NEURAL INFORMATION
PROCESSING SYSTEMS

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

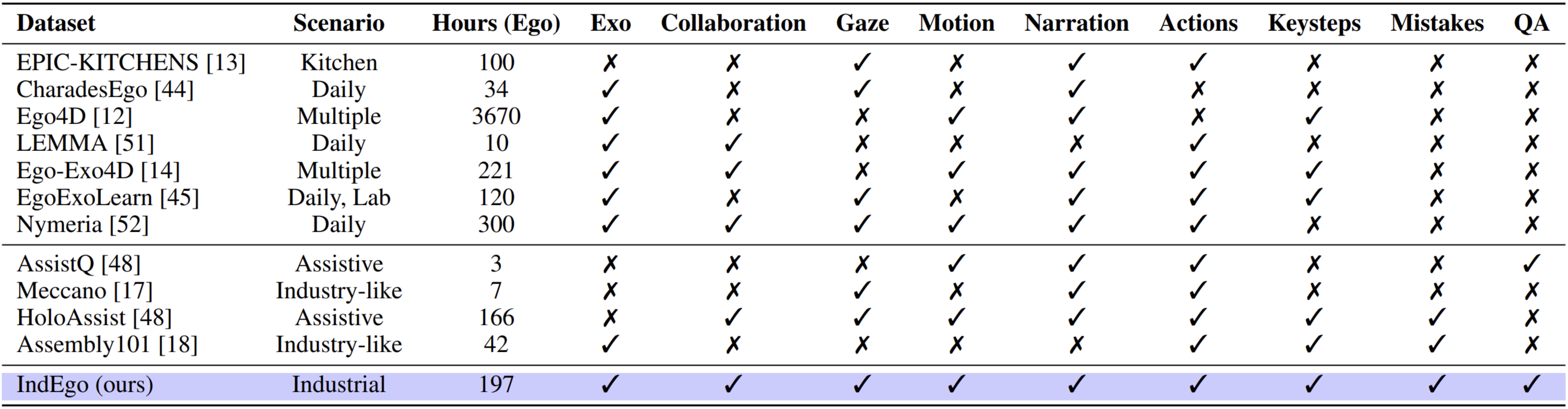
[5] Leibniz Universität Hannover

Motivation & Gaps

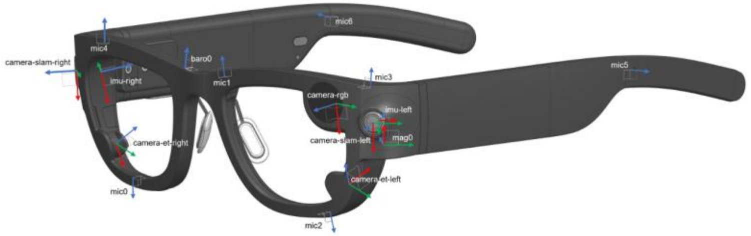
- Industrial settings need intelligent assistants
- Egocentric vision provides a natural interface
- Lack of representative data
- We need datasets that bridge academic vision and industrial use

Gaps:

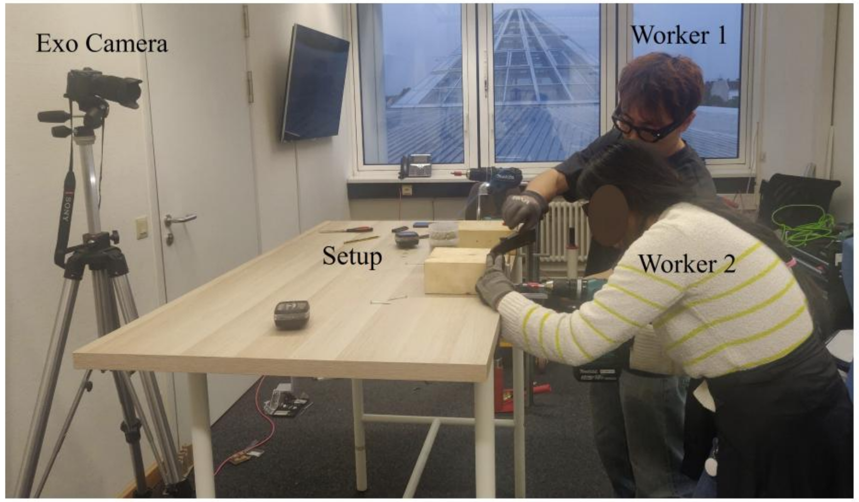
× Long sequence tasks
× Collaborative Scenarios
× Industrial setups and use-cases
× Multimodality in industry



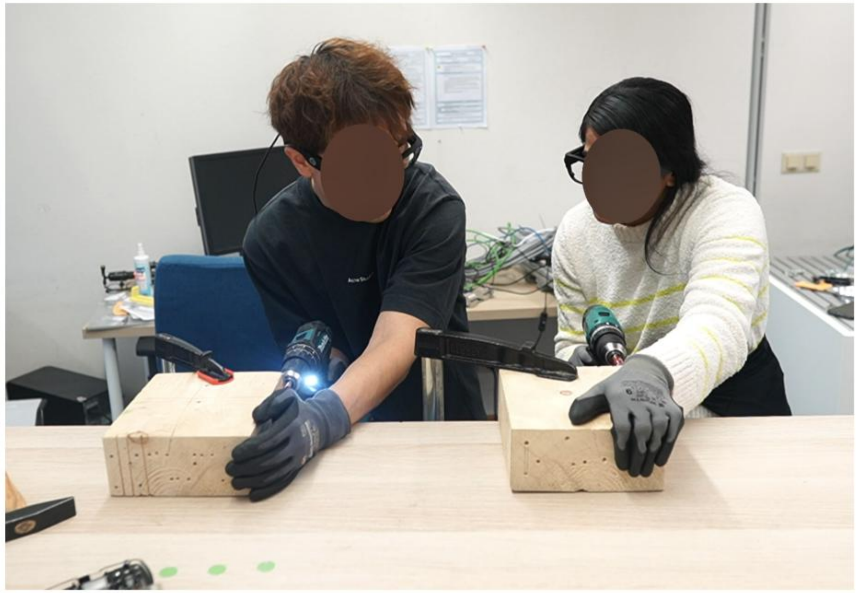
Dataset	Scenario	Hours (Ego)	Exo	Collaboration	Gaze	Motion	Narration	Actions	Keysteps	Mistakes	QA
EPIC-KITCHENS [13]	Kitchen	100	✗	✗	✓	✗	✓	✓	✗	✗	✗
CharadesEgo [44]	Daily	34	✓	✗	✓	✗	✓	✗	✗	✗	✗
Ego4D [12]	Multiple	3670	✓	✗	✗	✓	✓	✗	✓	✗	✗
LEMMA [51]	Daily	10	✓	✓	✗	✗	✗	✓	✗	✗	✗
Ego-Exo4D [14]	Multiple	221	✓	✓	✗	✓	✓	✓	✓	✗	✗
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)

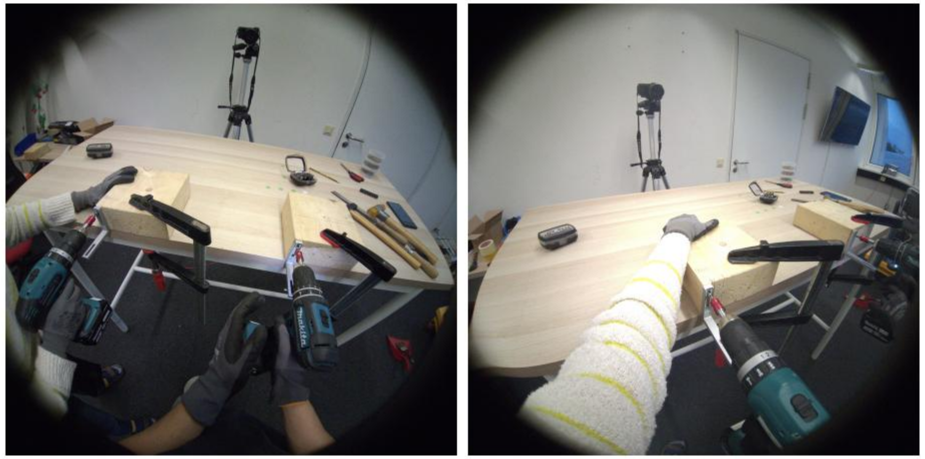


Setup



Exo Perspective

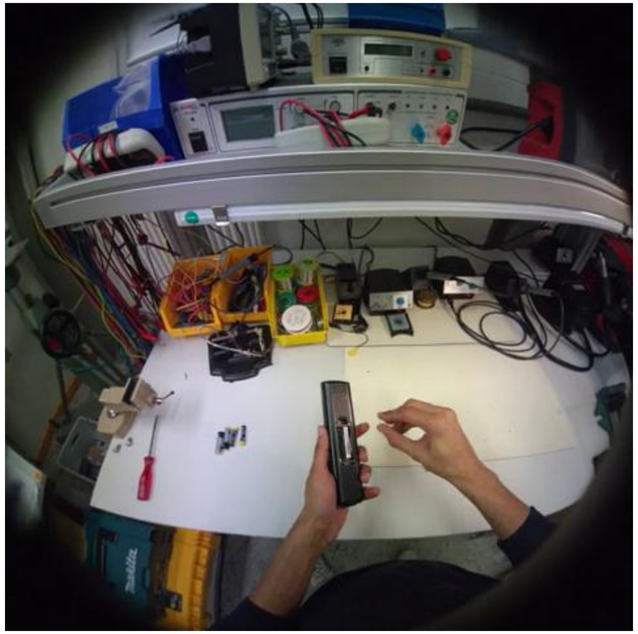
Industrial Tools and Devices



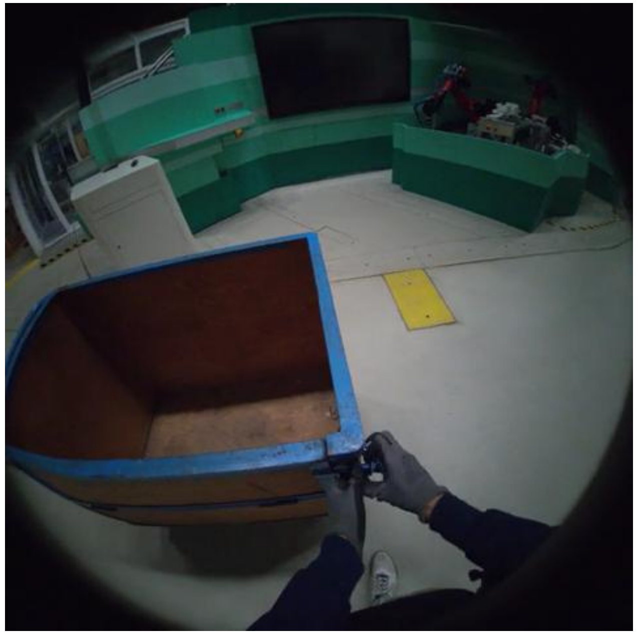
Ego Perspectives

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

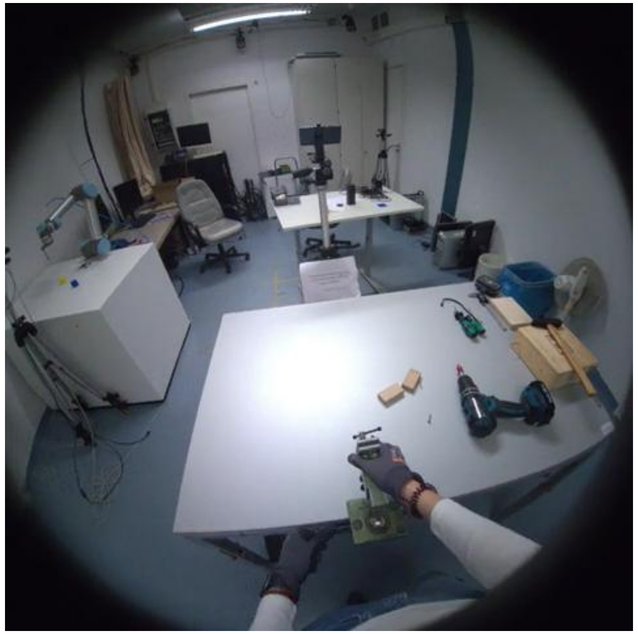
Assembly/Disassembly



Inspection/Repair



Logistics/Organisation

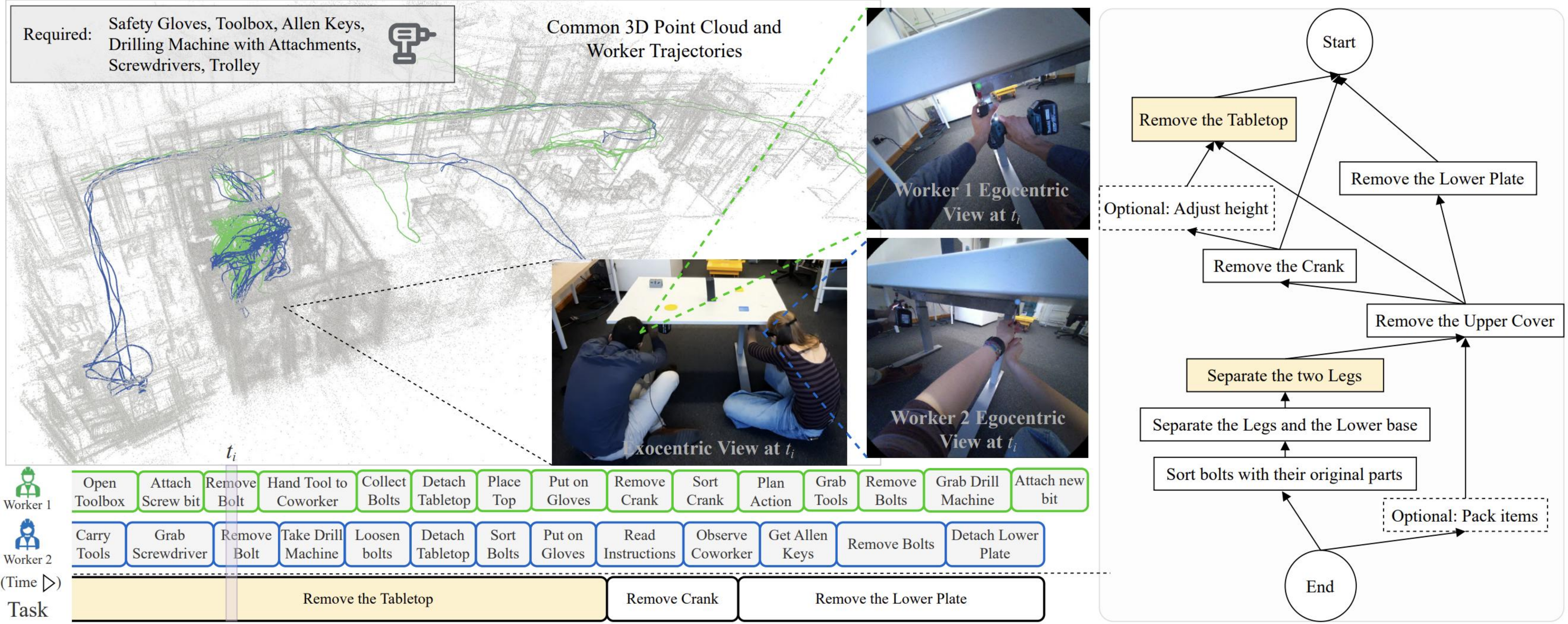


Woodworking

Miscellaneous

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

Annotations & Multimodality



Required: Safety Gloves, Toolbox, Allen Keys, Drilling Machine with Attachments, Screwdrivers, Trolley

Common 3D Point Cloud and Worker Trajectories

Worker 1 Egocentric View at t_i

Worker 2 Egocentric View at t_i

Exocentric View at t_i

t_i

Worker 1: Open Toolbox | Attach Screw bit | Remove Bolt | Hand Tool to Coworker | Collect Bolts | Detach Tabletop | Place Top | Put on Gloves | Remove Crank | Sort Crank | Plan Action | Grab Tools | Remove Bolts | Grab Drill Machine | Attach new bit

Worker 2: Carry Tools | Grab Screwdriver | Remove Bolt | Take Drill Machine | Loosen bolts | Detach Tabletop | Sort Bolts | Put on Gloves | Read Instructions | Observe Coworker | Get Allen Keys | Remove Bolts | Detach Lower Plate

(Time ▷)

Task: Remove the Tabletop | Remove Crank | Remove the Lower Plate

Start

Remove the Tabletop

Remove the Lower Plate

Optional: Adjust height

Remove the Crank

Remove the Upper Cover

Separate the two Legs

Separate the Legs and the Lower base

Sort bolts with their original parts

Optional: Pack items

End

Ego

Exo

Narration/Audio

Gaze

Motion*

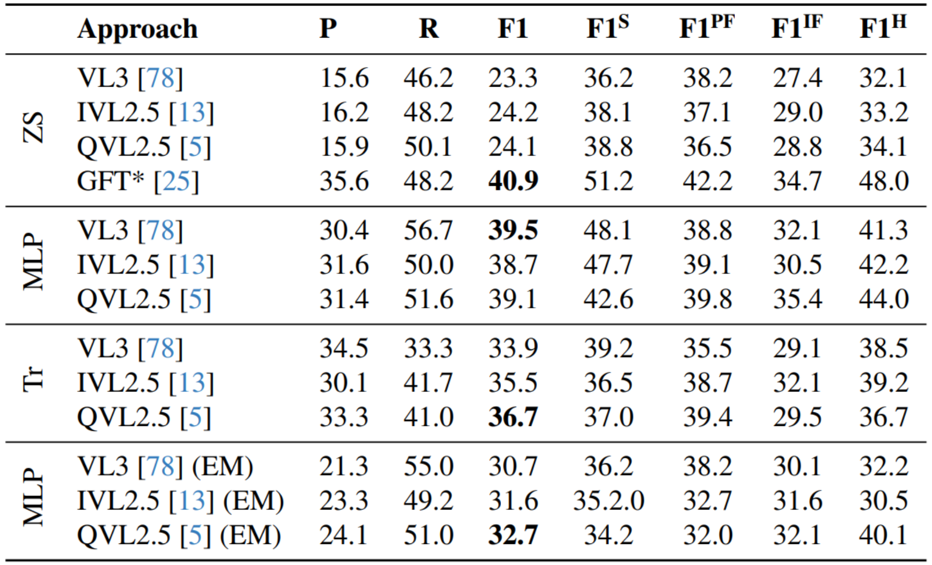
Hand Pose*

SLAM*

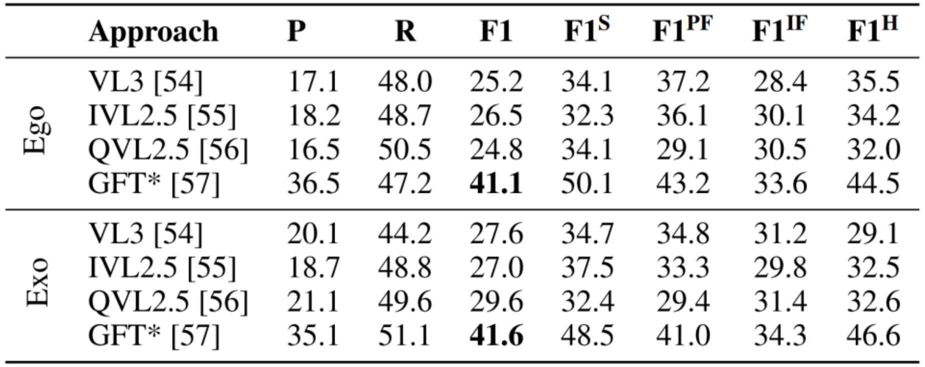
*Processed

Context

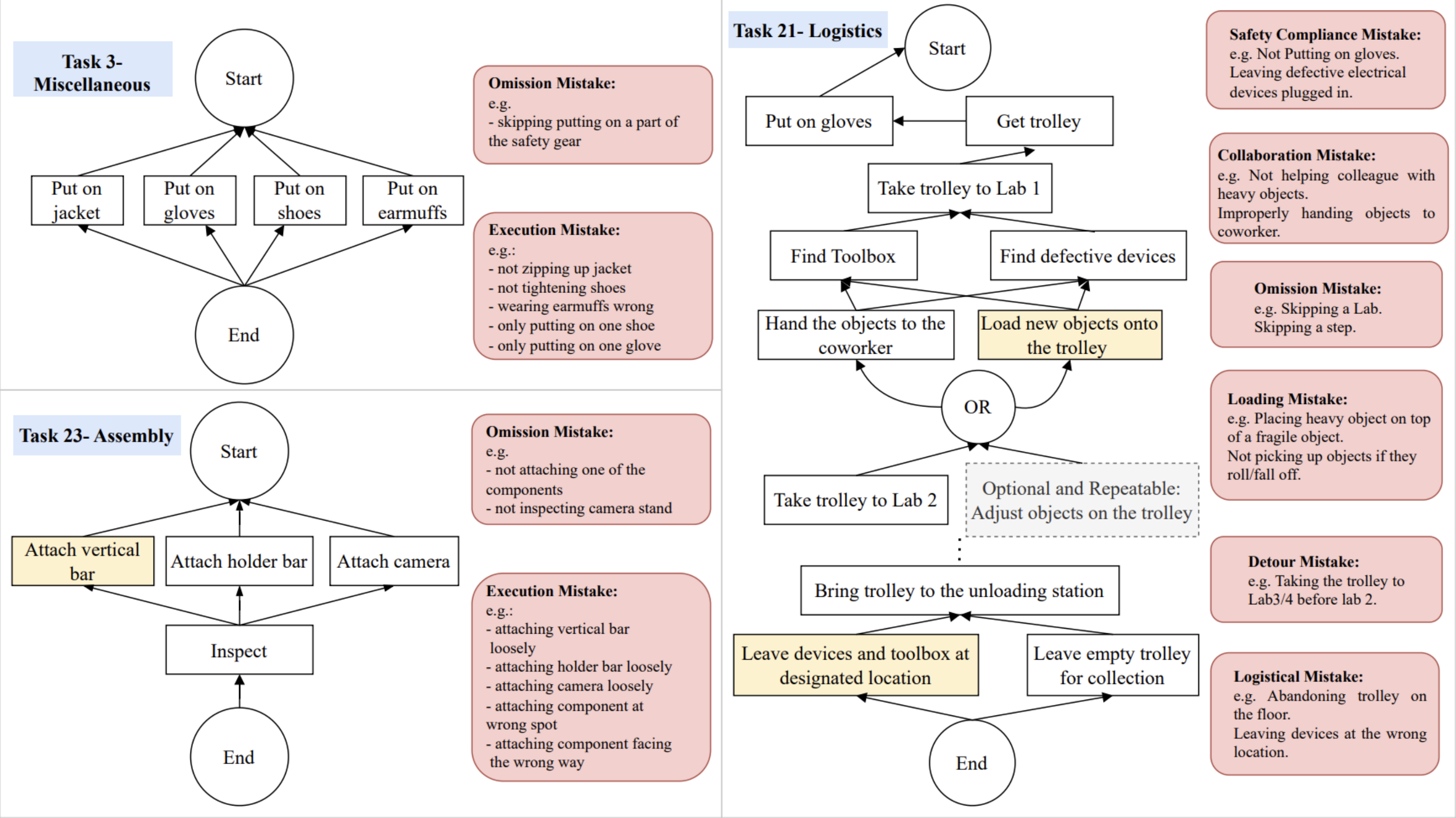
Benchmark: Mistake Detection



	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



Omission Mistake:
e.g.
- skipping putting on a part of the safety gear

Execution Mistake:
e.g.:
- not zipping up jacket
- not tightening shoes
- wearing earmuffs wrong
- only putting on one shoe
- only putting on one glove

Omission Mistake:
e.g.
- not attaching one of the components
- not inspecting camera stand

Execution Mistake:
e.g.:
- attaching vertical bar loosely
- attaching holder bar loosely
- attaching camera loosely
- attaching component at wrong spot
- attaching component facing the wrong way

Safety Compliance Mistake:
e.g. Not Putting on gloves.
Leaving defective electrical devices plugged in.

Collaboration Mistake:
e.g. Not helping colleague with heavy objects.
Improperly handing objects to coworker.

Omission Mistake:
e.g. Skipping a Lab.
Skipping a step.

Loading Mistake:
e.g. Placing heavy object on top of a fragile object.
Not picking up objects if they roll/fall off.

Detour Mistake:
e.g. Taking the trolley to Lab3/4 before lab 2.

Logistical Mistake:
e.g. Abandoning trolley on the floor.
Leaving devices at the wrong location.

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

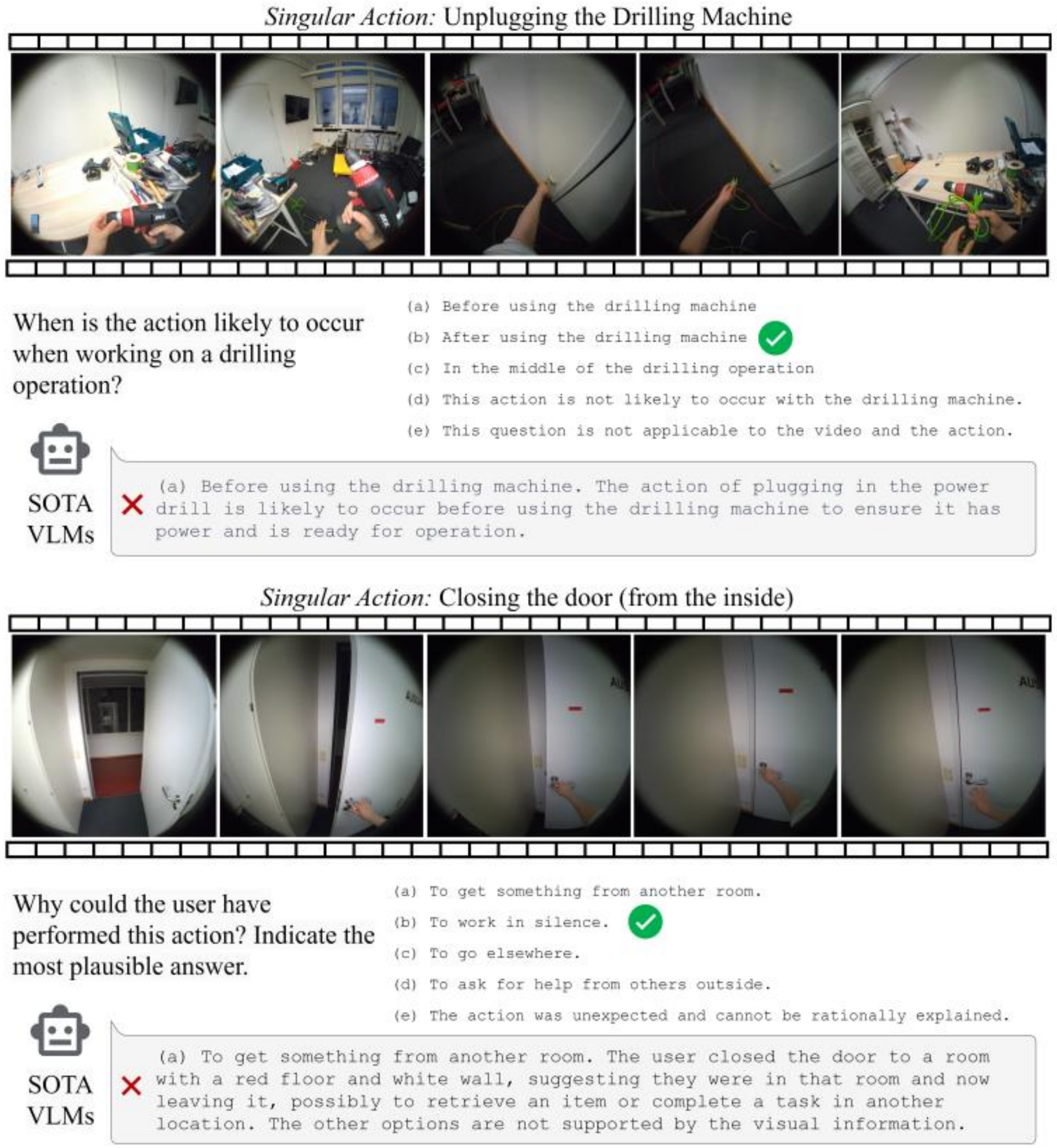
VL3: Video-LLaMA3 IVL2.5: InternVL2.5 QVL2.5: Qwen2.5-VL GFT: Gemini 2.0 Flash Thinking ZS: Zero-Shot

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



Singular Action: Unplugging the Drilling Machine

When is the action likely to occur when working on a drilling operation?

(a) Before using the drilling machine
(b) After using the drilling machine ✓
(c) In the middle of the drilling operation
(d) This action is not likely to occur with the drilling machine.
(e) This question is not applicable to the video and the action.

SOTA VLMs ✗ (a) Before using the drilling machine. The action of plugging in the power drill is likely to occur before using the drilling machine to ensure it has power and is ready for operation.

Singular Action: Closing the door (from the inside)

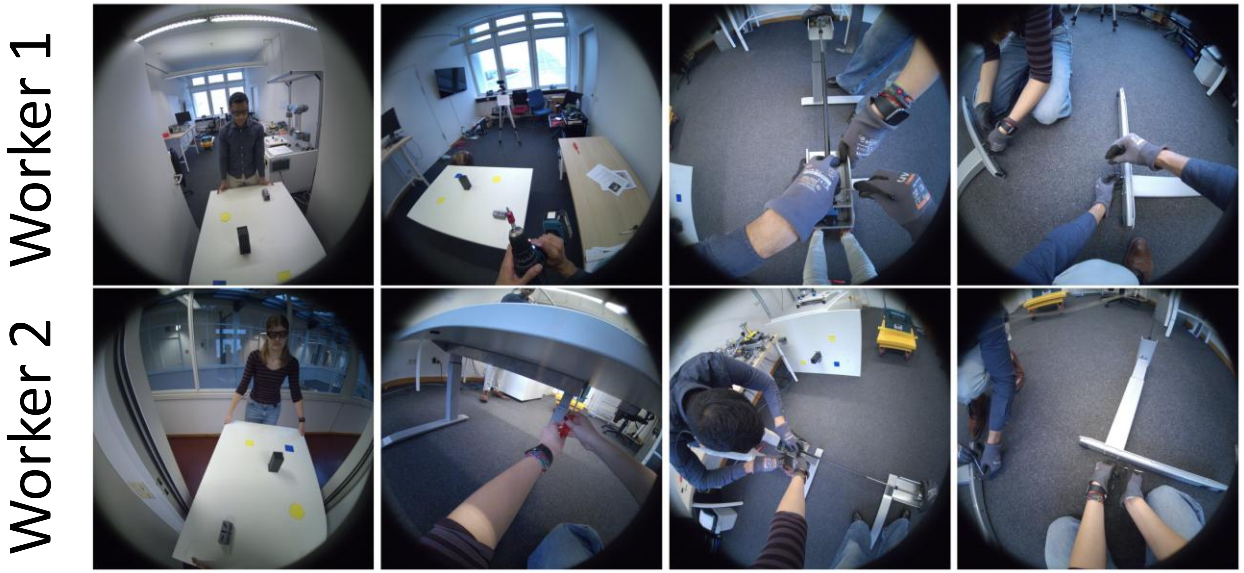
Why could the user have performed this action? Indicate the most plausible answer.

(a) To get something from another room.
(b) To work in silence. ✓
(c) To go elsewhere.
(d) To ask for help from others outside.
(e) The action was unexpected and cannot be rationally explained.

SOTA VLMs ✗ (a) To get something from another room. The user closed the door to a room with a red floor and white wall, suggesting they were in that room and now leaving it, possibly to retrieve an item or complete a task in another location. The other options are not supported by the visual information.

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

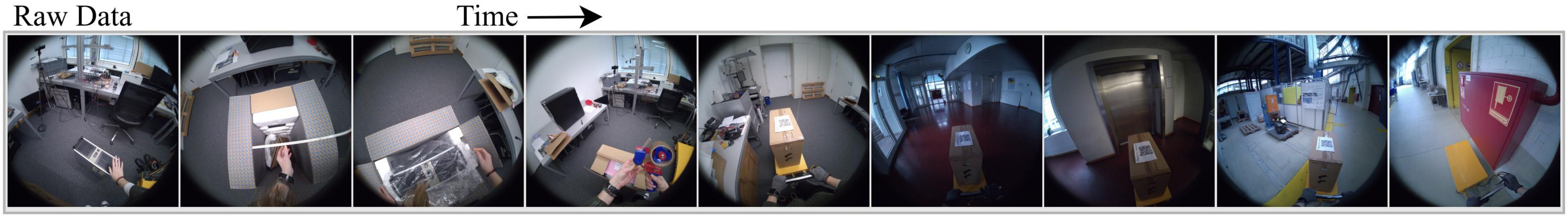
Task Understanding in a Collaborative Setting



- ✓ 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

KIKERP (Grant No. 16IS23055C)

DLR Projektträger

IWF
INSTITUTE MACHINE TOOLS
AND FACTORY MANAGEMENT
TECHNISCHE UNIVERSITÄT BERLIN

REALITY LABS

Hugging Face