenario (b) (3D printer assembly)—key wearable modalities remain consistent across both tasks. Most notably, inertial measurement units (IMUs) are placed on the wrists, wearable LiDAR units are mounted on the chest, and stereo microphones are consistently used. These shared sensor positions enable meaningful multimodal comparisons, particularly for HAR and captioning benchmarks. As illustrated in 3, this overlap is visually evident and supports comparative analysis despite scenario-specific adaptations.

3. Statistic

The study involved a diverse group of participants whose demographic and professional characteristics are summarized in Figure 5. In terms of height, participants ranged from 150 to 193 cm, with the largest group (31%) falling within the 181–193 cm range. The age distribution was skewed toward younger individuals, with 47% aged 22–24, 25% aged 25–29, and 28% aged 30–37. Participants self-reported their experience levels in assembly-related tasks: 57% identified as beginners, 37% as intermediate, and 6% as advanced. Regarding academic and professional background, engineers represented the majority at 72%, followed by computer scientists (14%), biologists (8%), physicists (3%), and managers (3%). This composition reflects a technically oriented, predominantly early-career participant pool with varying levels of practical experience in assembly tasks. The international diversity of the participants, who come from more than 20 countries, contributes to the cultural and experiential diversity of the study. Moreover, the majority of participants (31 individuals) were

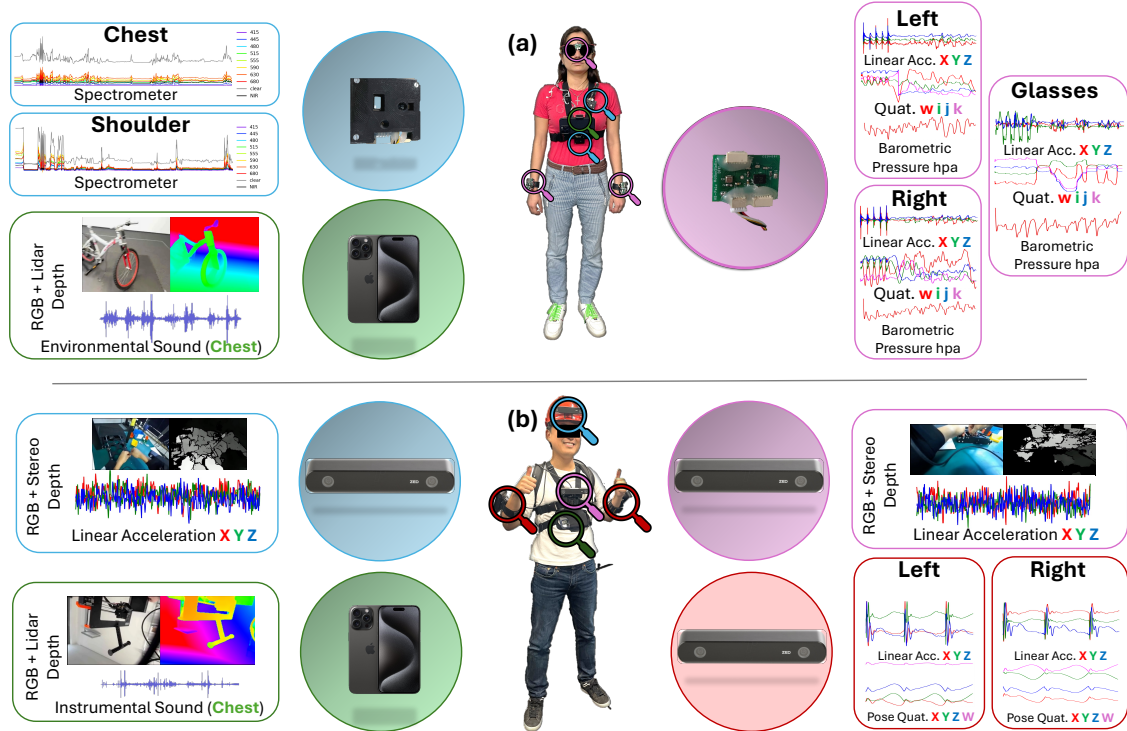


Figure 3. Participant wearable setup and sensor signals. **(a)** Participant in the ad-hoc Scenario: bicycle assembly. **(b)** Participant in the procedural Scenario: 3D printer assembly. Wrist-mounted ZEDs provided only IMU and barometer data (stereo cameras disabled), while head and chest cameras captured visual streams. Sensor placements were adapted to the ergonomics and task demands of each scenario.

right-handed, while a smaller subset (5 individuals) identified as left-handed. While the participant cohort predominantly comprises right-handed individuals with engineering backgrounds—potentially limiting demographic generalization—this demographic reflects the typical industrial workforce, thus enhancing ecological validity. Participants reported their ownership and usage habits related to bicycles and 3D printers (see Figure 6). A majority indicated that they owned a bicycle, while only a small fraction owned a 3D printer. Usage patterns reflected this ownership disparity: bicycles were used more frequently than 3D printers, with participants reporting daily, weekly, or monthly use. In contrast, the majority of participants reported either very infrequent use or no use of 3D printers. These differences highlight a greater familiarity and hands-on experience with bicycles among participants, making bicycle-related tasks more intuitive for the group, while 3D printer tasks likely required more instruction or exploration due to limited prior exposure. The participant characteristics provide valuable context for interpreting task performance and interaction behavior observed during the study. The combination of technical backgrounds, varied assembly experience levels, and broad international representation contributes to a rich and diverse dataset. Furthermore, the participants’ greater familiarity with bicycles compared to 3D printers offers a

natural contrast between ad-hoc and procedural tasks, reinforcing the relevance of the chosen scenarios for studying goal-oriented activities across different levels of prior knowledge and skill. Figure 7 represents an example of activity distribution within OpenMarcie. These actions represent a diverse set of activities that humans are expected to perform in industrial scenarios, specifically during assembly tasks. The dataset focuses on two distinct scenarios. Fig. 7 **Left** is the distribution of activities for the ad-hoc bicycle scenario (**Scenario (a)**). The most frequent activities are screwing/unscrewing and inspecting, as these actions are an integral part of handling bicycle parts. In contrast, hammering or lying down occurred at more defined stages of the assembly process. Fig. 7 **Right** shows the distribution of activities for the 3D printer assembly procedural scenario. Moving objects (“Move”), Adjust, and Read are the dominant categories. In contrast to the ad-hoc scenario, the Unlabeled class is less prominent in the 3D printer case.

4. Validation Benchmarks

To assess the utility and versatility of OpenMarcie, we establish validation benchmarks across three core tasks: Human Activity Recognition (HAR), Open Vocabulary Captioning, and Cross-Modal Alignment. These tasks are cho-

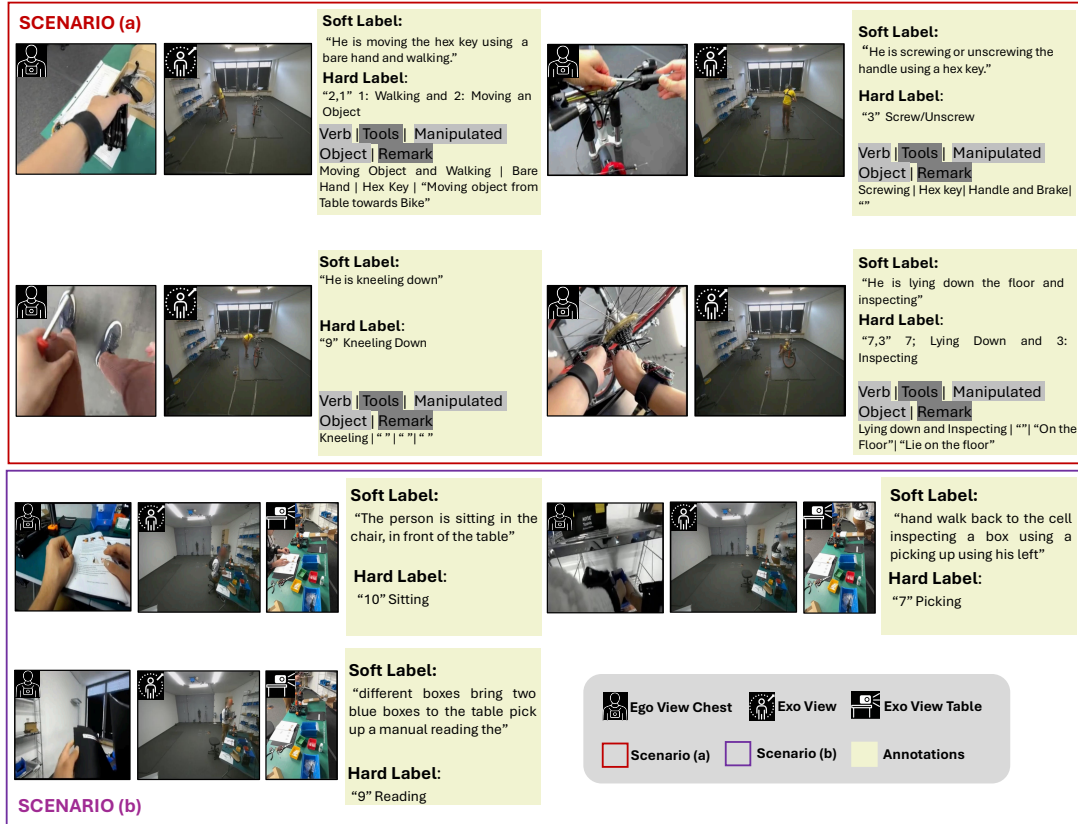


Figure 4. Egocentric and exocentric views of activity examples, accompanied by annotations (see ??). These include both soft and hard labels for two scenarios: (a) the ad-hoc bicycle assembly/disassembly task, and (b) the procedural 3D printer construction task.

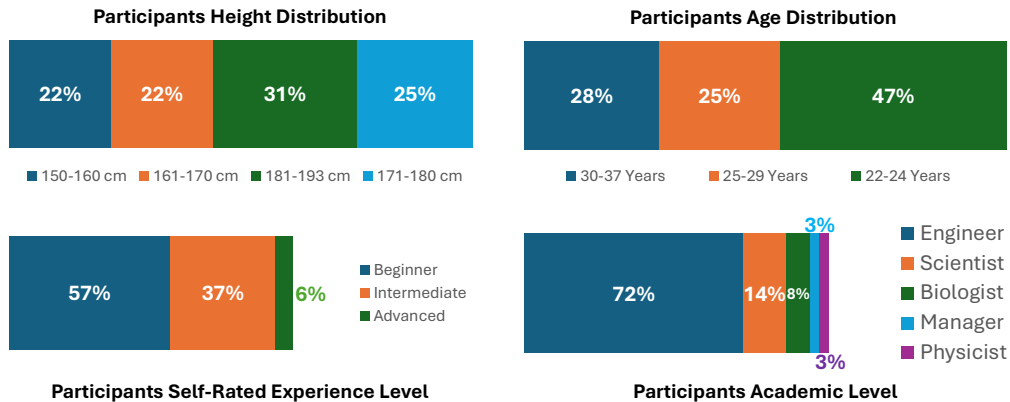


Figure 5. Participants statistic. **Top left** Height distribution ranging from 150 to 193 cm. **Top right** Age distribution spanning 22 to 37 years. **Bottom left** Self-reported experience levels in assembly tasks, categorized as beginner, intermediate, and advanced. **Bottom right** academic level of participants, with engineers representing the majority and managerial roles accounting for only 3% of the sample.

sen to reflect key challenges in industrial and task-driven environments. HAR supports downstream applications like safety monitoring, skill assessment, and robotic imitation. Open vocabulary captioning enables procedural documentation and naturalistic human-robot interaction. Cross-

modal alignment is essential for sensor substitution, retrieval, and transfer in environments where sensing modalities may be limited or asynchronous. These tasks also address known gaps in prior datasets—such as short action windows, lack of verbal grounding, and limited sen-

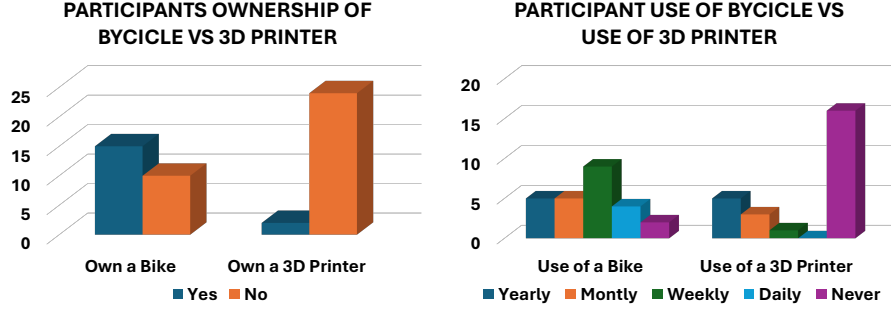


Figure 6. Participants’ ownership and usage habits related to bicycles and 3D printers.

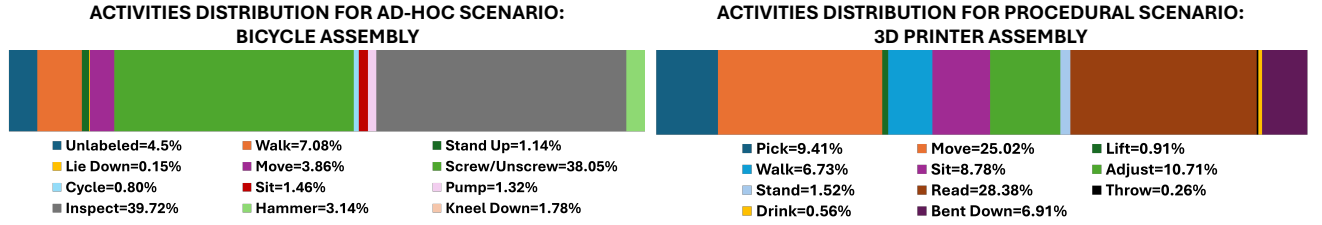


Figure 7. OpenMarcie activities distribution for the ad-hoc scenario and the procedural scenario.

Table 3. Macro F1 scores for human activity recognition in Scenario (a): Bicycle Assembly and Scenario (b): 3D Printer Assembly, with and without the null class.

Modality	Scenario (a)		Scenario (b)	
	No Null	Null	No Null	Null
Inertial(I)	0.834±0.007	0.811±0.007	0.750±0.015	0.674±0.003
Acoustic(A)	0.489±0.018	0.469±0.017	0.425±0.004	0.432±0.005
Vision(V)	0.757±0.011	0.729±0.011	0.705±0.004	0.655±0.003
I + A	0.803±0.012	0.782±0.010	0.744±0.004	0.666±0.003
A + V	0.739±0.016	0.714±0.013	0.695±0.003	0.646±0.003
I + V	0.882±0.009	0.851±0.009	0.773±0.000	0.685±0.000
I + A + V	0.859±0.010	0.831±0.011	0.763±0.003	0.676±0.003

Table 4. Open vocabulary captioning results for Scenario (a): Bicycle Assembly and Scenario (b): 3D Printer Assembly, including cosine similarity values with and without the null class.

Modality	Scenario (a)		Scenario (b)	
	No Null	Null	No Null	Null
Inertial(I)	0.518±0.023	0.501±0.022	0.642±0.002	0.640±0.002
Acoustic(A)	0.361±0.030	0.341±0.018	0.316±0.003	0.323±0.004
Vision(V)	0.479±0.016	0.463±0.014	0.632±0.002	0.631±0.003
I + A	0.512±0.021	0.493±0.020	0.644±0.002	0.641±0.003
A + V	0.466±0.025	0.444±0.012	0.626±0.003	0.625±0.004
I + V	0.561±0.016	0.531±0.014	0.655±0.000	0.655±0.000
I + A + V	0.547±0.020	0.519±0.017	0.647±0.001	0.646±0.003

sor diversity—positioning OpenMarcie as a testbed for real-world, multimodal intelligence.

Human Activity Recognition in industrial settings is challenging due to procedural, goal-driven, and concurrent activities.

Unlike prior datasets [28] with short, scripted actions,

OpenMarcie captures long, unsegmented sequences with natural variation, tool use, and diverse action paths. Its rich multimodal setup with multipositional sensors, synchronized ego/exo views, and verbal narrations offers a realistic testbed for human-centered automation. We evaluate HAR using three modalities: egocentric video, right-hand IMU, and instrumental sound. Each is tested independently with a modality-specific model, ViT [11] for video, DeepConvLSTM [26] for IMU, and EnCodec [10] with a temporal classifier for audio on a 12-class activity task with subject-disjoint splits. For multimodal evaluation, we explore all pairwise combinations and full three-way fusion using a late-fusion transformer [19] that integrates temporally encoded embeddings from each stream. This setup quantifies both the individual discriminative power and the complementarity across modalities. Tab. 3 shows Macro F1 scores across modalities for two scenarios (a, b) with and without null class inclusion. Inertial + Vision consistently outperforms all other combinations, achieving the highest Macro F1 scores. Multimodal fusion generally improves performance over unimodal inputs, especially over Acoustic alone, which performs the worst across conditions.

Open Vocabulary Captioning generates free-form descriptions of actions and scenes, crucial for industrial tasks like documentation, operator feedback, and human-machine handovers. Unlike prior datasets focused on short, generic activities with templated labels [16], OpenMarcie provides unscripted spoken narrations aligned with visual and audio streams. This enables grounded language generation in fine-grained, goal-oriented settings and allows in-

Table 5. Cross-modal alignment results for Scenario (a): Bicycle Assembly and Scenario (b): 3D Printer Assembly.

Modality	Scenario (a)			Scenario (b)		
	R@1	R@5	Top-1	R@1	R@5	Top-1
Inertial (I) + Text (T)	0.324±0.016	0.655±0.025	0.481±0.018	0.312±0.016	0.642±0.026	0.468±0.019
Acoustic (A) + T	0.241±0.014	0.583±0.025	0.342±0.016	0.227±0.013	0.567±0.022	0.329±0.015
Vision (V) + T	0.437±0.015	0.768±0.017	0.556±0.016	0.421±0.013	0.751±0.018	0.541±0.014
I + A + T	0.347±0.014	0.679±0.019	0.495±0.017	0.334±0.015	0.663±0.017	0.479±0.018
A + V + T	0.412±0.013	0.740±0.020	0.533±0.015	0.395±0.014	0.723±0.019	0.517±0.014
I + V + T	0.485±0.014	0.803±0.019	0.587±0.016	0.467±0.013	0.787±0.015	0.570±0.016
I + A + V + T	0.470±0.015	0.795±0.019	0.579±0.016	0.453±0.014	0.779±0.018	0.563±0.016

vestigation of how multimodal context, especially verbal intent, enhances captioning performance. Similar to HAR, we use modality-specific encoders to regress sentence embeddings derived from participant narrations as proposed in OV-HAR [22]. Captioning is framed as embedding prediction, with outputs decoded via Vec2Text [17] using embedding retrieval. We evaluate unimodal, pairwise, and fused setups. This approach enables efficient, open-vocabulary captioning without large language models. Tab. 4 reports cosine similarity scores for captioning, showing trends similar to HAR. Inertial + Vision achieves the best performance across both datasets, while Acoustic alone performs the worst. Adding Acoustic to other modalities provides marginal or no improvement, indicating its limited contribution compared to Inertial and Vision.

Cross Modal Alignment aims to learn a shared representation across sensory modalities, enabling retrieval, recognition, and transfer. While models like CLIP [20], Multi3Net [12] succeed on web-scale data, they lack structured, sensor-rich recordings. OpenMarcie offers a realistic testbed with synchronized video, audio, and IMU from goal-driven tasks, supporting alignment studies in complex, real-world settings beyond internet-scale benchmarks. Inspired by ImageBind [13], we use contrastive learning to align egocentric video, IMU, audio, and language into a shared embedding space. Each modality is encoded with a dedicated backbone and trained with a multi-modal InfoNCE loss over synchronized positive pairs. We evaluate all 2, 3, and 4 modality combinations, enabling fine-grained analysis of alignment and shared information across modalities. As given in Tab. 5, for both scenarios, combining inertial and vision modalities with text yields the strongest alignment performance. Vision and text perform well alone; while acoustic features contribute less individually, their performance improves when combined with other modalities. Overall, richer modality fusion improves alignment, though gains taper with full combinations. Embedding-based approaches enable efficient, scalable evaluation across tasks, highlighting OpenMarcie’s value for studying real-world, sensor-rich human activity understanding. Across all tasks, HAR, open vocabulary captioning, and cross-modal alignment, multimodal fusion consistently outperforms unimodal inputs, with inertial and

visual modalities providing the strongest signal. Audio alone provides limited information but contributes meaningfully when combined with other modalities. Moreover, because data was collected in a test-bench rather than a real factory, it lacks authentic industrial noise (e.g., machinery, vibrations). While constrained, the modality remains useful in multimodal fusion and highlights opportunities for privacy-preserving sensing under realistic conditions. Additional experiments further probing the causes of audio’s limited performance are provided in the supplementary material.

5. Conclusion

OpenMarcie is a large-scale multimodal dataset for human activity recognition, cross-modal learning, and industrial automation. It captures both ad-hoc and procedural assembly tasks with synchronized egocentric/exocentric video, wearable sensors, anonymized audio, and textual narration, enabling robust benchmarking for embodied AI. The dataset reflects realistic industrial workflows through sequential, collaborative activities involving error correction and task monitoring. While participants are predominantly right-handed engineers—limiting demographic generalization—this aligns with typical industrial workforces, enhancing ecological validity. Annotations cover only part of the dataset’s potential, as its multi-view recordings support future labeling of objects, actions, interactions, and poses. Sensor placements were ergonomically adapted per scenario but key modalities (e.g., wrist IMUs, chest-mounted LiDAR, stereo microphones) remain consistent, allowing multimodal comparisons. Overall, it provides a flexible foundation for research in fine-grained activity recognition, human–robot collaboration, and context-aware AI in industrial settings.

Acknowledgments

This work was supported by the German Federal Ministry of Education and Research (BMBF) through the Cross-Act project (Grant Agreement No. 01IW25001). This work was also supported by the IITP grant funded by the MSIT (RS-2019-II190079, AI Graduate School Program (Korea University) and IITP-2026-RS-2024-00436857, IITP-ITRC), and by the Technology Innovation Program (RS-2025-25453819) funded by the MOTIR.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3, 4
- [2] Hymalai Bello. Unimodal and multimodal sensor fusion for wearable activity recognition. In *2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pages 364–365. IEEE, 2024. 1
- [3] Hymalai Bello, Luis Alfredo Sanchez Marin, Sungho Suh, Bo Zhou, and Paul Lukowicz. Inmyface: Inertial and mechanomyography-based sensor fusion for wearable facial activity recognition. *Information Fusion*, 99:101886, 2023. 1
- [4] Hymalai Bello, Daniel Geißler, Sungho Suh, Bo Zhou, and Paul Lukowicz. Tsak: Two-stage semantic-aware knowledge distillation for efficient wearable modality and model optimization in manufacturing lines. In *International Conference on Pattern Recognition*, pages 201–216. Springer, 2025. 1
- [5] Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 847–859, 2021. 2
- [6] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–970, 2015. 1
- [7] Grazia Cicirelli, Roberto Marani, Laura Romeo, Manuel García Domínguez, Jónathan Heras, Anna G Perri, and Tiziana D’Orazio. The ha4m dataset: Multimodal monitoring of an assembly task for human action recognition in manufacturing. *Scientific Data*, 9(1):745, 2022. 2
- [8] MMAAction2 Contributors. Openmmlab’s next generation video understanding toolbox and benchmark. <https://github.com/open-mmlab/mmaaction2>, 2020. 1
- [9] Mejdí Dallel, Vincent Havard, David Baudry, and Xavier Savatier. Inhard-industrial human action recognition dataset in the context of industrial collaborative robotics. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, pages 1–6. IEEE, 2020. 1, 2
- [10] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022. 7
- [11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 7
- [12] Vitor Fortes Rey, Lala Shakti Swarup Ray, Qingxin Xia, Kaishun Wu, and Paul Lukowicz. Enhancing inertial hand based har through joint representation of language, pose and synthetic mus. In *Proceedings of the 2024 ACM International Symposium on Wearable Computers*, pages 25–31, 2024. 8
- [13] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023. 8
- [14] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 2
- [15] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 4
- [16] Huan Liu, Guangbin Wang, Ting Huang, Ping He, Martin Skitmore, and Xiaochun Luo. Manifesting construction activity scenes via image captioning. *Automation in Construction*, 119:103334, 2020. 7
- [17] John X Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander M Rush. Text embeddings reveal (almost) as much as text. *arXiv preprint arXiv:2310.06816*, 2023. 8
- [18] Friedrich Niemann, Christopher Reining, Fernando Moya Rueda, Nilah Ravi Nair, Janine Anika Steffens, Gernot A Fink, and Michael Ten Hompel. Lara: Creating a dataset for human activity recognition in logistics using semantic attributes. *Sensors*, 20(15):4083, 2020. 2
- [19] Yagya Raj Pandeya and Joonwhoan Lee. Deep learning-based late fusion of multimodal information for emotion classification of music video. *Multimedia Tools and Applications*, 80(2):2887–2905, 2021. 7
- [20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. Pmlr, 2021. 8
- [21] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. 4
- [22] Lala Shakti Swarup Ray, Bo Zhou, Sungho Suh, and Paul Lukowicz. Initial findings on sensor based open vocabulary activity recognition via text embedding inversion. In *2025 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pages 685–688. IEEE, 2025. 8
- [23] Karen L Schmidt and Jeffrey F Cohn. Human facial expressions as adaptations: Evolutionary questions in facial expression research. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, 116(S33):3–24, 2001. 1

- [24] Tim J Schoonbeek, Tim Houben, Hans Onvlee, Fons Van der Sommen, et al. Industreal: A dataset for procedure step recognition handling execution errors in egocentric videos in an industrial-like setting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4365–4374, 2024. [2](#)
- [25] Fadime Sener, Dibyadip Chatterjee, Daniel Sheleпов, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022. [2](#)
- [26] Satya P Singh, Madan Kumar Sharma, Aimé Lay-Ekuakille, Deepak Gangwar, and Sukrit Gupta. Deep convlstm with self-attention for human activity decoding using wearable sensors. *IEEE Sensors Journal*, 21(6):8575–8582, 2020. [7](#)
- [27] Xiao Wang, Zongzhen Wu, Bo Jiang, Zhimin Bao, Lin Zhu, Guoqi Li, Yaowei Wang, and Yonghong Tian. Hardvs: Re-visiting human activity recognition with dynamic vision sensors. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5615–5623, 2024. [1](#)
- [28] Naoya Yoshimura, Jaime Morales, Takuya Maekawa, and Takahiro Hara. Openpack: A large-scale dataset for recognizing packaging works in iot-enabled logistic environments. In *2024 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pages 90–97. IEEE, 2024. [1](#), [2](#), [7](#)
- [29] Hao Zheng, Regina Lee, and Yuqian Lu. Ha-vid: A human assembly video dataset for comprehensive assembly knowledge understanding. *Advances in Neural Information Processing Systems*, 36:67069–67081, 2023. [1](#), [2](#)