OpenMarcie: Dataset for Multimodal Action Recognition in Industrial Environments

Supplementary Material

This supplementary document contains additional information about OpenMarcie. Sec. 6 provides a detailed comparison to prior datasets, outlining differences in sensing coverage, industrial fidelity, and annotation design. Section 7 presents the metadata information about each participant independently. Section 8 declares the ethical consideration and societal impact of the OpenMarcie dataset with emphasis in user privacy and anonymization. Section 9 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 10 shows the object distribution for both experimental scenarios and illustrates example scenes with object segmentation and human skeleton poses. Section 11 reports tool-handling statistics in the bicycle scenario and highlights the multimodal cues involved in tool-contact events. Section 12 details the implementation of the proposed benchmarks and presents the evaluation metrics used. Sec. 13 reports additional audio-only experiments designed to probe the modality's limitations under different conditions, and analyzes audio's role in tool-contact detection. 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 14 describes the rationale behind the selected sensor modalities and their complementary physical signals. Sec. 15 reports results from modern multimodal architectures, including Perceiver IO and AnyMAL, demonstrating improvements over late-fusion baselines. Sec. 16 outlines several research directions that extend beyond the benchmarks presented in this work.

6. Detailed Comparison to Prior Datasets, Scope, and Generalization

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 [11], 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 [24] 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 [37] 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 [32] 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 [7] 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 [9] 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 [38] 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 [31] 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 [17] 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 Tab. 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.

6.1. Scope and Industrial Positioning

OpenMarcie focuses specifically on assembly-centric industrial workflows rather than attempting exhaustive coverage of all industrial activities. The dataset is structured around two complementary regimes: (i) ad-hoc, experience-driven assembly and disassembly (bicycle), capturing exploration, corrective behavior, and goal-oriented problem solving; and (ii) instruction-following procedural assembly (3D printer), reflecting standardized production and onboarding scenarios.

While we do not claim coverage of activities such as large-scale packaging, heavy machinery inspection, or conveyor-based logistics, the procedural scenario naturally incorporates industrially relevant behaviors beyond assembly itself, including unpacking, part organization, manual reading, workspace preparation, object transport, and sequential collaborative continuation of partially completed tasks. These behaviors reflect structured production-line dynamics within a controlled environment.

The design choice prioritizes controlled variability and multimodal richness over maximal task breadth, enabling systematic study of modality complementarity and procedural reasoning within assembly workflows.

6.2. Ecological Validity and Controlled Test-Bench Environment

Data collection was conducted in a controlled test-bench setting rather than an operational factory. Consequently, factors such as heavy machinery noise, large-scale environmental clutter, strict safety constraints, and unplanned workflow interruptions are underrepresented. We acknowledge this as a limitation.

However, the controlled environment provides several methodological advantages: synchronized multimodal capture, reduced confounding variables, stable camera calibration, and high annotation fidelity. This setup enables precise temporal alignment across modalities and systematic investigation of cross-modal robustness, which would be significantly more difficult in uncontrolled factory conditions.

OpenMarcie should therefore be understood as a controlled yet industrially inspired benchmark for multimodal perception in assembly contexts, rather than as a direct replica of a live factory floor.

6.3. Generalization and Domain Shift Considerations

With respect to generalization, models trained on OpenMarcie are expected to transfer most directly to assembly-oriented tasks involving fine motor manipulation, procedural reasoning, and multimodal perception under moderate environmental variability. Transfer to substantially different industrial domains—such as high-noise machining environ-

ments, high-speed packaging lines, or large-scale quality inspection pipelines—is not guaranteed.

Future evaluations should explicitly investigate robustness under domain shift conditions, including simulated machinery noise injection, environmental clutter augmentation, and cross-site validation in operational factory environments. Such experiments would provide quantitative assessment of real-world transferability.

In this context, OpenMarcie serves as a structured foundation for studying multimodal robustness, modality complementarity, and procedural activity modeling in assembly-centric industrial workflows, while recognizing the boundaries of its ecological validity.

7. 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.

In addition to the demographic information, we also report upper-body anthropometric measurements for Scenario (b), summarized in Table 8.

These measurements include shoulder-to-shoulder (SS) and shoulder-to-wrist lengths (SR-WR and SL-WL), expressed in centimeters, and provide a quantitative characterization of participant body proportions.

The physical dimensions of a person in a video—particularly their absolute scale, bounding box size, and relative segment proportions—are highly relevant to pose estimation performance. Although many modern algorithms aim to be scale-invariant, the pixel-level representation of the body directly influences keypoint localization accuracy and subsequent 3D reconstruction. In our setup, cameras are installed in the environment to estimate pose while participants perform assembly tasks and move freely within the room, continuously changing their distance to the cameras and thus their apparent scale. Providing subject-specific anthropometric measurements therefore enables more precise metric reconstruction in a centimeter-based reference frame, supports accurate multi-camera triangulation, and offers reliable ground-truth information for evaluating and improving pose estimation methods.

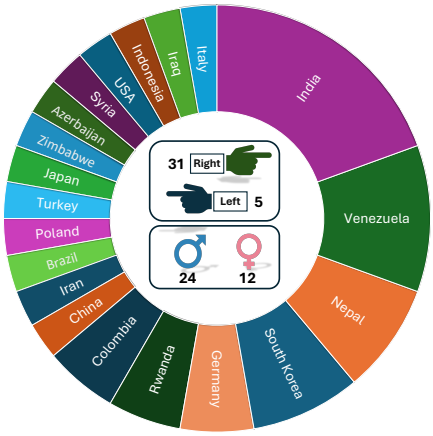

Figure 8. Distribution of participants by nationality, dominant hand and self identified sex.

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 9 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

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

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

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 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.

8. 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, and the study was reviewed and approved by the Ethics Board of the German Research Center for Artificial Intelligence under protocols HRW-35/24 and SMD-30/24. 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

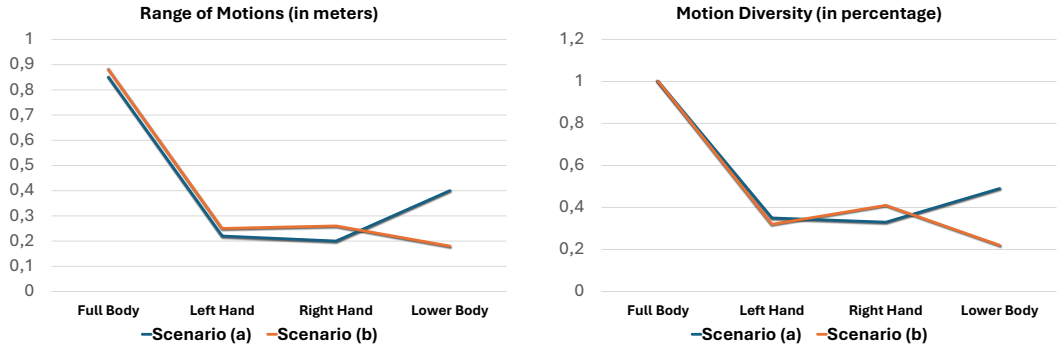

Figure 9. Comparison of range and diversity of body motions across Scenario (a): bicycle assembly and Scenario (b): 3D printer assembly.

Table 8. Upper-body anthropometric measurements: Shoulder-to-Shoulder length (SS), Shoulder-Right to Wrist-Right length (SR-WR), and Shoulder-Left to Wrist-Left length (SL-WL), measured in cm. Values are presented for each participant along with Mean ± Standard Deviation.

ID	SS (cm)	SR-WR (cm)	SL-WL (cm)
P1-D	51	55	55
P2-D	34	49	48
P3-D	51	61	61
P4-D	36	57	57
P5-D	36	54	54
P6-D	38	53	53
P7-D	46	59	59
P8-D	44	55	54
P9-D	40	55	54
P10-D	36	50	50
P11-D	38	53	53
P12-D	47	65	64
P13-D	44	55	55
P14-D	50	61	61
P15-D	38	46	46
P16-D	45	54	54
P17-D	41	56	57
P18-D	43	56	56
P19-D	32	52	52
P20-D	42	56	56
P21-D	47	59	59
P22-D	47	57	58
P23-D	38	49	49
P24-D	42	57	57
Mean ± SD	41.33 ± 5.82	55.04 ± 4.42	54.96 ± 4.43

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 [28], and only the resulting text descriptions were used in the benchmark tasks and released in OpenMarcie.

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) [25, 36]. Second, all videos were manually reviewed, and any remaining visible facial features were fully obscured using DaVinci Resolve [13]. For audio, only the instrumental components were released. Voices were intentionally removed from the audio tracks using the OpenVINO Music Separation plugin in Audacity [19], 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.

Table 9. 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

9. LLM-based Annotation Translation and Ground Truth Strategy

9.1. 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 METEOR[2]. 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.

9.2. 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.

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

In the first stage, we used reasoning model DeepSeeker1 [18] 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 [1] 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."
```

Table 10. Open-vocabulary cosine similarity when training on VLM auto-labels vs. human annotations. VLM labels yield scores only marginally above chance, substantially below the manual-label baselines.

Modality	Scenario (a)		Scenario (b)	
	Cosine Similarity (↑)			
	No Null	Null	No Null	Null
I	0.137±0.041	0.129±0.038	0.168±0.012	0.164±0.011
A	0.098±0.035	0.091±0.033	0.103±0.009	0.099±0.010
V	0.152±0.029	0.143±0.027	0.181±0.008	0.177±0.009
I + A	0.141±0.039	0.133±0.036	0.172±0.010	0.168±0.011
A + V	0.148±0.033	0.139±0.030	0.179±0.009	0.174±0.009
I + V	0.163±0.028	0.154±0.026	0.194±0.007	0.189±0.008
I + A + V	0.159±0.031	0.150±0.029	0.190±0.008	0.185±0.009

`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.

Ground Truth Acquisition Strategy. The ground-truth annotations described above rely on synchronized multi-view capture rather than dedicated motion-capture systems. While high-accuracy optical mocap could in principle provide precise kinematic measurements, we intentionally avoided marker-based setups. Such systems are intrusive for long-horizon tool use, may restrict natural manipulation, and are particularly susceptible to occlusions during close-range assembly involving frequent hand–object interactions. These limitations can compromise ecological validity in assembly-centric workflows. Instead, we leverage synchronized multi-view RGB-D acquisition using ZED stereo cameras, which provides strong 3D scene perception while remaining minimally intrusive and scalable. Manual, intent-aware annotation is performed on the most informative exocentric view and temporally aligned across modalities, ensuring semantically consistent ground truth while preserving natural interaction behavior.

VLM-Based Automatic Labelling To assess whether Vision–Language Models (VLMs) can replace or reduce manual annotation effort, we generated automatic soft labels by prompting a state-of-the-art VLM Qwen2-VL [34] to caption each egocentric video segment using the same temporal boundaries as the manual annotations. Table 10 reports the open-vocabulary cosine similarity achieved when training the regression pipeline on these VLM-generated labels instead of human-written soft labels.

Across all modality combinations and both scenarios, VLM auto-labels produce cosine similarity scores in the range 0.09–0.19, only marginally above a random-embedding baseline (≈0.0). By comparison, the same regression model trained on human-written soft labels achieves 0.36–0.56 in Scenario (a) and 0.32–0.66 in Scenario (b), a gap of 3–4×. Qualitative inspection reveals that VLM captions tend to describe generic scene context (*"a person standing in a workshop"*) rather than the fine-grained tool–object interactions captured by human annotators (*"He is screwing the brake using a hex key"*). This discrepancy is especially pronounced for short action segments where the VLM lacks sufficient temporal context to distinguish between visually similar activities. Given these findings, we omit VLM auto-labels from the main benchmarks.

Table 11. Tool vocabulary for Scenario (a) (Bicycle Assembly). Percentages are computed over the 1,769 tool-annotated action instances aggregated across all users.

Code	Tool	Usage in %
bh	Bare hand	72.5%
hx	Hex key	16.2%
sd	Screwdriver	3.3%
wr	Wrench	2.9%
hk	Hook	2.2%
pu	Pump	1.0%
pl	Pliers	1.0%
hm	Hammer	0.6%
sc	Scissors	0.2%

10. Object Tracking

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.

11. Tool Handling

Table 11 lists the nine-tool vocabulary used in the Scenario (a) verb–tool–object annotation scheme. Bare-hand

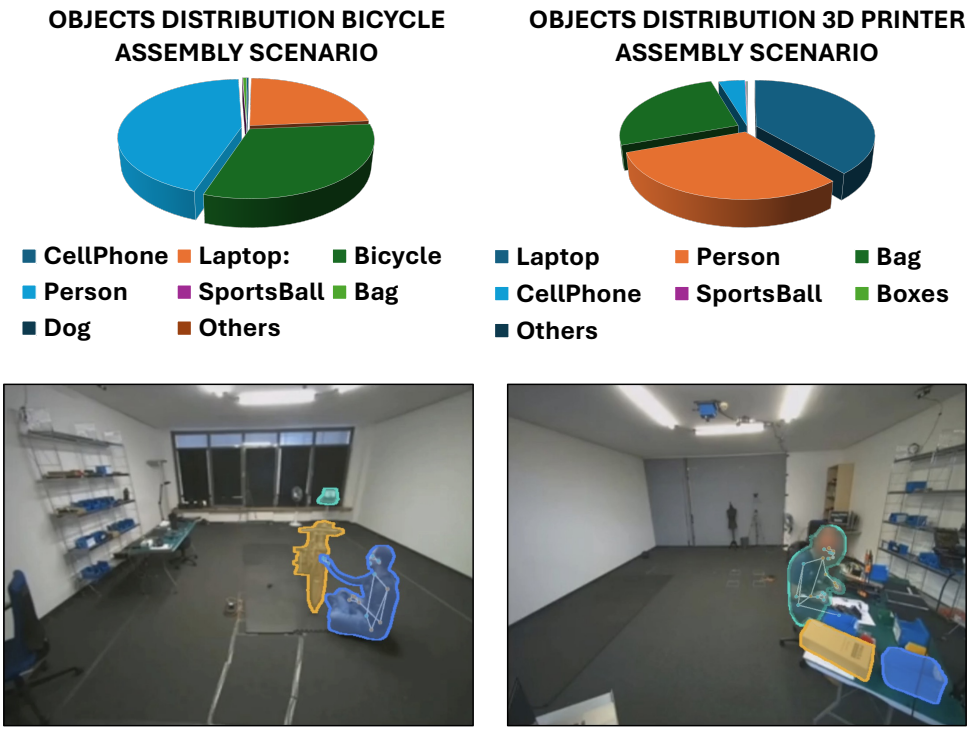

Figure 10. Object distributions in the bicycle and 3D printer scenarios from exocentric views, accompanied by scene visualizations illustrating object segmentation (masks) and human pose estimation in both environments.

manipulation dominates (72.5%), reflecting the prevalence of positioning and inspection steps during bicycle assembly. The hex key (Allen key) accounts for 16.2% of tool instances and is the most frequent specialised tool, consistent with its ubiquitous use in tightening bicycle bolts. The remaining tools—screwdriver, wrench, hook lever, pump, pliers, hammer, and scissors—collectively cover 11.3% of instances. As shown in Table 18, the acoustic modality is especially valuable for detecting these tool-contact events: adding audio to V+I raises Macro F1 from 0.905 to 0.923 in Scenario (a) and from 0.898 to 0.914 in Scenario (b), because transient impact and friction sounds produced by tools such as the hammer, wrench, and hex key carry discriminative signatures that complement the inertial and visual channels.

12. 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 [14] $\mathcal{E}_v$, IMU signals $x_i \in \mathbb{R}^{100\times 6}$ via a DeepConvLSTM [33] encoder $\mathcal{E}_i$, and 1-second audio $x_a \in \mathbb{R}^{16000}$ via EnCodec [12] $\mathcal{E}_a$, 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 [26] $\mathcal{F}$: $z = \mathcal{F}(z_{m_1}, z_{m_2}, ...)$, 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

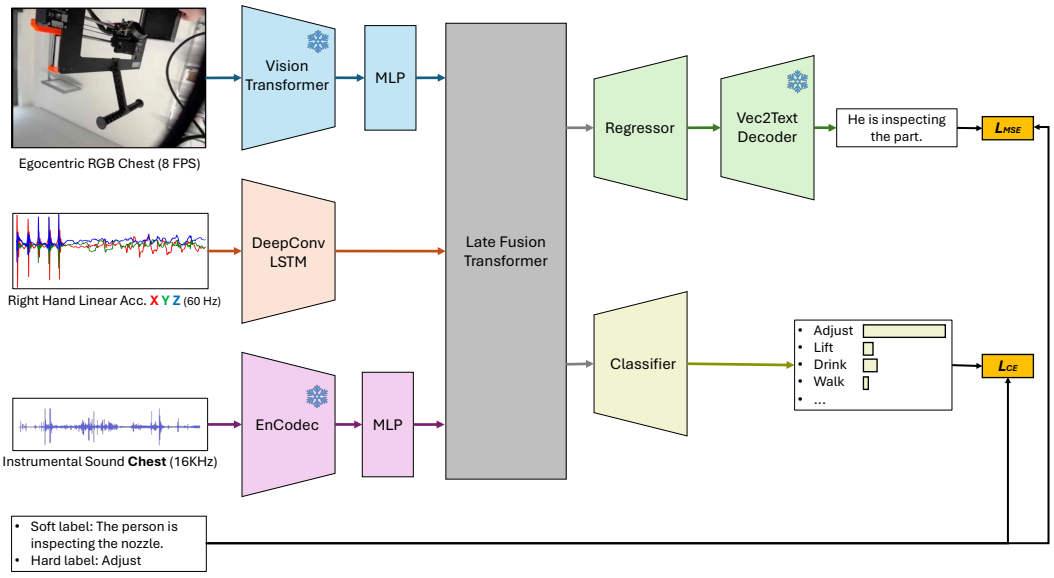

Figure 11. Architectures for classification and regression in human activity recognition and open-vocabulary captioning benchmarks.

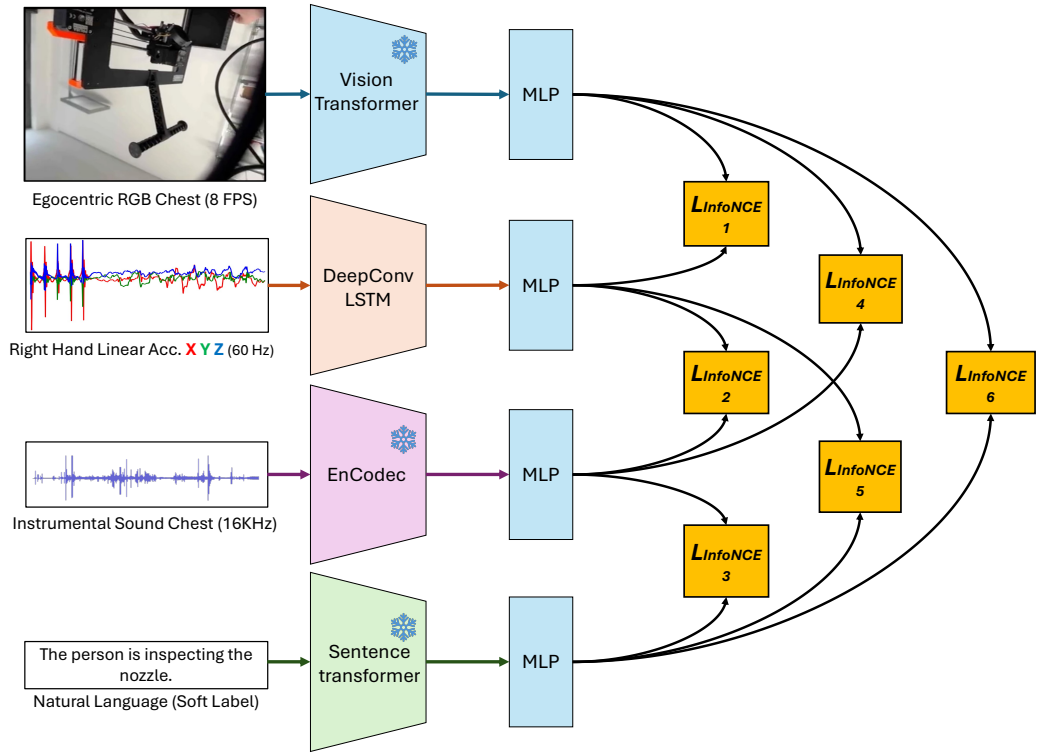

Figure 12. Architecture for cross-modal alignment benchmark.

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 [23] 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 [16]. 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.

12.1. 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.

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 [2]): We compare user-annotated captions s and LLM-generated captions $\hat{s}$ using the METEOR metric:

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

12.2. Training Details

All models are trained using a fixed sliding-window strategy to ensure consistent temporal context across modalities. Any segment shorter than the required window length is skipped during dataset construction to guarantee complete temporal context for every sample.

Across all models, a batch size of 128 and an initial learning rate of 1×10^{-4} are used. The Late Fusion (classification and regression) and JLR contrastive models are optimised with Adam and no weight decay. All models are trained for 300 epochs with early stopping. Learning-rate scheduling follows a ReduceLROnPlateau strategy, with a patience of 5 epochs, and a decay factor of 0.5.

Architecturally, the Late Fusion and JLR models use a hidden dimension of 128 with two Transformer layers, four attention heads, and a feed-forward dimension of 256. The classification model uses a BCEWithLogits objective, the regression model uses mean squared error (MSE), and the JLR model employs an InfoNCE contrastive objective with six pairwise terms and a temperature parameter $\tau = 0.1$.

For weight initialisation, all models rely on PyTorch's Kaiming uniform initialisation for Linear and Conv layers, with biases initialised to zero. LSTM forget-gate biases are set to 1.0 to encourage stable long-range gradient flow.

The framework supports flexible modality handling. Optional modalities are instantiated only when enabled via configuration flags. During batching, optional tensors are retrieved , returning None when a modality is not present, ensuring robustness when certain data streams are unavailable. Dropout is applied with a rate of 0.3 in MLP encoder hidden layers and 0.1 within Transformer encoder layers.

12.3. 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 Tab. 12, Tab. 13, and Tab. 14).

13. 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,

Table 12. 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 13. 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}

Table 14. 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	1.70×10^{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}

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 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 (Tab. 15), open-vocabulary captioning (Tab. 16), and cross-modal alignment (Tab. 17), 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.

Contact Detection using Audio Table 18 reports tool-contact/impact detection results. The binary contact labels were generated by classifying each temporally-segmented annotation from the original open-vocabulary soft labels into *tool-contact* or *tool-non contact* classes using a hybrid pipeline: a rule-based first pass matches tool-related keywords and action patterns (e.g., *hammering*, *tightening with wrench*), followed by an LLM-based refinement step for ambiguous instances. Notably, while acoustic features alone perform poorly for general activity recognition, they prove highly informative for detecting tool-contact events, reaching 0.783 and 0.772 as a single modality. Moreover, adding audio to the best bimodal combination V + I consistently improves tool-contact F1: from 0.905 to 0.923 in Scenario (a) and from 0.898 to 0.914 in Scenario (b). This suggests that the acoustic channel captures transient impact signatures such as hammering, clicking, and screwing sounds that are largely redundant with IMU vibrations for coarse activity classes but become discriminative when the task requires detecting physical contact between a tool and

Table 15. 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**

Table 16. 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 17. 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**

Table 18. Tool-contact/impact detection Macro F1 scores for Scenario (a): Bicycle Assembly and Scenario (b): 3D Printer Assembly, with and without the null class. Unlike general HAR, audio is complementary for detecting tool-contact events: adding acoustic data to V + I improves F1 from 0.905 to 0.923 in (a) and from 0.898 to 0.914 in (b).

Modality	**Scenario (a)**		**Scenario (b)**	
	Macro F1 (↑)			
	No Null	Null	No Null	Null
Inertial(I)	0.872±0.008	0.847±0.009	0.861±0.006	0.836±0.006
Acoustic(A)	0.783±0.012	0.762±0.013	0.772±0.005	0.750±0.006
Vision(V)	0.841±0.009	0.817±0.010	0.833±0.004	0.808±0.005
I + A	0.898±0.007	0.871±0.008	0.889±0.004	0.863±0.004
A + V	0.863±0.009	0.838±0.010	0.855±0.004	0.830±0.005
I + V	0.905±0.006	0.879±0.007	0.898±0.003	0.872±0.004
I + A + V	**0.923±0.005**	**0.897±0.006**	**0.914±0.003**	**0.888±0.004**

a workpiece.

14. Sensor Design Rationale

As summarized in Tab. 19, the sensing modalities in OpenMarcie were selected based on physical complementarity rather than performance maximization.

IMUs form the primary wearable motion modality, capturing fine-grained hand and body dynamics essential for distinguishing manipulation patterns; however, inertial integration is subject to drift, motivating complementary signals. Magnetometers support orientation stabilization and reduce long-term rotational drift. Barometers provide vertical motion cues related to posture transitions (e.g., standing, kneeling, bending), offering an independent signal correlated with altitude.

Although explicit sensor fusion is not performed in the current benchmarks, barometers are deployed at multiple body locations to support potential relative height reasoning. In such a configuration, sudden changes pressure variations would affect all sensors in a similar (common-mode) manner, enabling differential pressure measurements to provide more stable vertical displacement cues. Ambient temperature is additionally recorded to ensure physically consistent altitude estimation through the barometric (hypsometric) relation, which models how air pressure varies with height [4]. In this context, temperature serves as an auxiliary environmental variable supporting principled vertical motion interpretation rather than as an independent activity recognition signal.

Egocentric RGB-D sensing provides semantic grounding by capturing object identity, spatial relationships, and hand–object interactions, while exocentric multi-view RGB-D cameras offer global scene context and a stable reference for annotation and cross-modal alignment. LiDAR depth improves spatial robustness under varying lighting

Table 19. Sensor Modalities and Their Intended Roles in OpenMarcie.

Modality	Physical Signal	Assembly-Relevant Attributes	Contribution
IMU (Wrists, Head)	Acceleration, angular velocity, orientation	Tool manipulation, temporal segmentation	High
Magnetometer	Magnetic field orientation	Heading stabilization, orientation consistency	Supportive
Barometer (multi-position)	Air pressure (hPa)	Posture transitions, vertical displacement cues	Medium
Temperature (ambient)	Ambient temperature	Environmental context, barometric formula	Low (supporting)
Spectrometer	Material spectral reflectance	Material differentiation (metal vs. plastic)	Low (Exploratory)
Thermal Camera	Surface temperature distribution	Contact duration, friction heat	Low (Exploratory)
Egocentric RGB-D	Visual appearance + depth	Object identity, hand–object interaction	High
Exocentric RGB-D (multi-view)	Scene-level RGB-D	Global spatial context	High
Stereo Audio	Instrumental/environmental sound	Tool–material interaction signatures	Medium
LiDAR Depth	Active depth sensing	3D spatial structure, object distance	High (with vision)

conditions and strengthens 3D reasoning.

Stereo audio captures tool–material interaction signatures that complement motion and vision, and although standalone acoustic performance is limited for activity recognition, it is highly informative for detecting tool-contact events, as shown in Tab. 18.

Finally, the spectrometer and thermal camera–based visual sensors are included as exploratory modalities to support future research on material-aware interaction modeling and the visual capture of physical signals outside the human-visible spectrum.

Overall, the multimodal design reflects a structured attempt to capture complementary physical aspects of industrial activity rather than an effort to maximize the number of modalities.

15. Modern Architectures

Tables 20 and 21 summarize the performance of modern multimodal architectures across human activity recognition (HAR) and open-vocabulary captioning for both Scenario (a) and (b). For HAR, the Perceiver IO inspired model [20] consistently outperforms late fusion baselines, with the best results achieved when leveraging all modalities, reaching a Macro F1 of 0.915 (no null) and 0.886 (null) in Scenario (a), and 0.842 (no null) and 0.746 (null) in Scenario (b). Similarly, in open-vocabulary captioning, the AnyMAL [22] LLM-based approach surpasses late fusion, particularly when incorporating all modalities, achieving cosine similarity scores of 0.643 (no null) and 0.610 (null) in Scenario (a), and 0.741 (no null) and 0.740 (null) in Scenario (b). Overall, the results demonstrate that modern multimodal fusion architectures and LLM-based approaches provide consistent improvements over late fusion baselines, especially when all modalities are utilized.

16. Future Research Directions

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

Table 20. Macro F1 scores for human activity recognition in Scenario (a) and Scenario (b) with and without the null class. Results are grouped by late fusion baselines, and Perceiver IO baseline with all modalilies including IMU, Audio, Video, Barometer, Temperature, Spectrometer, Thermal signals.

Modality	Scenario (a)		Scenario (b)	
	Macro F1 (↑)			
	No Null	Null	No Null	Null
Late Fusion				
I + A + V	0.859±0.010	0.831±0.011	0.763±0.003	0.676±0.003
All	0.891±0.008	0.862±0.008	0.825±0.003	0.731±0.004
Perceiver IO				
I+A+V	0.882±0.007	0.853±0.008	0.779±0.003	0.690±0.003
All	**0.915±0.005**	**0.886±0.006**	**0.842±0.002**	**0.746±0.003**

Table 21. Open vocabulary captioning results for Scenario (a) and Scenario (b) with and without the null class. Results are grouped by late fusion baselines, and the AnyMAL LLM-based baseline.

Modality	Scenario (a)		Scenario (b)	
	Cosine Similarity (↑)			
	No Null	Null	No Null	Null
Late Fusion				
I + A + V	0.547±0.020	0.519±0.017	0.647±0.001	0.646±0.003
All	0.593±0.015	0.563±0.014	0.691±0.002	0.690±0.002
AnyMAL				
I+A+V	0.622±0.012	0.562±0.012	0.714±0.002	0.693±0.002
All	**0.643±0.010**	**0.610±0.010**	**0.741±0.002**	**0.740±0.002**

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 multi-

modal 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.