IndEgo: A Dataset of Industrial Scenarios and Collaborative Work for Egocentric Assistants

Vivek Chavan✉[1,2], Yasmina Imgrund[2+], Tung Dao[2+], Sanwantri Bai[3+], Bosong Wang[4+], Ze Lu[5+], Oliver Heimann[1], Jörg Krüger[1,2]

Project Page: https://indego-dataset.github.io/

1, + Fraunhofer IPK
2 TECHNISCHE UNIVERSITÄT BERLIN
3 EBERHARD KARLS UNIVERSITÄT TÜBINGEN
4 RWTH AACHEN UNIVERSITY
5 Leibniz Universität Hannover

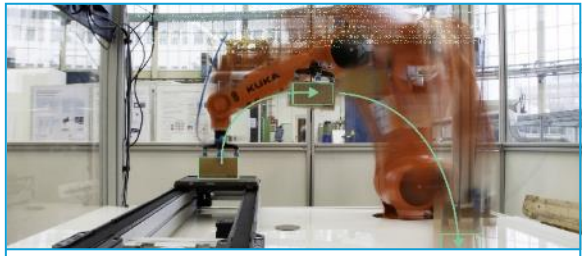
Industry Trends and Research Directions

Industry 4.0

- Individual technologies and system solutions for digitally integrated production
- Industrie 4.0 Transfer Center Berlin
- Berlin Center for Digital Transformation

Digital engineering

- Digital development of the future
- Digital twins
- Information factory for PLM and IoT
- Digital factory and inspection
- Smart products and services
- Model-based systems engineering



Automation

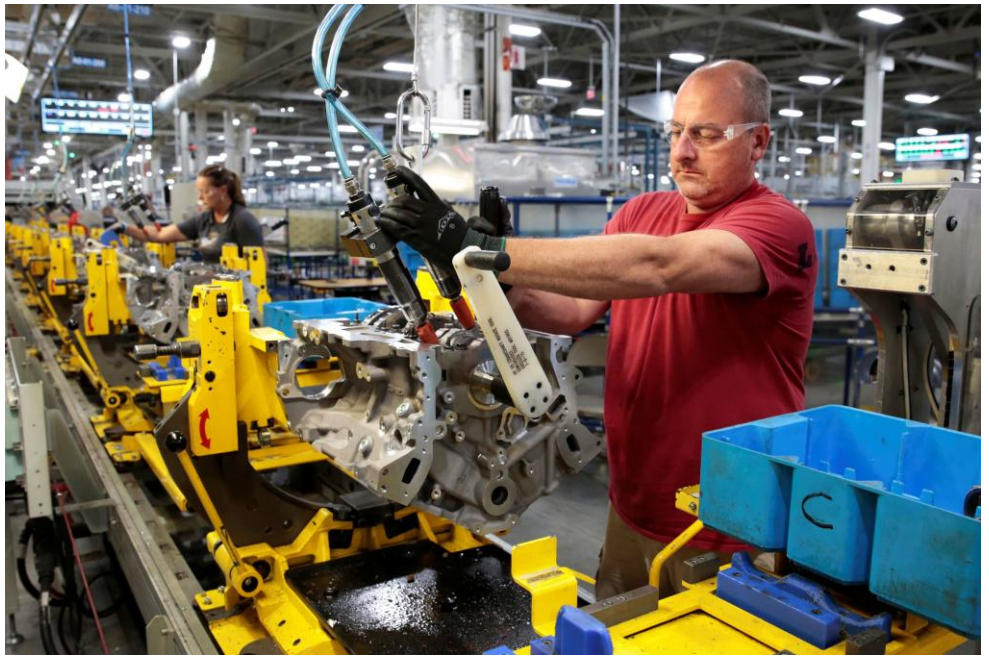
- Industrial image processing
- Virtual reconstruction
- Computer vision for safety and security
- Industrial robotics
- Process technology and optimization
- Intelligent systems for health and ergonomics

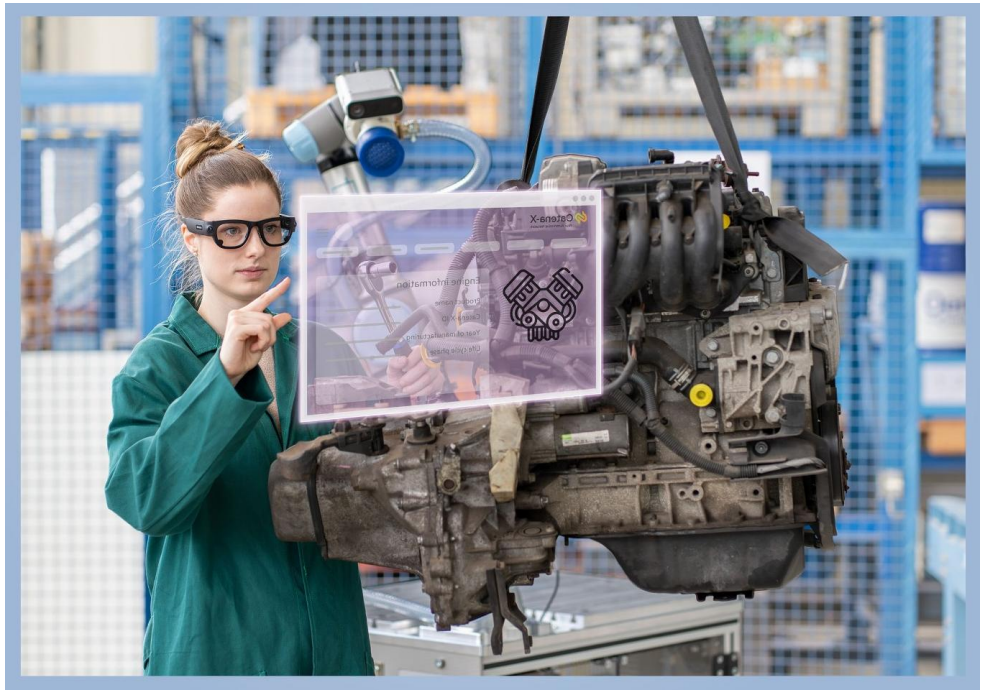
Artificial Intelligence

- Research and Development aligned with emerging trends and industry needs.
- Computer Vision, Robotics, Knowledge Representation.
- Image and Video Understanding, Data curation and benchmarking.
- Closing the loop on applied research using industry collaboration.

Motivation & Gaps

- 50% of the world‘s GDP is manual work, despite all the automation ($42T).
- These tasks and processes are highly context-dependent and cannot be programmed or automated by classical approaches.
- Multimodal data and Video Understanding open new avenues of industrial applications w.r.t. procedural data, digitising expert knowledge and downstream developments.
- Such industrial applications need context-aware intelligent assistants
- Egocentric vision provides a natural interface
- Application Potential: AR/VR, user guidance, cognitive robotics.



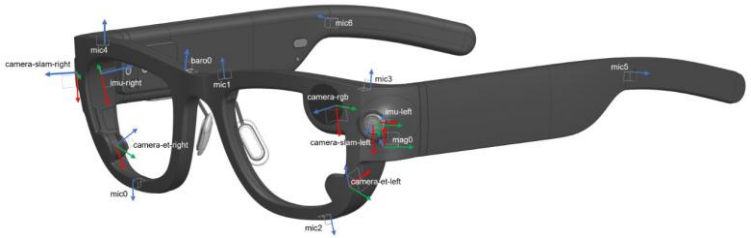


Motivation & Gaps

- Lack of representative data.
- Existing datasets poorly represent industrial use-cases and their challenging settings.
- We need datasets that bridge academic vision and industrial use

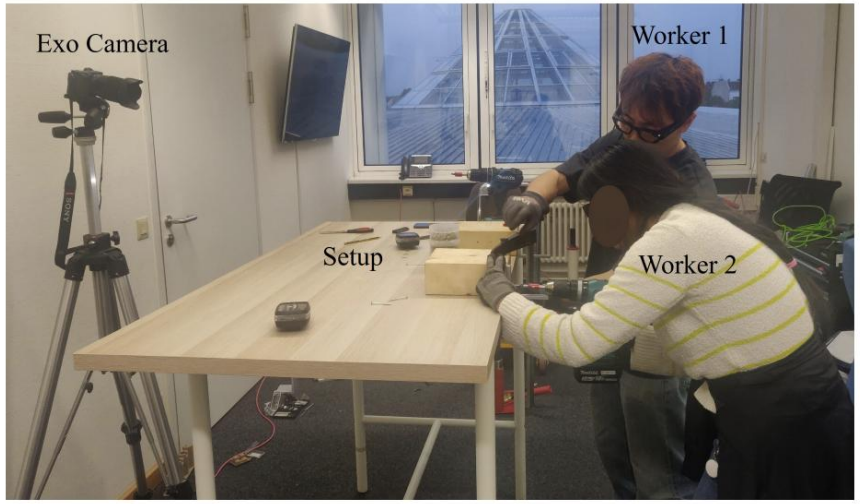
Gaps:
× Long sequence tasks
× Collaborative Scenarios
× Industrial setups and use-cases
× Multimodality in industry

Dataset	Scenario	Hours (Ego)	Exo	Collaboration	Gaze	Motion	Narration	Actions	Keysteps	Mistakes	QA
EPIC-KITCHENS [13]	Kitchen	100	✗	✗	✓	✗	✓	✓	✗	✗	✗
CharadesEgo [44]	Daily	34	✓	✗	✓	✗	✓	✗	✗	✗	✗
Ego4D [12]	Multiple	3670	✓	✗	✗	✓	✓	✗	✓	✗	✗
LEMMA [51]	Daily	10	✓	✓	✗	✗	✗	✓	✗	✗	✗
Ego-Exo4D [14]	Multiple	221	✓	✓	✗	✓	✓	✓	✓	✗	✗
EgoExoLearn [45]	Daily, Lab	120	✓	✗	✓	✗	✓	✓	✓	✗	✗
Nymeria [52]	Daily	300	✓	✓	✓	✓	✓	✓	✗	✗	✗
AssistQ [48]	Assistive	3	✗	✗	✗	✓	✓	✓	✗	✗	✓
Meccano [17]	Industry-like	7	✗	✗	✓	✗	✓	✓	✗	✗	✗
HoloAssist [48]	Assistive	166	✗	✓	✓	✓	✓	✓	✓	✓	✗
Assembly101 [18]	Industry-like	42	✓	✗	✗	✗	✗	✓	✓	✓	✗
IndEgo (ours)	Industrial	197	✓	✓	✓	✓	✓	✓	✓	✓	✓

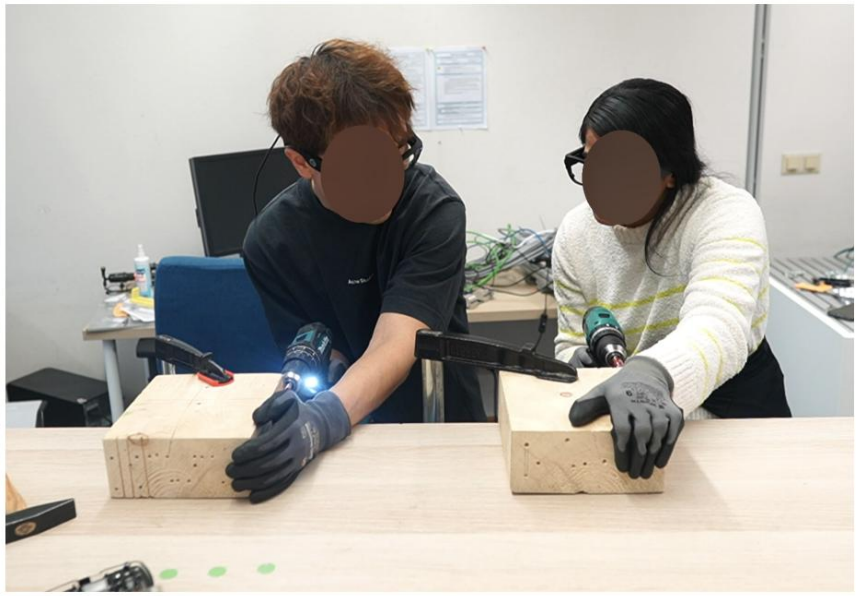


Project Aria Research Kit
(Meta Reality Labs)

Industrial Tools and Devices



Setup



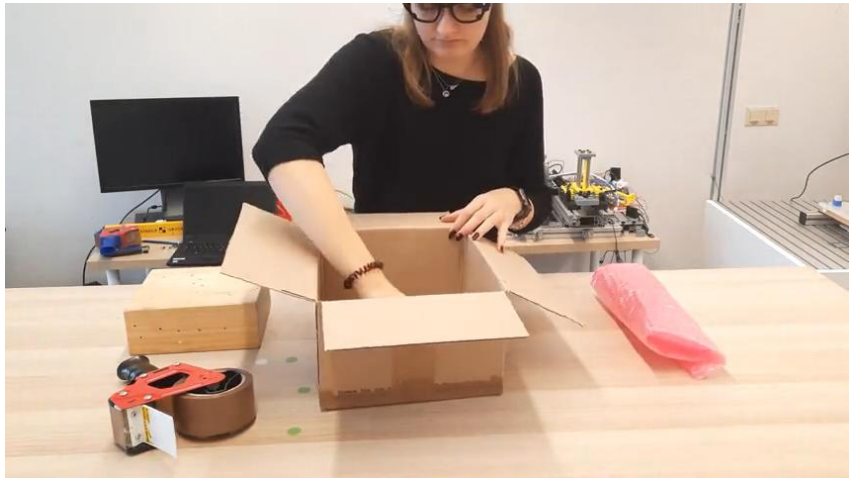
Exo Perspective

Ego Perspectives

- ✓ 20 Participants
- ✓ Different locations, including research labs and test fields
- ✓ Varying setups depending on the tasks & workflow

IndEgo Dataset

- The scenarios and use-cases were chosen to represent most likely application areas for real-world deployment.
- Both perspectives have varying advantages and limitations.



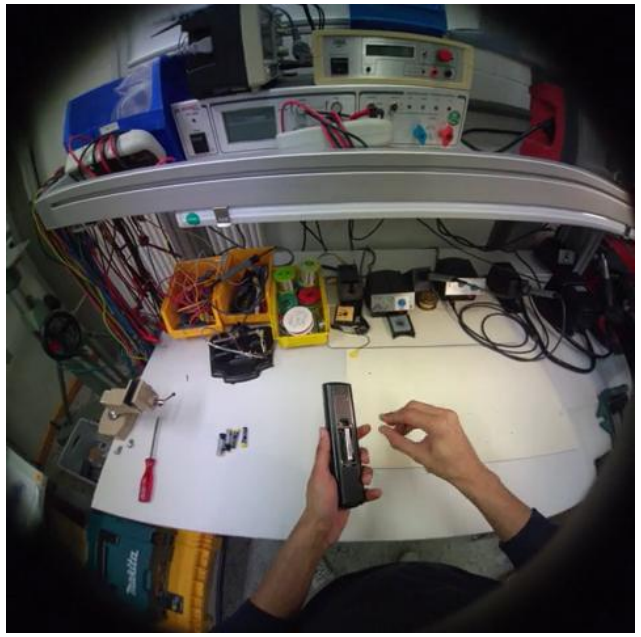
Exocentric View

Egocentric View

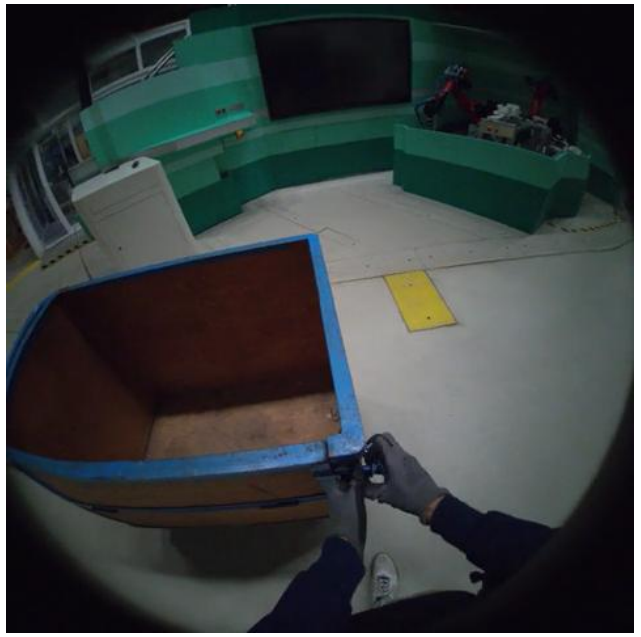
Category	T_{avg}	1-Person	Collaborative	Narration	#Ego	T_{Ego} (h)	#Exo	T_{Exo} (h)
Assembly	15.2 m	✓	✓	✓	188	47.5	152	30.4
Disassembly	11.1 m	✓	✓	✓	136	24.9	112	17.0
Inspection and Repair	7.8 m	✓	✓	✓	238	30.9	202	17.7
Logistics/Organisation	4.5 m	✓	✓	✓	456	35.4	158	8.1
Woodworking	7.5 m	✓	✓	✓	148	18.4	116	14.9
Miscellaneous	1.5 m	✓	✓	✗	378	9.4	352	8.7
Tools/Objects in Context	120 s	✓	✗	✗	604	20.1	–	–
Tools/Objects Demo	53 s	✓	✗	✓	302	4.5	–	–
Singular Actions	21 s	✓	✓	✗	1010	5.9	–	–
Total	205 s	✓	✓	✓	**3460**	**197.1**	**1092**	**96.8**

Table 1: A breakdown of the IndEgo dataset, showing the key categories and related statistics. T_{avg} gives the average duration of the recording, #Ego gives the number of videos from the Egocentric perspective, T_{Ego} gives the total cumulative time for egocentric data, #Exo gives the number of videos from the fixed exocentric perspective, T_{Exo} gives the total cumulative time for exocentric data.

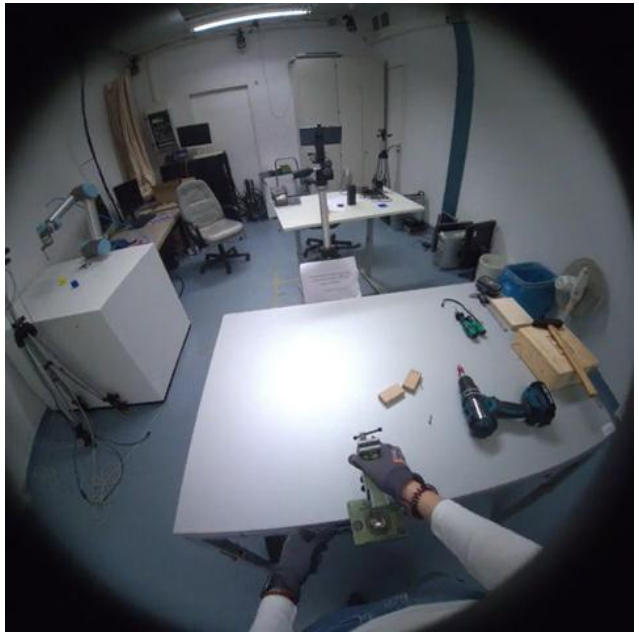
Assembly/Disassembly



Inspection/Repair



Logistics/Organisation

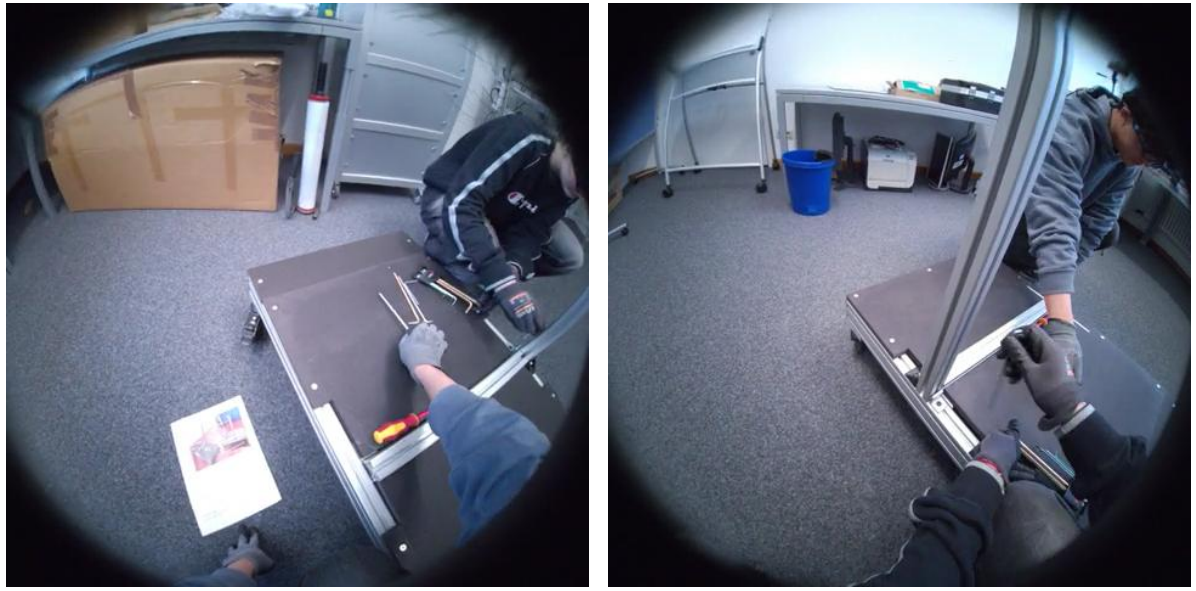


Woodworking

Miscellaneous

- ✓ 197 hours of Egocentric Data
- ✓ 97 hours of Exocentric Data
- ✓ Diverse industrial scenarios
- ✓ Collaborative work, physically and congitively demanding tasks

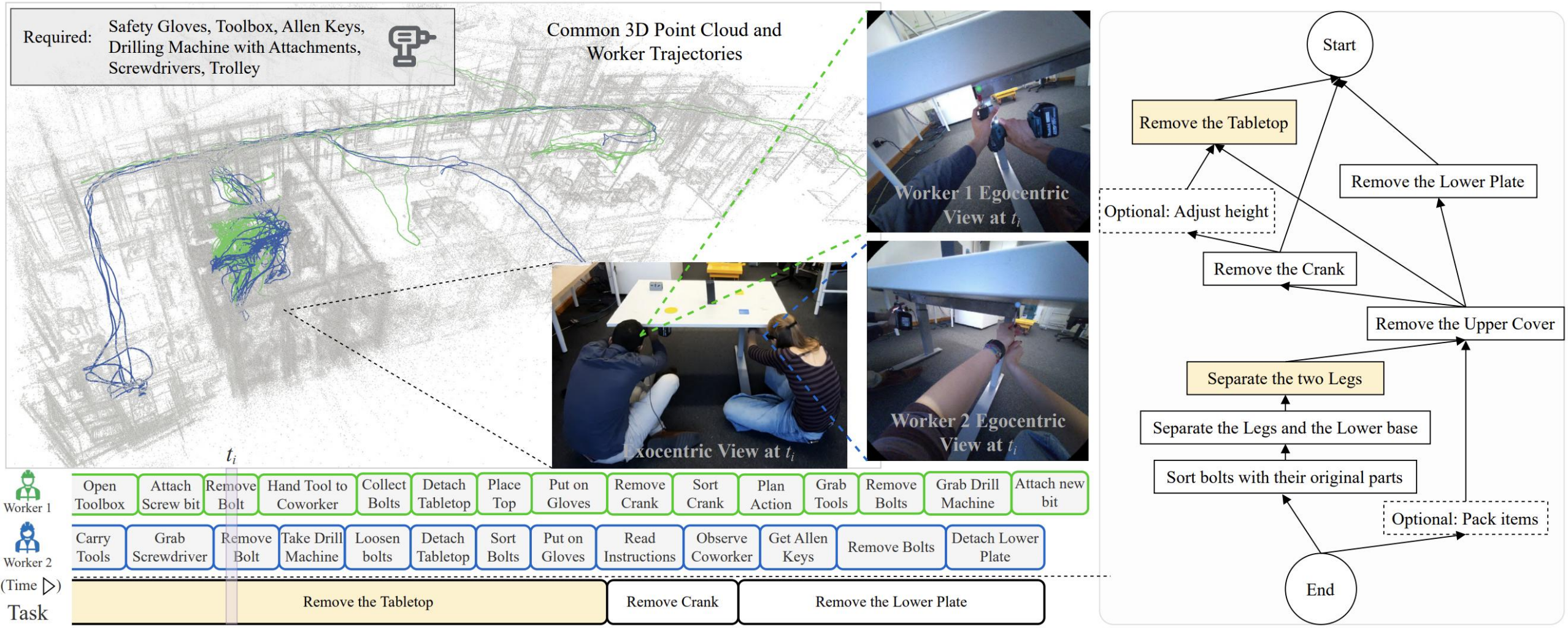
Collaborative Work



Coworkers

Leader-Follower/Training

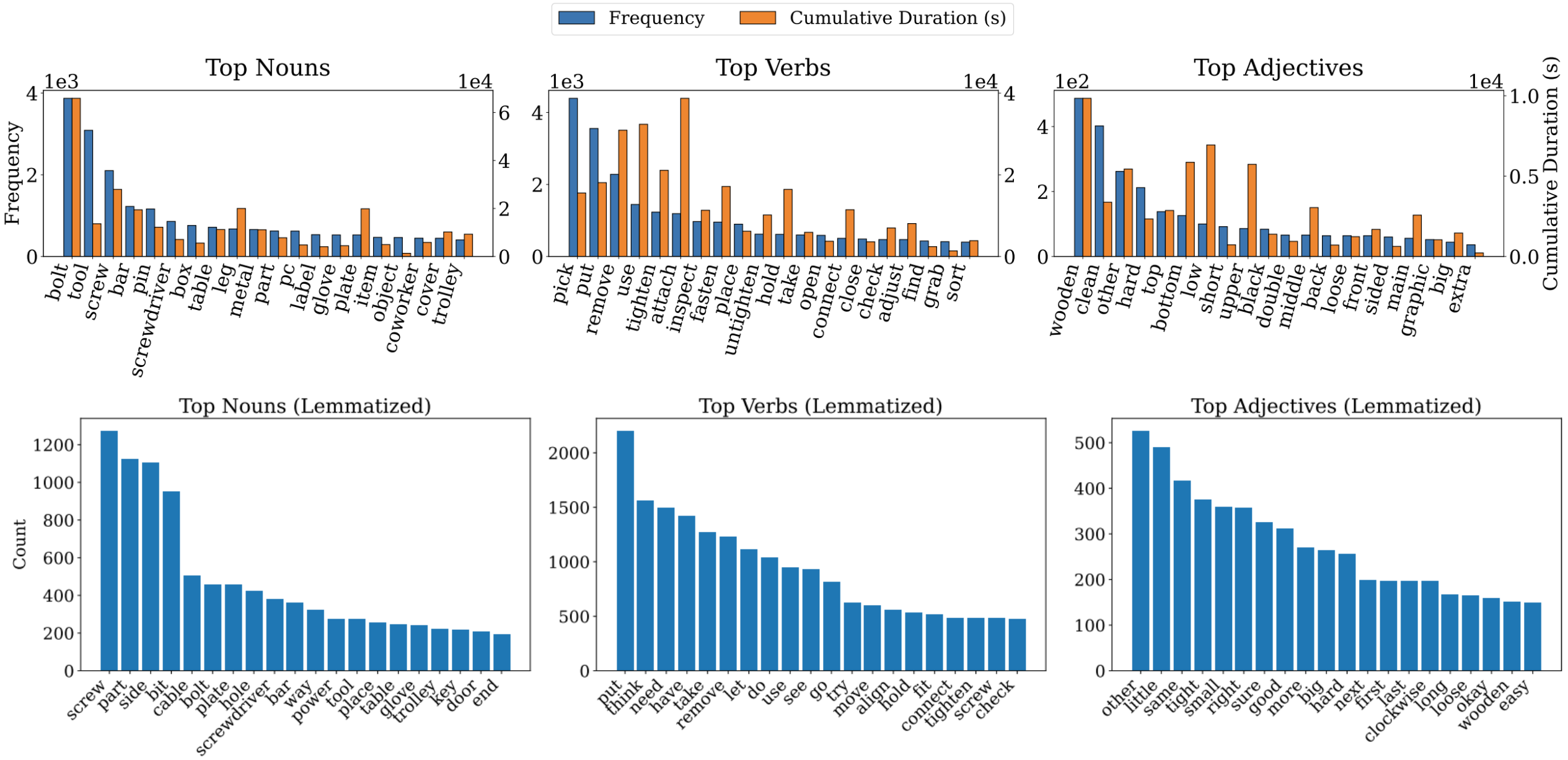
Annotations & Multimodality



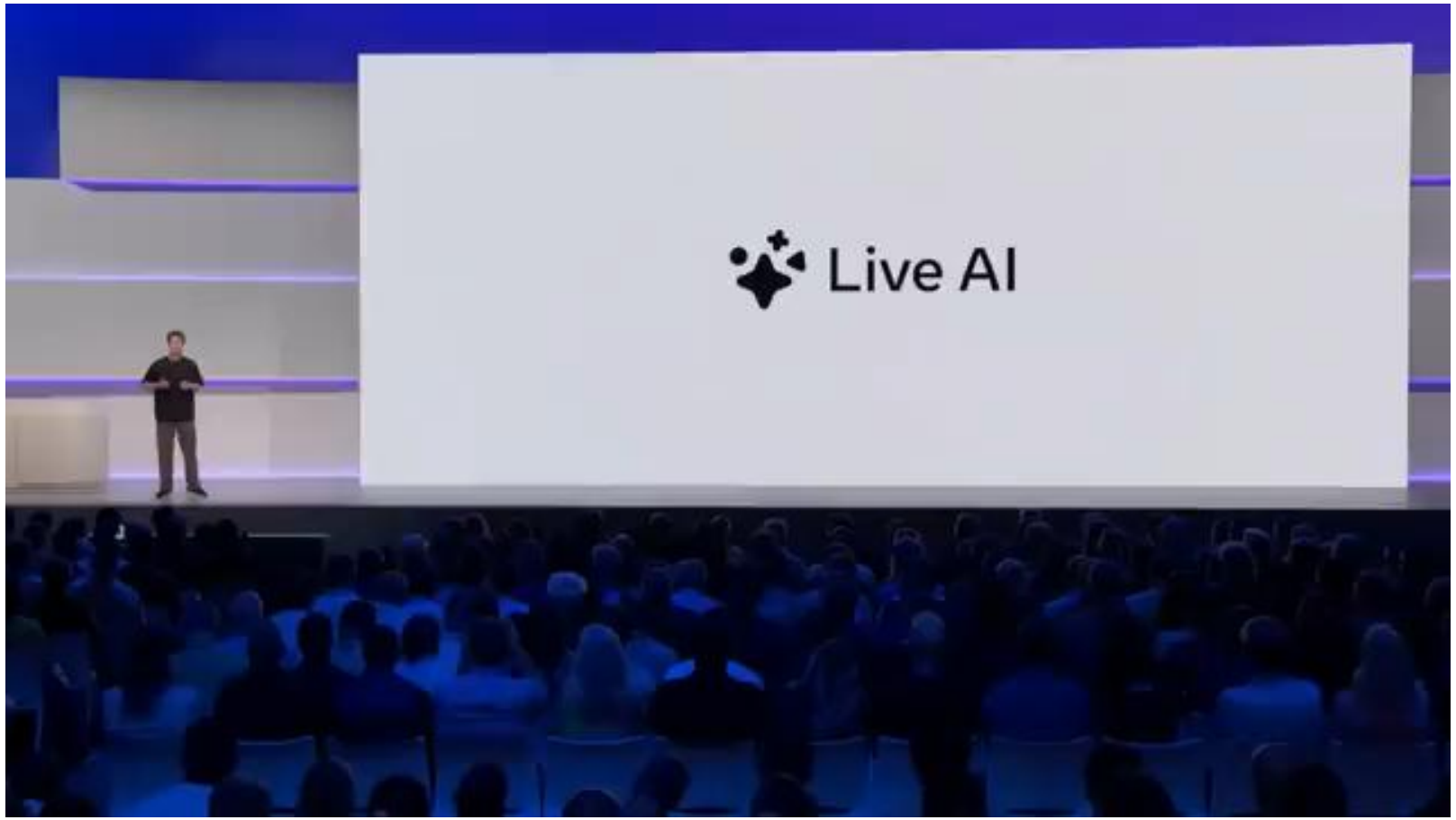
Ego · Exo · Narration/Audio · Gaze · Context · Motion* · Hand Pose* · SLAM*

*Processed

Annotations & Multimodality

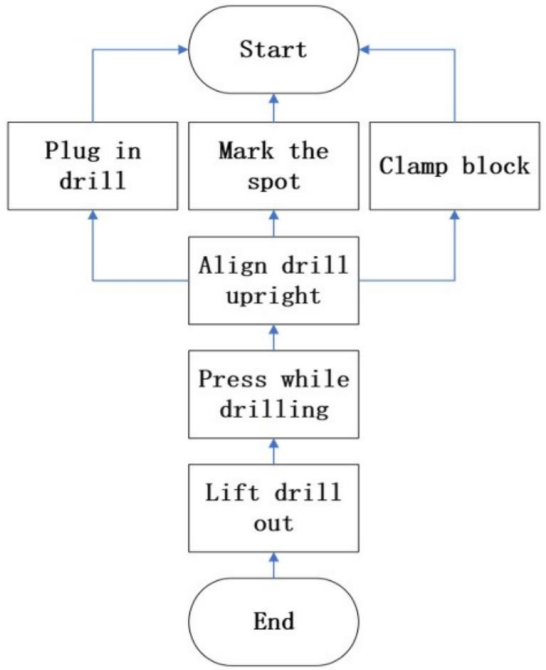


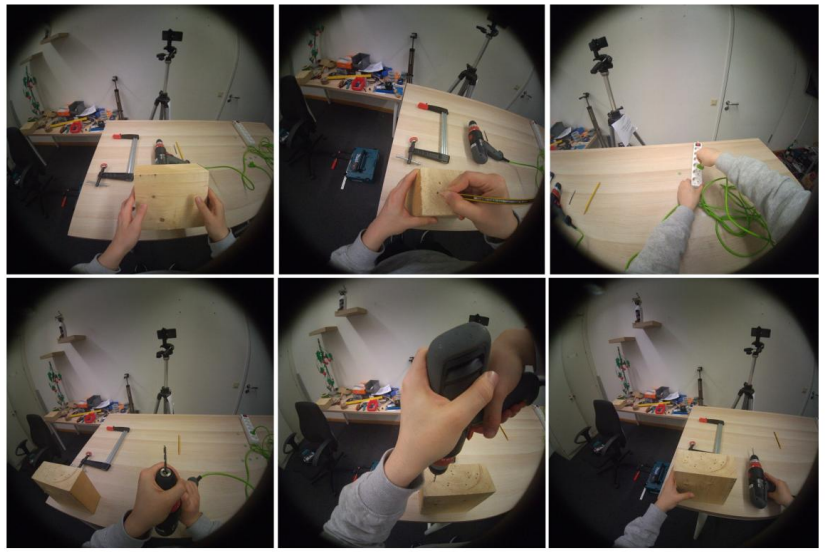
Egocentric Video Understanding



- Current Solution: Video frames processed by VLM.
- Task grounding is necessary to understand the process and guide the user.
- Several other design and engineering challenges exist (e.g. latency, deployment, output modalities).
- However, development and benchmarking of SOTA models on downstream tasks is important.

Benchmark: Mistake Detection





- **S:** Severe Mistakes, e.g. mishandling of a fragile object.
- **PF:** Process Failure, e.g. placing wrong item in a shipment.
- **IF:** Impact to Future Steps, e.g. forget opening a hatch.
- **H:** Risk of harm/injury, e.g. forgetting PPE.

	Approach	P	R	F1	$F1^S$	$F1^{PF}$	$F1^{IF}$	$F1^H$
ZS	VL3 [78]	15.6	46.2	23.3	36.2	38.2	27.4	32.1
	IVL2.5 [13]	16.2	48.2	24.2	38.1	37.1	29.0	33.2
	QVL2.5 [5]	15.9	50.1	24.1	38.8	36.5	28.8	34.1
	GFT* [25]	35.6	48.2	**40.9**	51.2	42.2	34.7	48.0
MLP	VL3 [78]	30.4	56.7	**39.5**	48.1	38.8	32.1	41.3
	IVL2.5 [13]	31.6	50.0	38.7	47.7	39.1	30.5	42.2
	QVL2.5 [5]	31.4	51.6	39.1	42.6	39.8	35.4	44.0
Tr	VL3 [78]	34.5	33.3	33.9	39.2	35.5	29.1	38.5
	IVL2.5 [13]	30.1	41.7	35.5	36.5	38.7	32.1	39.2
	QVL2.5 [5]	33.3	41.0	**36.7**	37.0	39.4	29.5	36.7
MLP	VL3 [78] (EM)	21.3	55.0	30.7	36.2	38.2	30.1	32.2
	IVL2.5 [13] (EM)	23.3	49.2	31.6	35.2.0	32.7	31.6	30.5
	QVL2.5 [5] (EM)	24.1	51.0	**32.7**	34.2	32.0	32.1	40.1

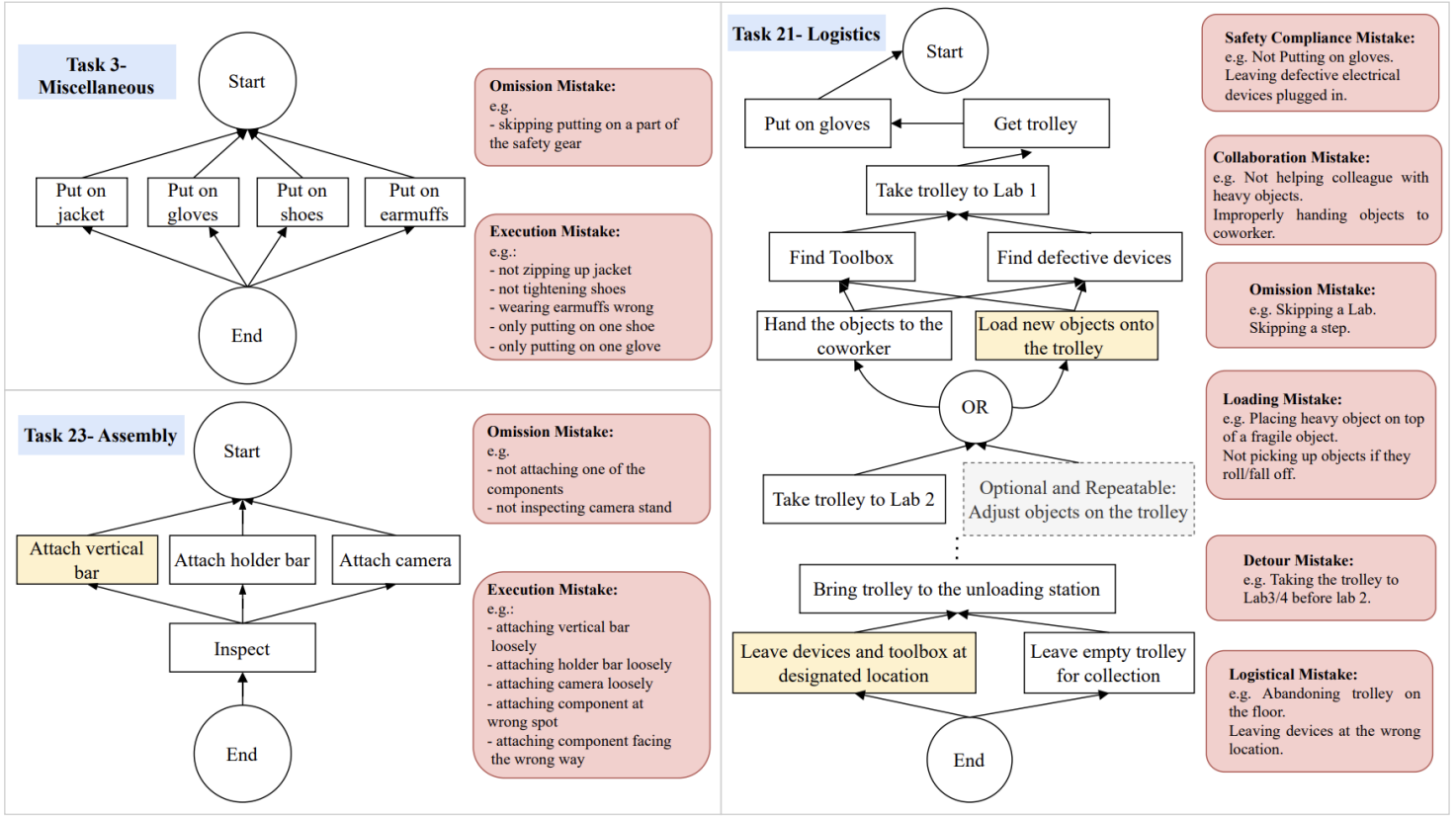
	Approach	P	R	F1	$F1^S$	$F1^{PF}$	$F1^{IF}$	$F1^H$
Ego	VL3 [54]	17.1	48.0	25.2	34.1	37.2	28.4	35.5
	IVL2.5 [55]	18.2	48.7	26.5	32.3	36.1	30.1	34.2
	QVL2.5 [56]	16.5	50.5	24.8	34.1	29.1	30.5	32.0
	GFT* [57]	36.5	47.2	**41.1**	50.1	43.2	33.6	44.5
Exo	VL3 [54]	20.1	44.2	27.6	34.7	34.8	31.2	29.1
	IVL2.5 [55]	18.7	48.8	27.0	37.5	33.3	29.8	32.5
	QVL2.5 [56]	21.1	49.6	29.6	32.4	29.4	31.4	32.6
	GFT* [57]	35.1	51.1	**41.6**	48.5	41.0	34.3	46.6

VL3: Video-LLaMA3 IVL2.5: InternVL2.5 QVL2.5: Qwen2.5-VL GFT: Gemini 2.0 Flash Thinking ZS: Zero-Shot

Mistake Detection

Model	Ego	Exo	Ego + Exo
GFT (ZS) [56]	0.43	0.39	0.44
VL3 + Tr [53]	0.33	0.32	0.37
IVL2.5 + Tr [54]	0.29	0.30	0.33

Ego/Exo Fusion



Intentional and Unintentional Errors in procedural and non-procedural tasks across all scenarios.

Model	RGB only	RGB + Audio	RGB + Gaze	RGB + Audio + Gaze
GFT (ZS) [56]	0.38	0.41	0.39	0.42
VL3 [53]	0.27	0.26	0.28	0.30
IVL2.5 [54]	0.30	0.28	0.29	0.29

Modality Ablation

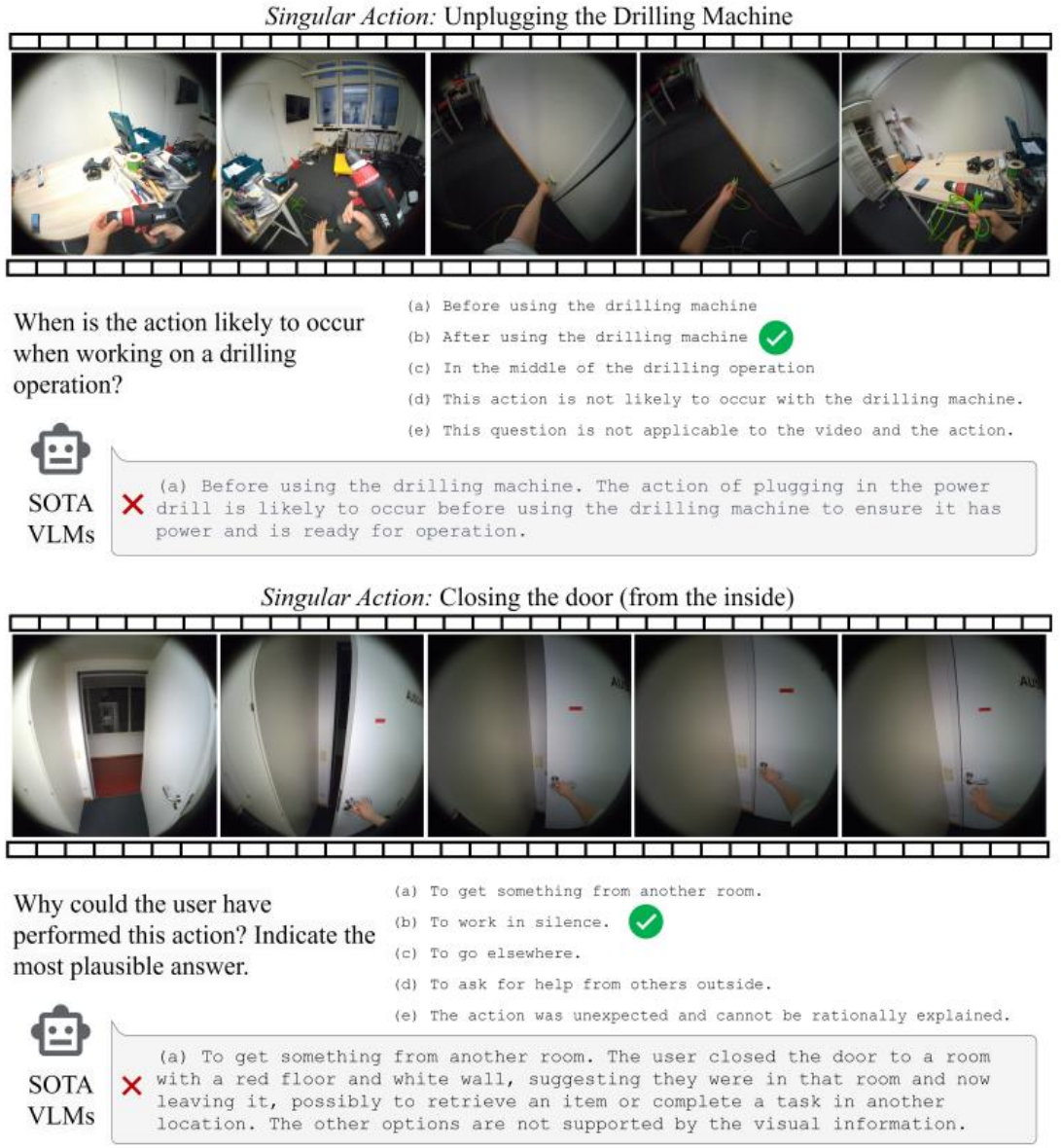
VL3: Video-LLaMA3 IVL2.5: InternVL2.5 QVL2.5: Qwen2.5-VL GFT: Gemini 2.0 Flash Thinking ZS: Zero-Shot

Benchmark: VQA

For Long and Short Tasks

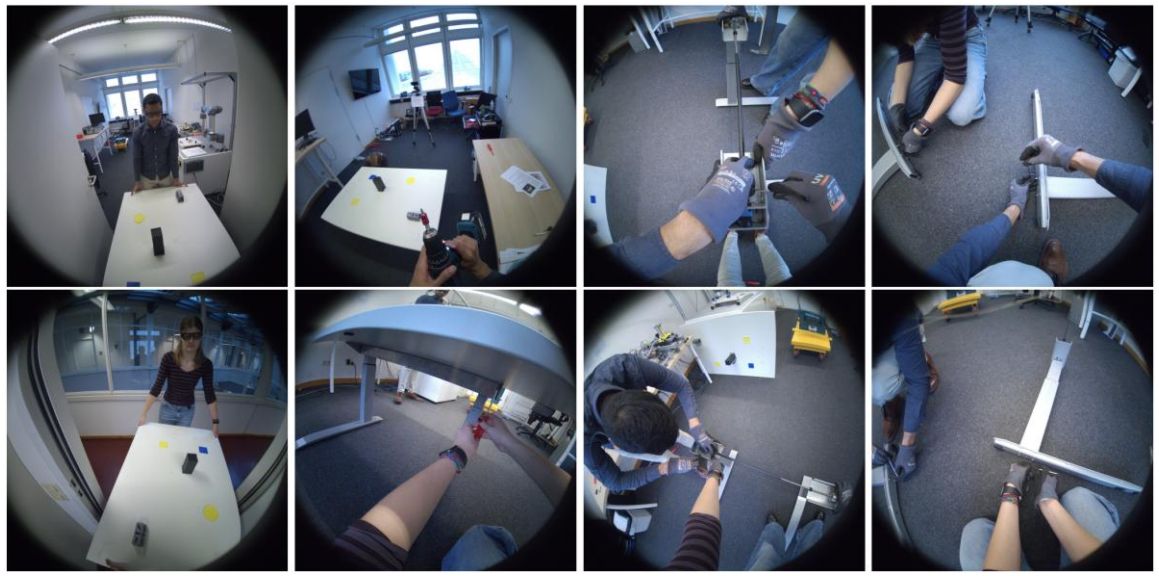
- Temporal Understanding (Tm)
- Situated Reasoning (Si)
- Visual Recognition (Re)
- Analogical/Abductive Reasoning (A)

Model	Acc^{Tm}	Acc^{Si}	Acc^{Re}	Acc^{A}	Acc
VL3 [78]	52.2	60.3	59.4	57.5	58.2
IVL2.5 [13]	51.7	61.1	58.2	56.0	57.6
QVL2.5 [5]	53.2	60.8	59.3	56.5	58.1
GFT* [25]	55.4	62.1	67.2	68.3	64.1
ML2 [47] + Label	92.3	51.4	42.8	78.3	61.4
Human	92.6	89.6	90.4	88.6	90.0



VL3: Video-LLaMA3 IVL2.5: InternVL2.5 QVL2.5: Qwen2.5-VL GFT: Gemini 2.0 Flash Thinking ZS: Zero-Shot

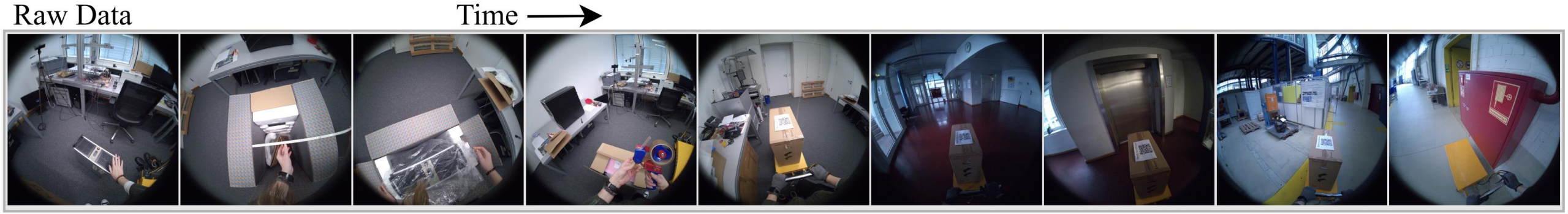
Task Understanding in a Collaborative Setting



- ✓ Anticipate co-worker's action
- ✓ Understand worker's role

GFT: 35.2% (action anticipation)

Summarisation



VL3: Video-LLaMA3 IVL2.5: InternVL2.5 QVL2.5: Qwen2.5-VL GFT: Gemini 2.0 Flash Thinking ZS: Zero-Shot

Project Page: https://indego-dataset.github.io/

Dataset: https://huggingface.co/datasets/FraunhoferIPK/IndEgo

Acknowledgements:

Federal Ministry of Research, Technology and Space

DLR Projektträger

KIKERP (Grant No. 16IS23055C)

INSTITUTE MACHINE TOOLS AND FACTORY MANAGEMENT TECHNISCHE UNIVERSITÄT BERLIN

REALITY LABS

Hugging Face