OPENMARCIE: DATASET FOR MULTIMODAL ACTION RECOGNITION IN INDUSTRIAL ENVIRONMENTS

Anonymous authors
Paper under double-blind review

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 integration 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 (Caba Heilbron et al., 2015; Contributors, 2020), including event cameras (Wang et al., 2024). Table 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 (Schmidt & Cohn, 2001). Variability in motion, context dependence, and environmental interactions add further complexity, making multimodal analysis essential for accurate interpretation (Bello et al., 2023; Bello, 2024). 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 (Yoshimura et al., 2024; Dallel et al., 2020; Zheng et al., 2023). This is especially true in industrial environments, where vision-only-based methods

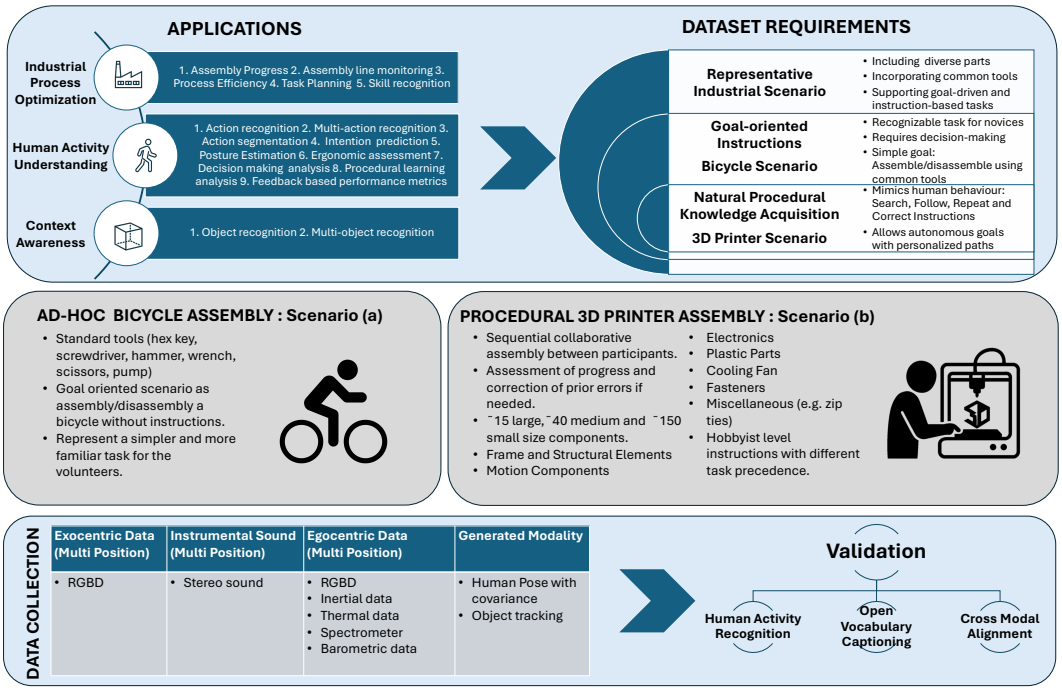

Figure 1: OpenMarcie is an open dataset for multimodal action recognition in industrial environments. **Top** Its applications include industrial process optimization, human activity understanding, and context awareness. **Middle** two assembly scenarios represent ad-hoc goal-oriented action and a procedural 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.

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 (Dallel et al., 2020)	✓	✓	x	✓	✓	x	HRC assembly, RGB+IMU
LARa (Niemann et al., 2020)	✓	✓	x	✓	x	x	Logistics, IMU+mocap
OpenPack (Yoshimura et al., 2024)	✓	✓	x	✓	x	x	Logistics, 50h+ IMU/IoT
Assembly101 (Sener et al., 2022)	✓	x	✓	✓	✓	x	Vision-only, procedural assembly
IKEA-ASM (Ben-Shabat et al., 2021)	✓	x	x	✓	✓	x	Exo RGB-D, furniture assembly
HA4M (Cicirelli et al., 2022)	✓	x	x	✓	x	x	Exo RGB-D+IR+pose
HA-ViD (Zheng et al., 2023)	✓	x	x	✓	✓	x	Rich semantic labels, multi-route
IndustReal (Schoonbeek et al., 2024)	✓	x	✓	x	x	x	PSR, ego-only, error modeling
Ego-Exo4D (Grauman et al., 2024)	**(~6% ind.)**	✓	✓	✓	✓	x	1,200h skilled activities
OpenMarcie (ours)	**(100%)**✓	✓	✓	✓	✓	✓	Industrial multitasking, 8 modalities

may raise privacy concerns and increase the risk of technology leakage (Bello et al., 2025). Accordingly, we present OpenMarcie (see Figure 1). Our contributions are threefold. First, we introduce OpenMarcie, a multimodal dataset with eight sensing 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

This section presents the process of building OpenMarcie and provides essential statistics. 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 Table 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.

2.1 Ad-hoc Scenario (a): Bicycle disassembly/assembly

The assembly and disassembly of a bicycle is proposed as structured, goal-oriented tasks that allow 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 temporally aligned across all modalities (egocentric video, inertial, thermal, audio). We additionally generate concise natural-language descriptions with GPT-4o (Achiam et al., 2023), providing soft labels that improve consistency and interpretability while preserving human-created ground truth.

2.2 Procedural Scenario (b): 3D Printer Assembly/Disassembly

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 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=`

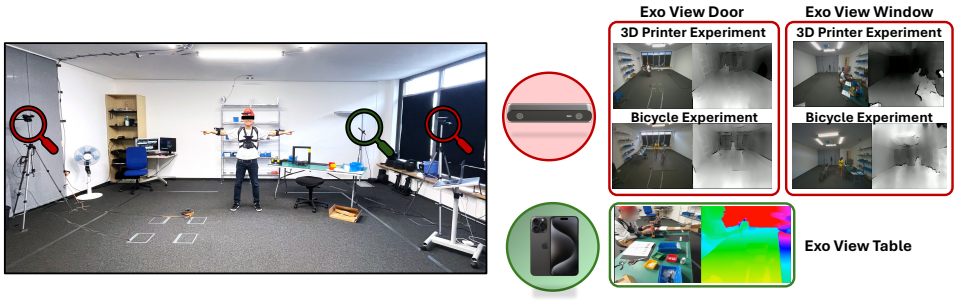

Figure 2: Experiment room setting with example views of the exocentric RGBD cameras.

Table 2: Sensing modalities information for each scenario

Scenario	Modality	Position	Total Channels
Ad-Hoc **(a)**: Bicycle Assembly	IMU	Wrists and Forehead	21
	Magnetometer	Wrists and Forehead	9
	Barometer	Wrists and Head	3
	Temperature	Wrists and Forehead	3
	Spectrometer	Chest and Right Shoulder	24
	Thermal Camera	Chest and Right Shoulder	(8x8)x2
	RGB-Lidar	Chest	4
	Stereo Sound	Chest	2
	RGBD	2 Exocentric cameras	8
Procedural **(b)**: 3D Printer Assembly	IMU	Wrists	20
	Magnetometer	Wrists	6
	Barometer	Wrists	2
	Temperature	Wrists, Forehead and Chest	4
	RGBD-IMU	Forehead and Chest	28
	RGB-LiDar	Chest and Table	8
	Stereo Sound	Chest and Table	4
	RGBD	2 Exocentric cameras	8
Both Scenarios	17	9	282

`uToqSlh64R4`. 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 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 on both wrists, and one ZED on the forehead by using a helmet, and another one on the chest by the use of a strap. 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 (Radford et al., 2022). 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 (Guo et al., 2025), then using GPT-4o (Achiam et al., 2023) 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

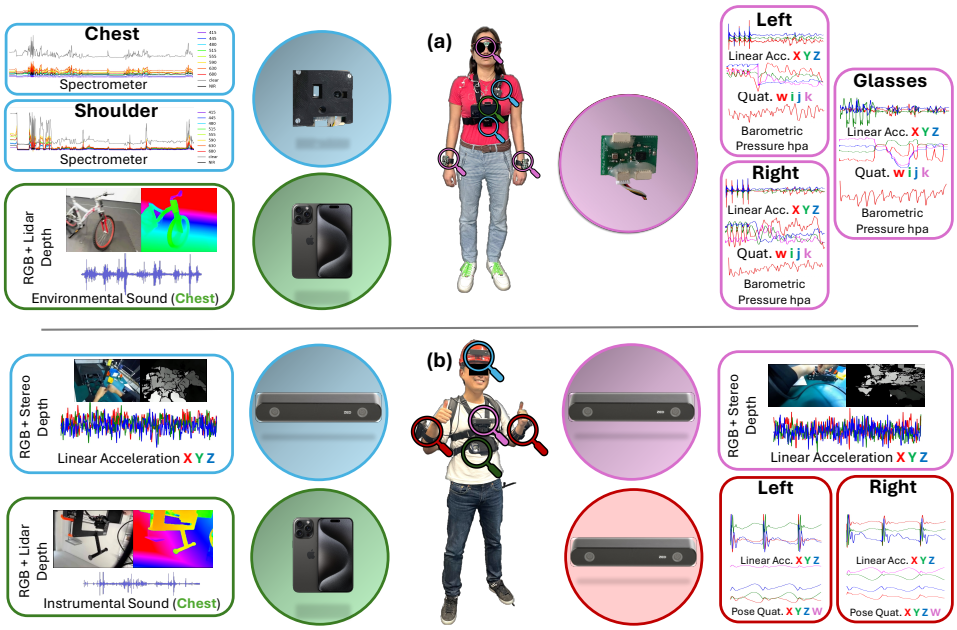

Figure 3: Participant wearable setup and example sensor signals. **(a)**Participant in the ad-hoc Scenario (a): bicycle assembly. **(b)** Participant in the procedural Scenario (b): 3D printer assembly. Sensor placements were adapted to suit the specific ergonomics and task demands of each scenario.

labels, we perform bidirectional consistency checks (structured→caption→structured), achieving strong alignment (Macro F1 = 0.715 in Scenario a; METEOR = 0.531 in Scenario b; see supplementary). 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 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.

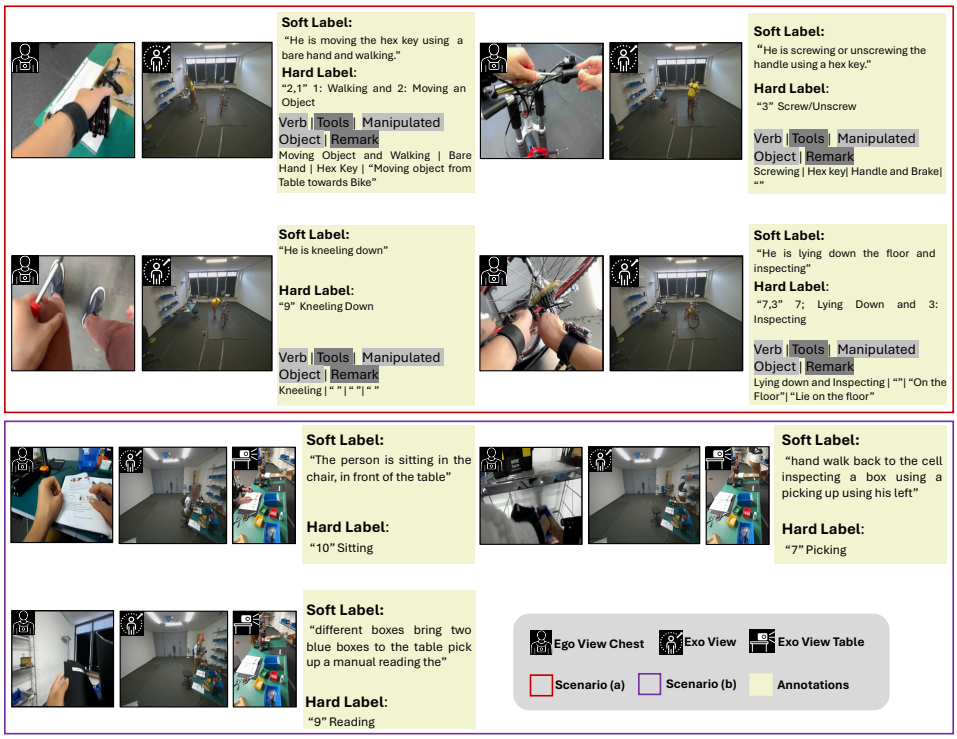

Figure 4: Egocentric and exocentric views of activity examples, accompanied by annotations. These include both soft and hard labels for two scenarios: **(a)** the ad-hoc bicycle assembly/disassembly task, and **(b)** the procedural 3D printer construction task.

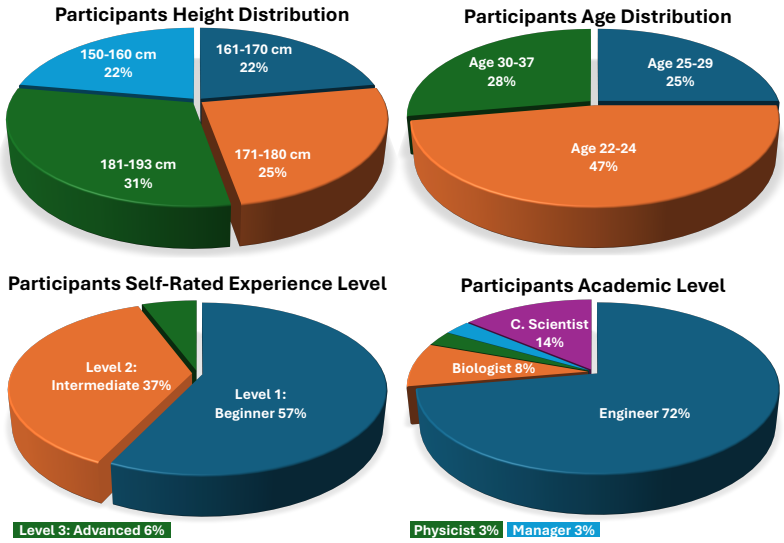

Figure 5: Participants statistic. **Top left** Height distribution ranging from 150 to 193 cm. **Top right** Age distribution spanning 22 to 37 years. **Bottom left** Self-reported experience levels in assembly tasks, categorized as beginner, intermediate, and advanced. **Bottom right** academic level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

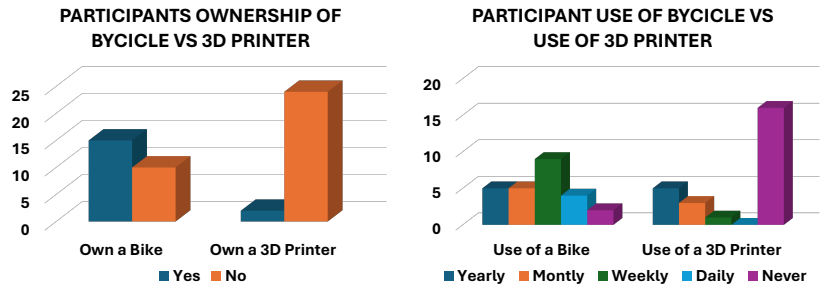

Figure 6: Participants' ownership and usage habits related to bicycles and 3D printers.

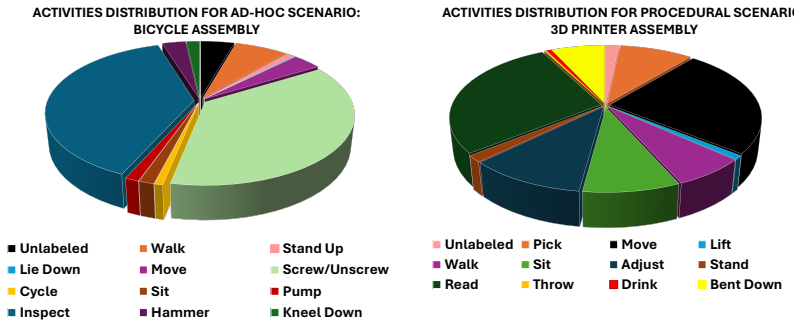

Figure 7: OpenMarcie activities distribution for the ad-hoc scenario and the procedural scenario.

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 assembly 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 chosen 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-

Table 3: Human activity recognition results for Scenario (a): Bicycle Assembly and Scenario (b): 3D Printer Assembly, including macro F1 scores with and without the null class.

Modality	**Scenario (a)**		**Scenario (b)**	
	Macro F1 (↑)			
	Without Null	With Null	Without Null	With Null
Inertial	0.834±0.007	0.811±0.007	0.750±0.015	0.674±0.003
Acoustic	0.489±0.018	0.469±0.017	0.425±0.004	0.432±0.005
Vision	0.757±0.011	0.729±0.011	0.705±0.004	0.655±0.003
Inertial + Acoustic	0.803±0.012	0.782±0.010	0.744±0.004	0.666±0.003
Acoustic + Vision	0.739±0.016	0.714±0.013	0.695±0.003	0.646±0.003
Inertial + Vision	**0.882±0.009**	**0.851±0.009**	**0.773±0.000**	**0.685±0.000**
Inertial + Acoustic + Vision	0.859±0.010	0.831±0.011	0.763±0.003	0.676±0.003

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

Human Activity Recognition HAR in industrial settings is challenging due to procedural, goal-driven, and concurrent activities. Unlike prior datasets (Yoshimura et al., 2024) 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 (Dosovitskiy et al., 2020) for video, DeepConvLSTM (Singh et al., 2020) for IMU, and EnCodec (Défossez et al., 2022) 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 (Pandeya & Lee, 2021) that integrates temporally encoded embeddings from each stream. This setup quantifies both the individual discriminative power and the complementarity across modalities. Table 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 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 (Liu et al., 2020), OpenMarcie provides unscripted spoken narrations aligned with visual and audio streams. This enables grounded language generation in fine-grained, goal-oriented settings and allows investigation 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 (Ray et al., 2025). Captioning is framed as embedding prediction, with outputs decoded via Vec2Text (Morris et al., 2023) using embedding retrieval. We evaluate unimodal, pairwise, and fused setups. This approach enables efficient, open-vocabulary captioning without large language models. Table 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 Cross-modal alignment aims to learn a shared representation across sensory modalities, enabling retrieval, recognition, and transfer. While models like CLIP (Radford et al., 2021), Multi³Net (Fortes Rey et al., 2024) 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 (Girdhar et al., 2023), we use contrastive learning to align egocentric video, IMU, audio, and language into a shared embedding space. Each modal-

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)**	
	Cosine Similarity (↑)			
	Without Null	With Null	Without Null	With Null
Inertial	0.518±0.023	0.501±0.022	0.642±0.002	0.640±0.002
Acoustic	0.361±0.030	0.341±0.018	0.316±0.003	0.323±0.004
Vision	0.479±0.016	0.463±0.014	0.632±0.002	0.631±0.003
Inertial + Acoustic	0.512±0.021	0.493±0.020	0.644±0.002	0.641±0.003
Acoustic + Vision	0.466±0.025	0.444±0.012	0.626±0.003	0.625±0.004
Inertial + Vision	**0.561±0.016**	**0.531±0.014**	**0.655±0.000**	**0.655±0.000**
Inertial + Acoustic + Vision	0.547±0.020	0.519±0.017	0.647±0.001	0.646±0.003

Table 5: Cross-modal alignment results for Scenario (a): Bicycle Assembly and Scenario (b): 3D Printer Assembly, including recall@5, recall@1, and top-1 metrics.

Modality	**Scenario (a)**			**Scenario (b)**		
	Recall@1 (↑)	**Recall@5 (↑)**	**Top-1 (↑)**	**Recall@1 (↑)**	**Recall@5 (↑)**	**Top-1 (↑)**
Inertial + Text	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 + Text	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 + Text	0.437±0.015	0.768±0.017	0.556±0.016	0.421±0.013	0.751±0.018	0.541±0.014
Inertial + Acoustic + Text	0.347±0.014	0.679±0.019	0.495±0.017	0.334±0.015	0.663±0.017	0.479±0.018
Acoustic + Vision + Text	0.412±0.013	0.740±0.020	0.533±0.015	0.395±0.014	0.723±0.019	0.517±0.014
Inertial + Vision + Text	**0.485±0.014**	**0.803±0.019**	**0.587±0.016**	**0.467±0.013**	**0.787±0.015**	**0.570±0.016**
Inertial + Acoustic + Vision + Text	0.470±0.015	0.795±0.019	0.579±0.016	0.453±0.014	0.779±0.018	0.563±0.016

ity 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 Table 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.

6 ETHICS STATEMENT

This work complies with the Declaration of Helsinki. All participants provided informed consent and retained the right to withdraw at any time. To mitigate privacy risks, faces were blurred, speech was excluded, and only anonymized instrumental audio was released. External narrations were provided by an observer, transcribed, and used in place of participants' private speech. Comprehensive demographic metadata is included to support fairness and bias analysis. Compensation was limited to voluntary vouchers in Scenario (b).

7 REPRODUCIBILITY STATEMENT

We release the OpenMarcie dataset, all annotations, and benchmark splits to ensure reproducibility. The benchmark baselines use publicly available architectures and libraries, with hyperparameters, training protocols, and evaluation metrics fully described in the main text and supplementary material. Extended audio-only experiments, model configurations, and preprocessing details are also documented in the supplementary for transparency. Together, these resources allow other researchers to replicate our experiments and build upon our benchmarks. As ICLR explicitly lists datasets and benchmarks as a core subject area, our contribution is aligned with this scope, providing both reproducible baselines and a foundation for future research in multimodal representation learning and industrial AI.

REFERENCES

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.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

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)*, pp. 364–365. IEEE, 2024.

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.

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*, pp. 201–216. Springer, 2025.

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*, pp. 847–859, 2021.

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*, pp. 961–970, 2015.

Grazia Cicirelli, Roberto Marani, Laura Romeo, Manuel García Domínguez, Jónathan Heras, Anna G Perri, and Tiziana D'Orazio. The ha4m dataset: Multi-modal monitoring of an assembly task for human action recognition in manufacturing. *Scientific Data*, 9(1):745, 2022.

MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. `https://github.com/open-mmlab/mmaction2`, 2020.

Mejdi 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)*, pp. 1–6. IEEE, 2020.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

Blackmagic Design. Davinci resolve. https://www.blackmagicdesign.com/products/davinciresolve/. Accessed: 2025-05-08.

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.

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 imus. In *Proceedings of the 2024 ACM International Symposium on Wearable Computers*, pp. 25–31, 2024.

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*, pp. 15180–15190, 2023.

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*, pp. 19383–19400, 2024.

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.

Intel. Openvino™ ai plugins for audacity: Music separation. https://github.com/intel/openvino-plugins-ai-audacity. Accessed: 2025-05-08.

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.

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.

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.

ORB-HD. deface. https://github.com/ORB-HD/deface. Accessed: 2025-05-08.

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.

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*, pp. 8748–8763. PmLR, 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

Lala Shakti Swarup Ray, Bo Zhou, Sungho Suh, and Paul Lukowicz. Initial findings on sensor based open vocabulary activity recognition via text embedding inversion. *arXiv preprint arXiv:2501.07408*, 2025.

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.

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*, pp. 4365–4374, 2024.

Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, 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*, pp. 21096–21106, 2022.

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.

Xiao Wang, Zongzhen Wu, Bo Jiang, Zhimin Bao, Lin Zhu, Guoqi Li, Yaowei Wang, and Yonghong Tian. Hardvs: Revisiting human activity recognition with dynamic vision sensors. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 5615–5623, 2024.

Yuanyuan Xu, Wan Yan, Haixin Sun, Genke Yang, and Jiliang Luo. Centerface: Joint face detection and alignment using face as point. In *arXiv:1911.03599*, 2019.

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)*, pp. 90–97. IEEE, 2024.

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.

SUPPLEMENTAL MATERIAL FOR OPENMARCIE: DATASET FOR MULTIMODAL ACTION RECOGNITION IN INDUSTRIAL ENVIRONMENTS

This supplementary document contains additional information about OpenMarcie. Section A compares in detail the prior datasets with OpenMarcie. Section B presents the metadata information about each participant independently. Section C declares the ethical consideration and societal impact of the OpenMarcie dataset with emphasis in user privacy and anonymization. Section D describes how a large language model semantically transforms human-written descriptions and vice-versa, serving as a structured translator rather than a primary labeling agent. Section E shows the object distribution for both experimental scenarios and illustrates example scenes with object segmentation and human skeleton poses. Section F details the implementation of the proposed benchmarks and presents the evaluation metrics used. Section G reports additional audio-only experiments designed to probe the modality's limitations under different conditions. Specifically, we (1) evaluate the full pipeline with pre-anonymized (raw) audio streams to isolate the effect of the anonymization process, and (2) replace Encodec embeddings with a non–ML-based mel-spectrogram classifier to disentangle the influence of the embedding method from the content itself. Section H outlines several research directions that extend beyond the benchmarks presented in this work.

A DETAILED COMPARISON TO PRIOR DATASETS

To provide a clearer perspective, we include a narrative comparison of each cited dataset, highlighting how their design choices converge with or diverge from OpenMarcie in terms of industrial fidelity, sensing breadth, and suitability for advanced research tasks.

- InHARD Dallel et al. (2020), developed for human–robot collaboration, integrates three exocentric RGB views with wearable inertial mocap, making it one of the few industrial datasets to include wearable sensing. Its scope, however, is limited to single-action recognition in fixed exocentric settings. In contrast, OpenMarcie extends beyond this by pairing wearables with synchronized egocentric and exocentric video and by providing multi-action labels, thereby capturing overlapping activities that more faithfully reflect real-world industrial workflows.
- LARa Niemann et al. (2020) captures logistics activities using IMUs, optical mocap, and a single RGB camera, providing valuable insight into worker variability in warehouse settings. Its scope, however, is primarily exocentric and lacks the multimodal depth of OpenMarcie. In the absence of egocentric streams or concurrent multi-action labels, LARa is best suited for controlled HAR scenarios. OpenMarcie advances this space by integrating egocentric video, wearable sensing, and multi-route task design, thereby enabling richer supervision for collaborative and multitasking contexts.
- OpenPack Yoshimura et al. (2024) is a large-scale logistics dataset comprising 53.8 hours of recordings that integrate IMUs, 2D keypoints, depth, and IoT data. Its primary strength lies in combining wearable sensing with process-level logging, though its perspective remains exclusively exocentric. OpenMarcie addresses these gaps by providing egocentric multimodal video aligned with exocentric multiviews and by introducing explicit multi-action annotation. This design enables models not only to classify activities but also to reason about simultaneity and task intent.
- Assembly101 Sener et al. (2022) offers multi-view egocentric and exocentric video with dense procedural annotations, establishing a strong benchmark for vision-based activity understanding. However, it does not incorporate wearable sensing or overlapping activity labels. OpenMarcie extends this direction by retaining ego–exo coverage while adding wearable IMUs, audio, thermal, and spectrometer signals, and by explicitly modeling concurrent actions. These additions make OpenMarcie particularly well-suited for advancing cross-modal learning.
- IKEA-ASM Ben-Shabat et al. (2021) is a furniture assembly dataset that provides RGB-D multiview recordings with annotations for atomic actions and manipulated objects. Its main strength is the inclusion of depth and pose data from multiple exocentric Kinect sensors, but it lacks egocentric perspectives and wearable sensing. OpenMarcie complements this

resource by combining egocentric video with wearables and additional non-visual sensors, thereby supporting cross-modal transfer beyond RGB-D alone.

- HA4M Cicirelli et al. (2022) focuses on gear-train assembly and provides six modalities captured with Azure Kinect, including RGB, depth, IR, and skeleton data. While it emphasizes vision-rich multimodality, it is restricted to exocentric viewpoints and does not include wearable sensing. OpenMarcie advances this space by integrating egocentric and exocentric cameras with wearables and multi-action labels, thereby supporting more realistic multitasking scenarios in industrial workflows.
- HA-ViD Zheng et al. (2023) provides fine-grained multi-view assembly recordings with detailed annotations covering subjects, verbs, objects, tools, and collaboration cues. It offers rich semantic supervision and supports alternative task routes. However, it does not include wearable sensing or egocentric perspectives. OpenMarcie complements HA-ViD by incorporating egocentric video, wearable streams, and explicit concurrent multi-action labels, thereby extending applicability to multitasking and multimodal alignment.
- IndustReal Schoonbeek et al. (2024) is designed for Procedure Step Recognition (PSR), using egocentric video to capture procedural errors and flexible subgoals. Its distinctive contribution is its focus on error modeling, though it remains limited to ego-only recordings without wearable sensing or multiview exocentric support. OpenMarcie, while currently benchmarked on HAR, captioning, and cross-modal alignment, integrates egocentric and exocentric video with wearables and multi-action labels, providing a complementary foundation for PSR and error modeling—both of which are included in our staged roadmap.
- Ego-Exo4D Grauman et al. (2024) is a large-scale dataset comprising over 1,200 hours of skilled activities, captured with synchronized egocentric and exocentric video, audio, IMU, gaze, and language streams. It is unmatched in scale and modality breadth, yet only a small portion of the recordings are industrial. OpenMarcie takes a complementary approach by focusing exclusively on industrial workflows, augmenting them with multi-action annotations and specialized wearables. In this way, OpenMarcie serves as a domain-focused counterpart to Ego-Exo4D, prioritizing depth over scale in factory settings.

OpenMarcie is the only dataset to jointly provide wearables, egocentric and exocentric multiview recordings, explicit concurrent multi-action labels, and complete industrial coverage (see Table 1). While prior datasets emphasize individual strengths—such as rich semantic annotations (HA-ViD), large scale (Ego-Exo4D), wearable integration (InHARD, LARa, OpenPack), or multi-view video (Assembly101, IKEA-ASM)—none combine all of these elements within an industrial setting. This unique positioning makes OpenMarcie particularly well-suited for the advanced benchmarks outlined in our roadmap, including procedural planning, skill assessment, intent prediction, fine-grained segmentation, pose reasoning, cross-modal transfer, and cross-modal generation.

B User Metadata

Based on Figure 8, Table 6, and Table 7, the participant data across the two scenarios—bicycle assembly (Scenario (a)) and 3D printer assembly (Scenario (b))—reflects a diverse and well-balanced cohort in terms of demographics and professional backgrounds. Scenario (a) includes 12 participants, while Scenario (b) features 24, offering a broader basis for analysis. Males form the majority in both groups, although Scenario (b) exhibits greater gender diversity, with a higher number of female participants.

The majority of participants are right-handed, with only four identifying as left-handed, as shown in the summarized visualization, indicating a strong dominance of right-handed individuals across the dataset. Participants range in age from their early 20s to late 30s, and most hold academic degrees in engineering. Additional disciplines represented include computer science, biology, physics, and management, demonstrating multidisciplinary relevance.

Geographically, the dataset includes participants from multiple continents—South America, Asia, Europe, Africa, and North America—highlighting broad international representation. Scenario (b) shows particularly rich demographic variety, with over 15 distinct national origins. Experience levels, self-reported on a scale from 1 to 3, vary across participants, with most indicating beginner to intermediate familiarity with the task domain.

Overall, the metadata illustrates the inclusiveness and diversity of the participant pool, supporting the dataset's utility for developing generalizable models in embodied AI, human-robot interaction, and vision-based behavior analysis.

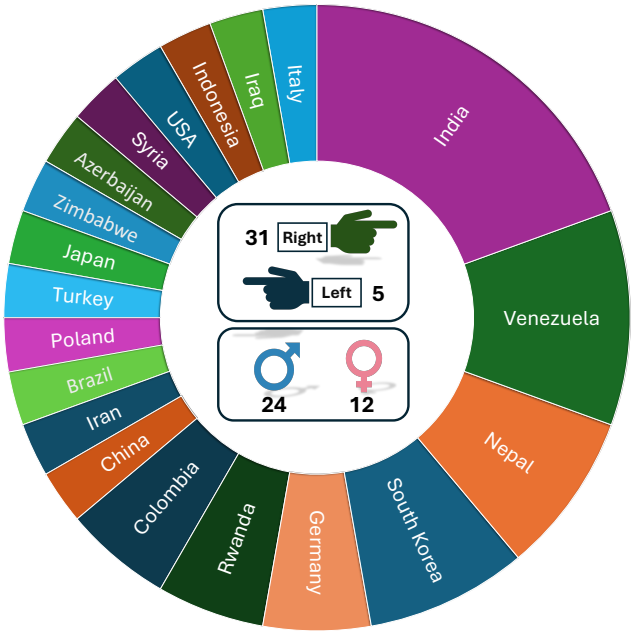

Figure 8: Distribution of participants by nationality, dominant hand and self identified sex.

Table 6: Participants' metadata information for the Ad-hoc Scenario (a).

ID	Sex	Age	Height	Dominant Hand	Academic Level	Demographic	Experience Level (1-3)
P1-B	F	33	160	R	Engineer	South America (Venezuela)	2
P2-B	F	26	176	R	Engineer	Europe (Poland)	3
P3-B	M	29	175	R	Computer Scientist	Europe (Germany)	1
P4-B	M	26	175	L	Engineer	South Asia (India)	1
P5-B	M	33	189	R	Computer Scientist	South America (Brazil)	1
P6-B	M	36	188	R	Engineer	East Africa (Rwanda)	1
P7-B	M	25	168	R	Engineer	South Asia (India)	1
P8-B	M	37	178	R	Engineer	East Asia (South Korea)	2
P9-B	M	27	176	R	Engineer	Middle East (Iran)	2
P10-B	M	27	174	R	Engineer	South Asia(India)	1
P11-B	M	34	193	R	Engineer	South America (Venezuela)	1
P12-B	M	30	160	R	Engineer	South East Asia (China)	1

Figure 9 presents a comparative analysis between range of motion (Left) and motion diversity (Right) across two distinct assembly scenarios: Scenario (a), Bicycle assembly, and Scenario (b), 3D Printer assembly. Each metric is broken down by body region: Full Body, Left Hand, Right Hand, and Lower Body. Scenario (b) exhibits a greater range of motion compared to Scenario (a), particularly for the full body as well as the left and right hands. This increased movement may be attributed to the multiple instances of searching around the table, shelf, and floor observed in Scenario (b). Scenario (a) involves kneeling, lying on the ground, and standing up to inspect specific areas of the bicycle. Consequently, it demonstrates a greater lower body range of motion and movement diversity compared to the 3D printer scenario (Scenario b).

Table 8 presents a list of acronyms used to represent various activities and objects involved in Scenario (a): Bicycle assembly. The activities are denoted by single-letter acronyms such as W for Walking, M for Move (manipulating an object before the next action), U for Screw/Unscrew, and C for Cycling. Other physical actions include S for Sitting down, P for Pumping air into the tires, I for Inspecting with hands or eyes, H for Hammering, K for Kneeling down, A for Standing up, L for Lying down, and T for Cutting. Additionally, object-related acronyms are listed, including hx

Table 7: Participants metadata information for the procedural Scenario (b).

ID	Gender	Age	Height	Dominant Hand	Academic Level	Demographic	Experience Level (1-3)
P1-D	M	37	178	R	Engineer	East Asia (South Korea)	2
P2-D	F	25	159	R	Engineer	South East Asia Pacific (Nepal)	2
P3-D	M	34	193	R	Engineer	South America (Venezuela)	1
P4-D	F	26	165	R	Engineer	Europe (Italy)	1
P5-D	F	33	160	R	Engineer	South America (Venezuela)	2
P6-D	F	25	164	R	Engineer	South Asia (India)	1
P7-D	M	24	175	R	Engineer	Middle East (Iraq)	3
P8-D	M	25	168	R	Engineer	South Asia (India)	1
P9-D	F	25	162	R	Physicist	South America (Colombia)	2
P10-D	F	27	155	R	Engineer	South East Asia (Indonesia)	2
P11-D	F	24	160	R	Engineer	South Asia (India)	1
P12-D	M	36	188	R	Engineer	East Africa (Rwanda)	1
P13-D	M	24	168	R	Biologist	North America (USA)	2
P14-D	M	22	187	R	Management	Middle East (Syria)	2
P15-D	F	24	150	R	Engineer	Caucasus Asia (Azerbaijan)	1
P16-D	M	25	165	R	Engineer	South Africa (Zimbabwe)	1
P17-D	M	22	167	R	Engineer	South East Asia Pacific (Nepal)	2
P18-D	M	29	175	L	Computer Scientist	Europe (Germany)	1
P19-D	F	23	161	R	Biologist	East Asia (Japan)	1
P20-D	M	26	175	L	Engineer	South Asia (India)	1
P21-D	M	27	182	R	Computer Scientist	East Asia (South Korea)	2
P22-D	M	24	183	L	Engineer	South East Asia Pacific (Nepal)	2
P23-D	F	24	156	R	Microbiologist	South America (Colombia)	1
P24-D	M	25	178	L	Computer Scientist	Middle East (Turkey)	1

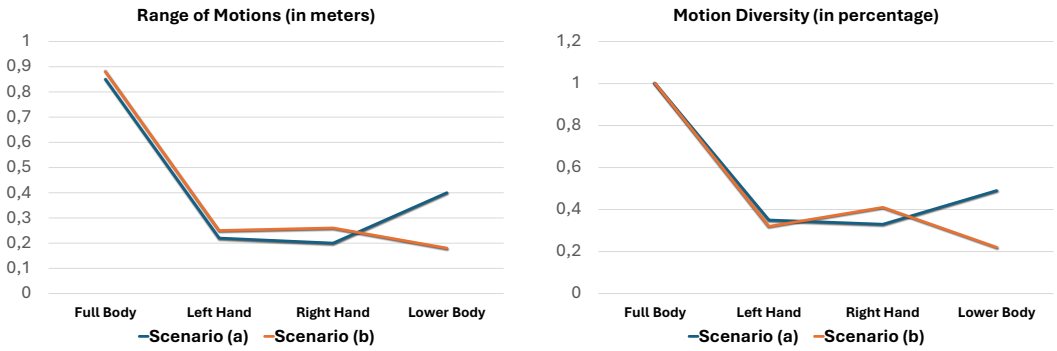

Figure 9: Comparison of range and diversity of body motions across Scenario (a): bicycle assembly and Scenario (b): 3D printer assembly.

for Hex key, wr for Wrench, sd for Screwdriver, hm for Hammer, sc for Scissors, pl for Plier, pu for Pump, and bh for Bare Hands. This table helps clarify shorthand notations used to describe the detailed steps and tools involved in the bicycle assembly process.

C Ethical Consideration and Societal Impact

The OpenMarcie dataset was developed with a strong emphasis on ethical research practices and societal responsibility. All participants provided informed consent in compliance with the Declaration of Helsinki. Participants were informed about the nature of data being collected—including egocentric video, wearable sensor data, and external audio narration, and retained the right to withdraw at any time. Participants received a 15-euro Amazon voucher for voluntary participation in Scenario (b). No compensation was provided for participation in Scenario (a).

To mitigate privacy risks, egocentric recordings were deliberately framed to capture task execution while minimizing exposure of participant identities. Verbal narrations were provided by an external observer to avoid capturing participants' private speech. These narrations were then transcribed using faster-whisper (v1.1.1) and ctranslate2 (v4.4.0), employing the "large-v3" model configuration Radford et al. (2022), and only the resulting text descriptions were used in the benchmark tasks and released in OpenMarcie.

Table 8: Activities and objects acronyms for the Scenario (a): Bicycle assembly.

Acronym	Meaning
W	Walking
M	Move: Manipulating an object until the next action
U	Screw/Unscrew
C	Cycling
S	Sitting down
P	Pump: Pumping air into the bicycle tires
I	Inspect: Inspecting an object with hand/eyes
H	Hammering
K	Kneeling down
A	Standing up
L	Lie Down
T	Cutting
hx	Hex key
wr	Wrench
sd	Screwdriver
hm	Hammer
sc	Scissors
pl	Plier
pu	Pump
bh	Bare Hands

Facial features and biometric identifiers were excluded from all labeling and analysis processes. A two-stage anonymization procedure was applied to the video data. First, participant faces were automatically detected and blurred using the open-source deface Python package (threshold set at 0.7) Xu et al. (2019); ORB-HD. Second, all videos were manually reviewed, and any remaining visible facial features were fully obscured using DaVinci Resolve Design. For audio, only the instrumental components were released. Voices were intentionally removed from the audio tracks using the OpenVINO Music Separation plugin in Audacity Intel, which separates recordings into vocal and instrumental stems. OpenMarcie includes only the instrumental tracks.

OpenMarcie is designed not only to advance research in human activity recognition but also to support transparency and fairness in machine learning. The dataset includes comprehensive metadata on participants' demographics, academic background, dominant hand, and self-assessed skill levels, enabling researchers to perform subgroup analyses and audit for potential biases. This supports the development of equitable multimodal systems in industrial and embodied AI applications.

From a broader societal perspective, OpenMarcie has the potential to benefit applications such as human-robot collaboration, workplace safety, ergonomic assessment, and adaptive training systems. Its real-world, goal-driven scenarios are ideal for advancing models that interpret human actions in complex environments. However, such systems may also carry risks if misused—for example, for excessive surveillance or performance monitoring without consent.

We encourage researchers using OpenMarcie to explicitly assess fairness, document performance across diverse subgroups, and consider the downstream implications of deploying human action recognition systems in human-centered settings. OpenMarcie aims to support the responsible development of AI by providing a rich yet ethically grounded testbed for real-world multimodal learning.

D LLM-BASED ANNOTATION TRANSLATION

D.1 LLM GENERATED LABEL VALIDATION

Our pipeline employs GPT-4o to translate human-authored soft activity descriptions into standardized formats—either discrete activity classes (hard labels) or continuous representations (soft labels). Importantly, the LLM is not used to generate annotations directly from visual input, but rather to semantically convert existing human-written annotations. Thus, it acts as a structured translator rather than a primary labeling agent.

Crucially, this translation process is not fully automated. Human oversight is incorporated throughout, particularly during the mapping of soft-label sentences to discrete classes. Annotators refine LLM prompts, inspect outputs for ambiguous cases, and resolve semantic inconsistencies. This human-in-the-loop approach helps ensure the accuracy and consistency of final labels used for model training.

To validate the quality of LLM-generated annotations, we perform a bidirectional consistency analysis under the two scenarios:

- **Ad hoc scenario (Hard labels → Captions → Hard labels):** Human-annotated discrete classes are transformed by the LLM into natural-language captions, then back-translated by the LLM into discrete labels.
- **Procedural scenario (Captions → Hard labels → Captions):** Human written soft label sentences are transformed into discrete classes by the LLM, then regenerated into captions.

We measure the consistency between original and recovered labels/captions using Macro F1 Score and METEORBanerjee & Lavie (2005). In both settings, we observe strong alignment between the classification and regression outputs having **0.715 Macro F1 score** for Scenario (a) and **0.531 METEOR score** for Scenario (b) respectively, suggesting that LLM-generated labels preserve semantic consistency and structural fidelity. This indirect validation supports the utility of LLMs as reliable semantic intermediaries within a partially supervised annotation pipeline.

D.2 Ad-hoc Scenario (a): Bicycle disassembly/assembly

For example, given:

```
Verb: Moving Object and Walking
Tools: Bare Hand
Manipulated Object: Hex Key
Remarks: "Moving object from Table towards Bike"
```

we issue the following prompt:

```
% System message to define assistant role and style
System:
You are an expert activity-description assistant.
Always produce a single declarative sentence
in present continuous tense,
third person, starting with a capital letter and ending
with a period.

% User message with the annotation payload and example
User:
Convert the following structured annotation into
one clear sentence.

Annotation:
{ Verb: Moving Object and Walking
{ Tools: Bare Hand
{ Manipulated Object: Hex Key
{ Remarks: Moving object from Table towards Bike

Now, please convert the above annotation.
```

`GPT-4o` then output:

```
\He is moving the hex key using a bare hand and walking."
```

This generated sentence is used as the soft label target for downstream model training.

D.3 PROCEDURAL SCENARIO (B): 3D PRINTER ASSEMBLY/DISASSEMBLY

In the first stage, we used reasoning model DeepSeeker1 Guo et al. (2025) on all soft-label sentences to predict candidate activity classes. We iteratively extracted the set of unique predicted classes, re-running the pipeline until convergence (i.e., no new classes appeared). Next, we manually verified and merged semantically similar classes to produce the final discrete set: Pick, Move, Lift, Walk, Sit, Adjust, Stand, Read, Throw, Drink, Bent Down.

For example, given the soft label:

```
\The person bends down to pick something off the floor."
```

we construct a prompt to DeepSeeker1 that encourages open-ended reasoning about the described action, such as:

```
System:
You are a reasoning assistant tasked with extracting
activity-related
verbs or action phrases from natural language
descriptions of human behavior.
User:
Extract the key activity-related labels
from the following sentence:
\The person bends down to pick something off the floor."
```

DeepSeeker1 may then return:

```
["Pick", "Bend", "Grab"]
```

Each returned label is treated as a candidate activity class. We check each one against the current working set of known classes. If a label is not already in the set, we add it. This iterative procedure continues over the entire dataset of soft-label sentences. After each pass, we re-run DeepSeeker1 on any newly discovered or ambiguous phrases to catch any missed classes. The process continues until convergence, meaning no new unique classes are added in a full iteration.

In our example, if ”Pick” and ”Grab” are already in the working set but ”Bend” is not, we would update the set as:

```
Existing class set: ["Pick", "Grab", ...]
Updated class set: ["Pick", "Grab", "Bend", ...]
```

After convergence, we manually verify and merge semantically similar or redundant labels (e.g., merging ”Bend” and ”BentDown”) to finalize a clean, discrete set of activity classes:

```
Final class set: ["Pick", "Move", "Lift", "Walk", "Sit",
"Adjust", "Stand", "Read", "Throw",
"Drink", "Bent Down"]
```

In the second stage, we used GPT-4o Achiam et al. (2023) with prompt engineering to convert each soft-label sentence into one of these classes, employing a “sticky” logic so that the predicted class persists until a new class is detected at a later timestamp.

For example, given the soft label: *“The person is sitting in the chair, in front of the table.”*

we issue the following prompt:

```
System:
You are an activity-classification assistant.
Candidate classes: Pick, Move, Lift, Walk, Sit, Adjust,
Stand, Read, Throw, Drink, BentDown, Others.
Sticky logic: retain previous label unless a new one is predicted.
```

```
User:
Classify the following sentence into one of the candidate classes:
\The person is sitting in the chair, in front of the table."
```

`GPT-4o` then outputs: "10 (Sit)"

where the integer "10" refers to the class `Sit.` where the integer "10" refers to the class Sit. This two-stage approach yields our final hard labels for downstream model training.

E OBJECT TRACKING

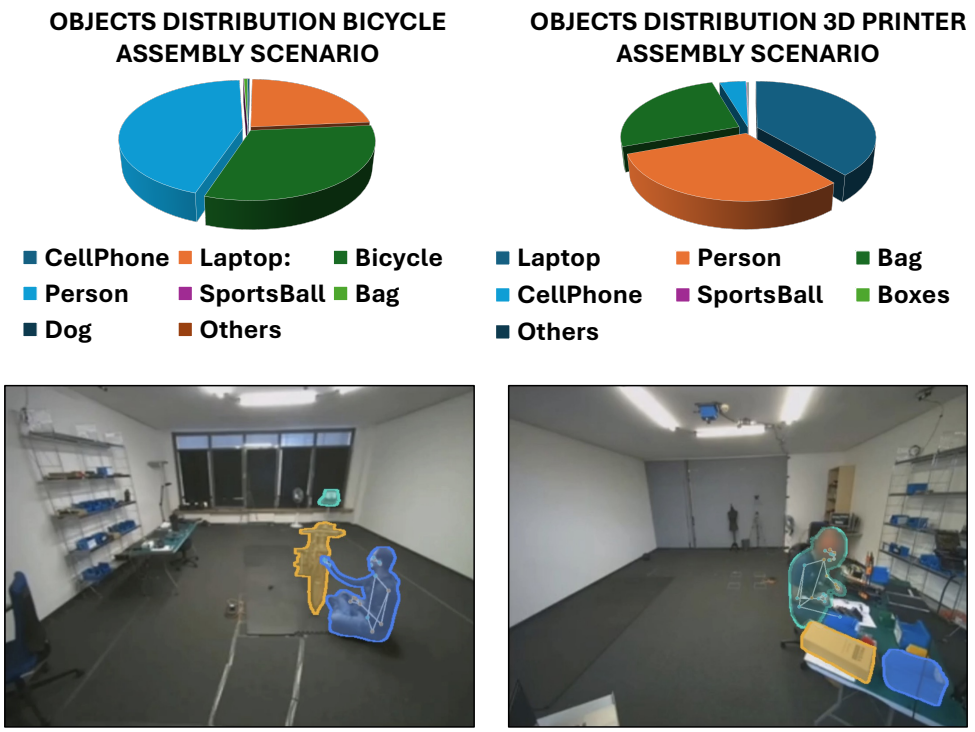

Figure 10: Object distributions in the bicycle and 3D printer scenarios from exocentric views, accompanied by scene visualizations illustrating object segmentation (masks) and human pose estimation in both environments.

OpenMarcie provides object segmentation and tracking data from multiple viewpoints, including an exocentric camera and two egocentric positions (head- and chest-mounted wearable cameras) for the 3D assembly experiment (Scenario (b)). Figure 10 **Top** shows the distribution of detected objects in the Scenario (a): bicycle, and Scenario (b): 3D printer assembly from exocentric views. In both cases, "Person" is the most frequently observed category, reflecting the consistent presence of participants during task execution. Common objects such as laptops, cellphones, and bags appear in both settings, likely reflecting typical work-related accessories.

Scenario-specific items highlight contextual differences: for example, the bicycle is unique to the bicycle assembly scenario, while boxes—potentially representing packaging or components—are exclusive to the 3D printer setup. Less relevant or incidental objects like sports balls and dogs are detected infrequently. Figure 10 **Bottom** presents example frames with object masks and human skeleton poses for both scenarios.

Overall, the figure underscores context-dependent variations in object presence, providing insights valuable for computer vision and robotics applications.

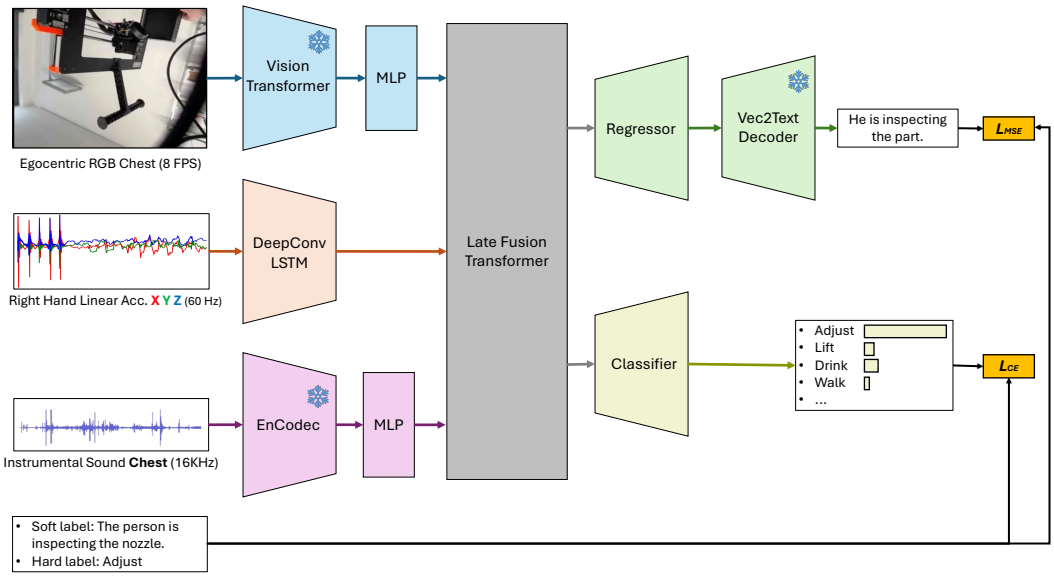

Figure 11: Architectures for classification and regression in human activity recognition and open-vocabulary captioning benchmarks.

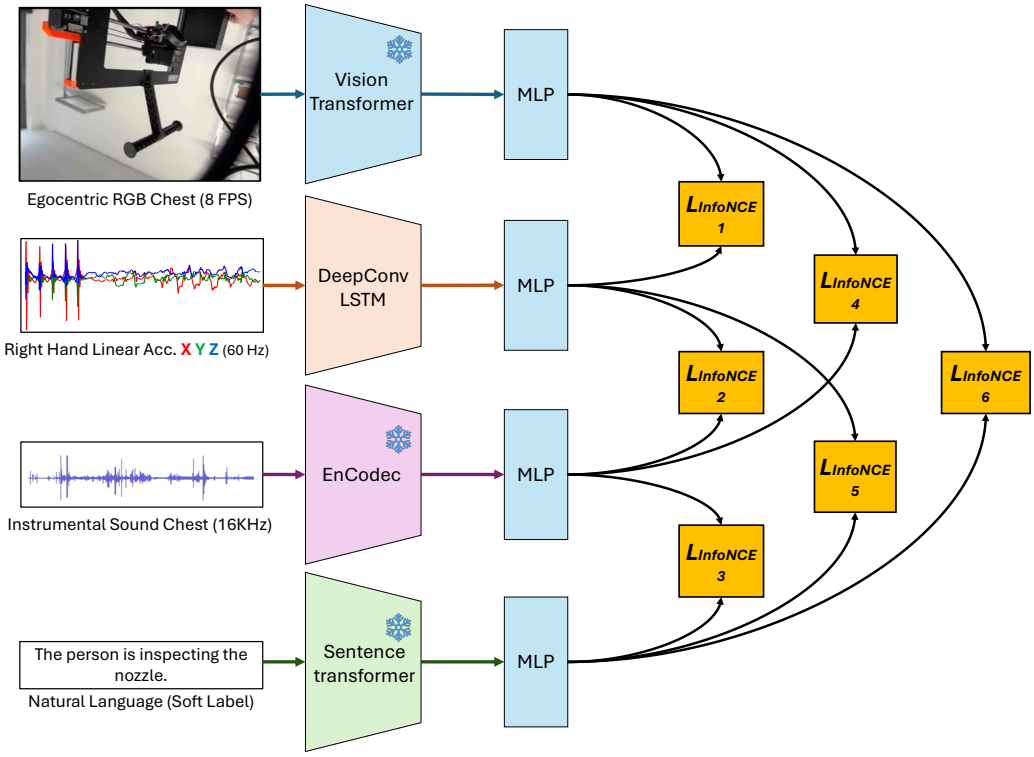

Figure 12: Architecture for cross-modal alignment benchmark.

F Benchmark Architecture

As shown in Figure 11, we segment data into synchronized 1-second windows and evaluate different modality combinations for human activity recognition. For each modality, input data is encoded independently: 3 video frames $\{x_v^t\}_{t=1}^3$ are processed via a Vision Transformer Dosovitskiy et al.

(2020) $\mathcal{E}_v$, IMU signals $x_i \in \mathbb{R}^{100\times 6}$ via a DeepConvLSTM Singh et al. (2020) encoder $\mathcal{E}_i$, and 1-second audio $x_a \in \mathbb{R}^{16000}$ via EnCodec Défossez et al. (2022) $\mathcal{E}_a$, producing embeddings $z_v = \mathcal{E}_v(\{x_v^t\})$, $z_i = \mathcal{E}_i(x_i)$, and $z_a = \mathcal{E}_a(x_a)$, respectively. For unimodal models, a classification head $\mathcal{C}$ maps each z_m to logits $\hat{y} = \mathcal{C}(z_m)$, where $m \in \{v, i, a\}$. For multimodal combinations (e.g., video+IMU, video+audio, or all three), embeddings are fused via a late-fusion transformer Pandeya & Lee (2021) $\mathcal{F}$: $z = \mathcal{F}(z_{m_1}, z_{m_2}, ...)$, followed by $\hat{y} = \mathcal{C}(z)$. All models are trained using cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{C} y_c \log \hat{y}_c \tag{1}$$

where $C = 12$ is the number of activity classes and y is the one-hot ground truth label.

We formulate open vocabulary captioning as a sentence embedding regression task, where models predict language representations from multimodal sensory inputs. Each modality is processed independently: 3 sampled video frames $\{x_v^t\}_{t=1}^3$ are passed through a Vision Transformer $\mathcal{E}_v$, 1-second audio $x_a \in \mathbb{R}^{16000}$ through EnCodec $\mathcal{E}_a$, and IMU signals $x_i \in \mathbb{R}^{100\times 6}$ through DeepConvLSTM $\mathcal{E}_i$, producing embeddings $z_v = \mathcal{E}_v(\{x_v^t\})$, $z_a = \mathcal{E}_a(x_a)$, and $z_i = \mathcal{E}_i(x_i)$. For unimodal or multimodal combinations, embeddings are fused via a transformer $\mathcal{F}$ to yield a shared representation $z = \mathcal{F}(z_{m_1}, z_{m_2}, \dots)$. A regression head $\mathcal{R}$ maps z to a predicted sentence embedding $\hat{s} = \mathcal{R}(z)$. Ground truth sentence embeddings s are obtained from a pretrained language encoder. Models are trained to minimize mean squared error (MSE):

$$\mathcal{L}_{\text{MSE}} = \|\hat{s} - s\|_2^2 \tag{2}$$

we perform caption retrieval using a Vec2Text Morris et al. (2023) decoder. This embedding-inversion approach enables scalable, low-latency caption generation without autoregressive decoding.

For cross-modal alignment (see Figure 12), we adopt a self-supervised contrastive learning approach using InfoNCE loss to map embeddings from different modalities into a shared representation space similar to ImageBind Girdhar et al. (2023). Using the same modality-specific encoders as before, we compute embeddings $z_m = \mathcal{E}_m(x_m)$ for each modality $m \in \{\text{video, audio, IMU, text}\}$. For each temporally aligned pair $(z_m, z_{m'})$, we apply a projection head and compute similarity with other samples in the batch. The InfoNCE loss is given by:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\text{sim}(z_m, z_{m'})/\tau)}{\sum_{z^-} \exp(\text{sim}(z_m, z^-)/\tau)} \tag{3}$$

where sim() is cosine similarity, τ is a temperature parameter, and z^- are negative samples.

F.1 Benchmark Metrics

We use task-specific metrics suited to the structure and goals of each benchmark:

HAR (Macro F1 Score): To account for class imbalance in multi-class activity recognition, we report macro-averaged F1 across all classes:

$$\text{F1}_{\text{macro}} = \frac{1}{C}\sum_{c=1}^{C} \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{4}$$

This gives equal weight to each class regardless of frequency.

Captioning (Cosine Similarity): We evaluate caption quality via cosine similarity between predicted and ground truth sentence embeddings:

$$\text{sim}(\hat{s}, s) = \frac{\hat{s} \cdot s}{\|\hat{s}\| \cdot \|s\|} \tag{5}$$

This reflects how well the model captures semantic similarity in the shared embedding space, without relying on exact lexical matches.

Table 9: Computational Resources for Joint Latent Representations Across Modality Combinations.

Combination	Parameters	MACs	FLOPS
Text + Audio	781,568	783,616	2.70×10^9
Text + Video	2,098,432	2,100,480	7.37×10^9
Text + IMU	712,512	22,797,312	2.07×10^{10}
Text + Video + IMU	2,351,552	24,437,376	2.00×10^{10}
Audio + Video + Text	2,420,608	2,423,680	5.91×10^9
Audio + Text + IMU	1,034,688	23,120,512	1.89×10^{10}

Table 10: Computational Resources for Classification Models Across Modality Combinations.

Combination	Params	MACs	FLOPS
Audio only	339,200	340,224	2.72×10^9
Video only	1,656,064	1,657,088	1.38×10^{10}
IMU only	270,144	22,353,920	2.39×10^{10}
Audio + Video	2,092,928	3,307,392	1.89×10^{10}
Audio + IMU	707,008	24,004,224	2.44×10^{10}
Video + IMU	2,023,872	25,321,088	2.57×10^{10}
Audio + Video + IMU	2,362,432	26,316,032	2.67×10^{10}

Cross-Modal Alignment (Retrieval Metrics): We assess alignment by retrieving the correct paired modality using Recall@k and Top-1 accuracy:

$$\text{Recall@}k = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[y_i \in \text{Top-}k\left(\hat{y}_i\right)\right] \tag{6}$$

$$\text{Top-1} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[\arg\max\left(\hat{y}_i\right) = y_i\right] \tag{7}$$

These metrics quantify retrieval quality in the shared embedding space, indicating how well modalities are aligned.

To validate the fidelity of our bidirectional annotation transforms, we compare original and recovered annotations using two complementary text-generation metrics, Macro F1 score for discrete labels and METEOR for captions.

Soft-Label Evaluation(METEOR Banerjee & Lavie (2005)): We compare user-annotated captions s and LLM-generated captions $\hat{s}$ using the METEOR metric:

$$\text{METEOR}(s, \hat{s}) = F_\alpha(s, \hat{s})\left(1 - \text{Pen}(s, \hat{s})\right) \tag{8}$$

F.2 Computational Resources and Model Efficiency

All experiments were conducted on an NVIDIA RTX 4090 GPU with 32GB VRAM. To assess the computational footprint of our models, we report the total number of parameters, multiply–accumulate operations (MACs), and FLOPs (floating-point operations per inference) for each experiment, along with the corresponding inference speed (see Table 9, Table 10, and Table 11).

G Audio Extended Experiments

Given the acoustic modality's limited standalone contribution, we provide additional experiments to better understand its behavior under different conditions. In particular, we analyze the impact of privacy-preserving anonymization (speech removal and replacement with instrumental textures) and feature extraction choices (Encodec embeddings vs. mel-spectrograms) across three tasks: human activity recognition, open-vocabulary captioning, and cross-modal alignment. These comparisons

Table 11: Computational Resources for Regression Models Across Modality Combinations.

Combination	Params	MACs	FLOPS
Audio only	746,624	744,000	5.95×10^9
Video only	1,952,768	1,952,000	1.63×10^{10}
IMU only	566,848	22,648,832	2.42×10^{10}
Audio + Video	2,270,080	2,268,992	1.30×10^{10}
Audio + IMU	884,160	22,965,824	2.33×10^{10}
Video + IMU	2,090,368	24,173,824	2.45×10^{10}
Audio + Video + IMU	2,407,616	24,490,816	2.49×10^{10}

Table 12: Human activity recognition results for Scenario (a): Bicycle Assembly and Scenario (b): 3D Printer Assembly, including macro F1 scores with and without the null class for Audio Extended Experiments.

Modality	Scenario (a)		Scenario (b)	
	Macro F1 (↑)			
	Without Null	With Null	Without Null	With Null
Acoustic-Anonymized Encoder	0.489±0.018	0.469±0.017	0.425±0.004	0.432±0.005
Acoustic-Non-Anonymized Encoder	0.509±0.002	0.488±0.001	0.460±0.002	0.453±0.003
Acoustic-Anonymized Mel-Spectrum	0.492±0.003	0.473±0.002	0.434±0.003	0.430±0.003
Acoustic-Non-Anonymized Mel-Spectrum	**0.517±0.004**	**0.493±0.002**	**0.466±0.003**	**0.455±0.002**

allow us to quantify how anonymization reduces semantic richness, whether alternative representations can recover useful signal, and to what extent audio still contributes in multimodal settings. The following tables report these extended results, clarifying both the challenges and the potential of acoustic data for privacy-preserving multimodal learning.

Across human activity recognition (Table 12), open-vocabulary captioning (Table 13), and cross-modal alignment (Table 14), a consistent trend emerges: non-anonymized audio outperforms anonymized variants, with Mel-spectrograms yielding the strongest results. For activity recognition, non-anonymized Mel-spectrograms achieve the highest macro F1 (0.517/0.493 in Scenario (a), 0.466/0.455 in Scenario (b)). In captioning, the same representation reaches the best cosine similarity (0.381/0.359 in Scenario (a), 0.330/0.340 in Scenario (b)). Finally, in cross-modal alignment, non-anonymized Mel-spectrograms again lead with recall@1/5 and top-1 scores (0.254/0.613/0.360 in Scenario (a); 0.238/0.597/0.345 in Scenario (b)). Scenario (b) consistently yields lower absolute performance, reflecting its greater procedural complexity. The performance gap between anonymized and non-anonymized streams (0.01–0.03 across metrics) highlights the trade-off between privacy and informativeness: anonymization systematically removes semantic richness, reducing discriminative power across all tasks. Still, audio retains complementary cues, particularly for grounding text, as seen in the stable improvements in cross-modal alignment even under anonymization.

The modality's limitations arise from four main factors:

- Privacy constraints: Speech removal and replacement with generic instrumental textures strip away contextual and semantic information.
- Feature extraction choices: Encodec embeddings, while general-purpose, may not capture fine-grained task-specific cues, especially on anonymized signals.
- Task-inherent challenges: Assembly sounds are subtle or intermittent, making them harder to discriminate.
- Environmental factors: Data collection in a test-bench environment lacks authentic industrial acoustics (e.g., machinery noise, vibrations), further constraining informativeness.

In sum, while acoustic data underperforms in isolation and can even introduce noise in classification tasks, it contributes positively in multimodal fusion and joint representation learning, and provides a testbed for exploring privacy-preserving sensing. The current experiments confirm both the utility of richer acoustic content and the limitations imposed by anonymization and feature choice.

Table 13: 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 for Audio Extended Experiments.

Modality	**Scenario (a)**		**Scenario (b)**	
	Cosine Similarity (↑)			
	Without Null	With Null	Without Null	With Null
Acoustic-Anonymized Encoder	0.361±0.030	0.341±0.018	0.316±0.003	0.323±0.004
Acoustic-Non-Anonymized Encoder	0.375±0.001	0.354±0.001	0.328±0.001	0.338±0.002
Acoustic-Anonymized Mel-Spectrum	0.364±0.003	0.348±0.002	0.326±0.002	0.322±0.001
Acoustic-Non-Anonymized Mel-Spectrum	**0.381±0.002**	**0.359±0.002**	**0.330±0.001**	**0.340±0.001**

Table 14: Cross-modal alignment results for Scenario (a): Bicycle Assembly and Scenario (b): 3D Printer Assembly, including recall@5, recall@1, and top-1 metrics.

Modality	**Scenario (a)**			**Scenario (b)**		
	Recall@1 (↑)	**Recall@5 (↑)**	**Top-1 (↑)**	**Recall@1 (↑)**	**Recall@5 (↑)**	**Top-1 (↑)**
Acoustic + Text (Anonymized Encoder)	0.241±0.014	0.583±0.025	0.342±0.016	0.227±0.013	0.567±0.022	0.329±0.015
Acoustic + Text (Non-Anonymized Encoder)	0.251± 0.001	0.605±0.002	0.355±0.002	0.237±0.001	0.589±0.001	0.341 ±0.001
Acoustic + Text (Anonymized Mel-Spectrum)	0.246 ± 0.001	0.595 ±0.002	0.350±0.001	0.233±0.001	0.578± 0.002	0.336 ±0.001
Acoustic + Text (Non-Anonymized Mel-Spectrum)	**0.254±0.001**	**0.613 ±0.002**	**0.360±0.001**	**0.238±0.001**	**0.597±0.001**	**0.345±0.001**

H Future Research Directions

Beyond the benchmarks presented in this work, our dataset opens several promising directions for the research community:

1. **Procedural Planning and Task Decomposition** Modeling the hierarchical structure of long-horizon industrial workflows, enabling systems to learn task graphs and segment complex sequences. Future work may explore methods for inferring workflow dependencies, optimizing decompositions, or generalizing across task variants.
2. **Skill Assessment and Expertise Modeling** Participant variability provides opportunities for modeling proficiency, efficiency, and learning progression. For example, future directions include automatic skill classification, modeling expertise transfer between agents, and designing adaptive training interventions guided by behavioral signals.
3. **Intent Prediction and Early Action Forecasting** Anticipating upcoming actions or goals from partial multimodal observations is essential for proactive assistance and collaboration. Potential research includes multimodal fusion strategies for early prediction, goal inference in partially observed sequences, and real-time assistive systems.
4. **Fine-Grained Action Segmentation and Role Understanding** Overlapping and concurrent actions offer a challenging testbed for segmentation and role inference. Open problems include modeling multi-label temporal boundaries, learning role dynamics in multi-agent or multi-phase tasks, and connecting segmentation to downstream planning.
5. **Pose Estimation and Body-Language Reasoning** With synchronized RGB-D and inertial data, future studies can advance full-body pose estimation, activity-conditioned pose forecasting, and non-verbal intent recognition in realistic industrial settings.
6. **Cross-Modal Knowledge Transfer** The dataset supports transfer learning across modalities—for instance, using vision or language to supervise inertial or acoustic models. This is especially relevant for privacy-sensitive or sensor-limited scenarios. Promising avenues include cross-modal distillation, modality dropout robustness, and unsupervised alignment.
7. **Cross-Modal Generation and Simulation** Generating one modality from another (e.g., IMU traces from instructions or reconstructing missing video) enables robust imitation learning and simulation. Future work may investigate generative modeling, simulation-to-reality transfer, and synthetic data augmentation for industrial tasks.