Egocentric Vision: Exploring User-Centric Perspectives

Michele Mazzamuto

LIVE Group @ UNICT - https://iplab.dmi.unict.it/live/

Next Vision - http://www.nextvisionlab.it/

Department of Mathematics and Computer Science - University of Catania

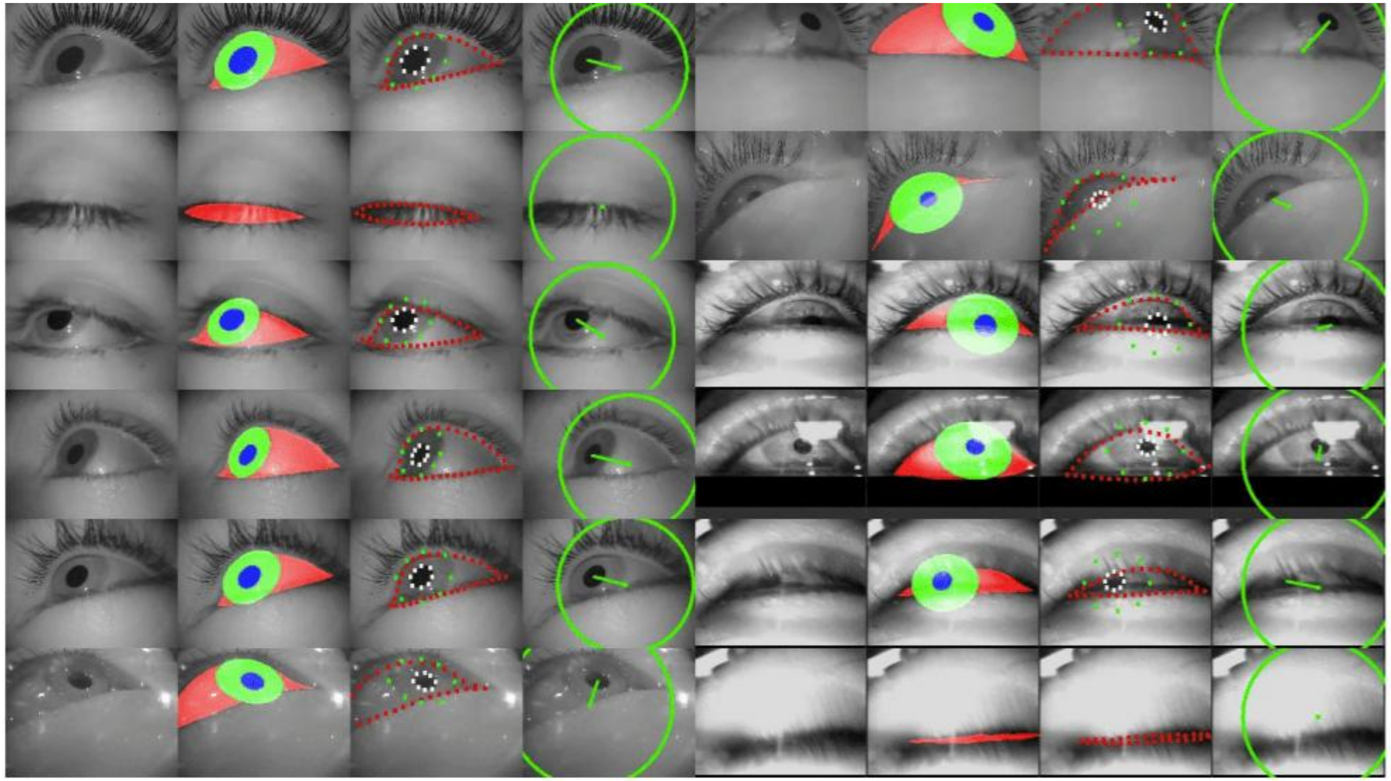
michele.mazzamuto@phd.unict.it - https://mikes95.github.io/

3) Part III: Gaze Understanding and Visual-Language Benchmarks

- Gaze Signal Fundamentals
  - Definitions
  - Tasks
  - Devices
- Gaze-Based Dataset
- Gaze signal in computer vision
  - Gaze prediction
  - Object Referring & Attended object detection
  - Foveation resolution
  - Gaze signal for mistake detection
- Building procedural assistant with VLLM
- Open Challenges and Future Directions

Gaze Definition

Gaze refers to the direction and focus of a person's visual attention, typically measured as the orientation of the eyes or head toward a specific point, object, or region in the environment.



Gaze and Mutual Gaze. Argyle, M., & Cook, M. (1976). Cambridge University Press.
Eye Movements During Scene Viewing: An Overview. Henderson, J. M., & Hollingworth, A. (1998). In *Eye Guidance in Reading and Scene Perception* (pp. 269–295). Elsevier.

Gaze

Gaze as a social signal: Conveys attention, interest, and intentions.

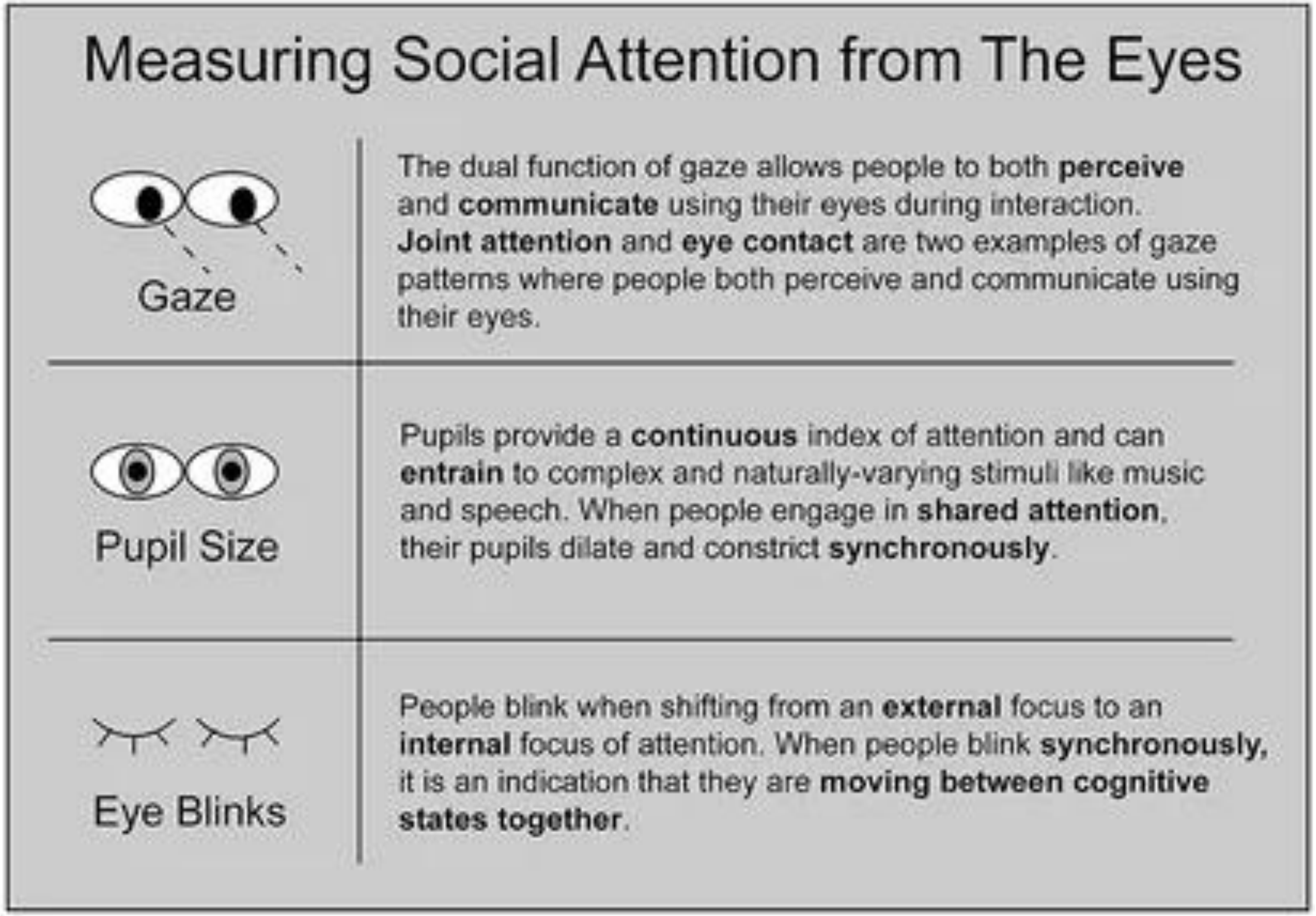
Communication channel: Guides interactions, signals understanding or disagreement.

Information carrier: Indicates objects or people of importance in the environment.

Relevance to Computer Vision: Predicting gaze allows automatic inference of attention and intention, enabling tasks like:
Attended object detection
Procedural assistance
Performance and mistake detection

Gaze and Mutual Gaze. Argyle, M., & Cook, M. (1976). Cambridge University Press.
Eye Movements During Scene Viewing: An Overview. Henderson, J. M., & Hollingworth, A. (1998). In *Eye Guidance in Reading and Scene Perception* (pp. 269–295). Elsevier.

Ocular Signals



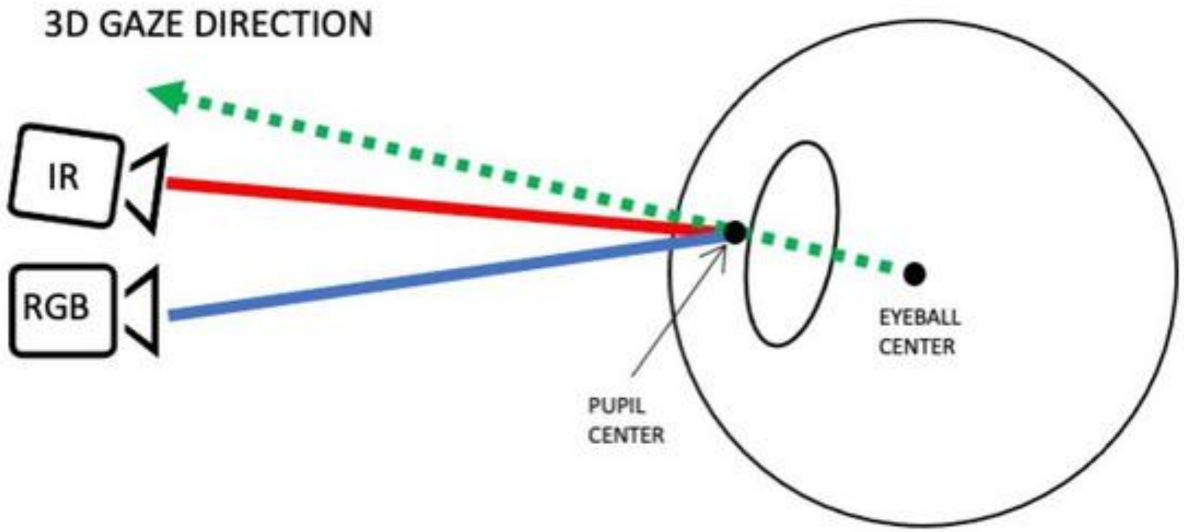
Interpersonal eye-tracking reveals the dynamics of interacting minds. Sophie Wohltjen, Thalia Wheatley. Department of Psychology, University of Wisconsin–Madison.

Gaze - Acquisition

Eye Data Acquisition
All eye-tracking devices use infrared (**IR**) cameras to illuminate the eyes and detect the pupil and corneal (Purkinje) reflections. In HMDs, cameras face inward toward the eyes, while in external trackers they are remote and point at the user. The IR images are processed to extract pupil center and corneal reflections, which are then used to compute gaze direction.

Geometric Eye Modeling
A **2D or 3D eye model** is built linking pupil and reflection positions to the gaze direction.
HMDs often use a **3D model** to combine eye tracking with head pose; external trackers typically use simpler models because the user is stationary in front of a screen.

Mapping in Space
In **AR/VR headsets**: the gaze vector is combined with **head tracking** to obtain a gaze direction in real-world or virtual space. In **external trackers**: the gaze vector is projected onto the **screen** to determine which point the user is looking at.



Eye Tracking Devices in Research

Acquisition Modalities:

- **External / Screen-based trackers** → high precision, lab-based.
- **Wearable glasses** → natural mobility, egocentric datasets.
- **AR/VR headsets** → integrated gaze with rich multimodal sensors.

Research Relevance:

Different devices shape the type of data collected and the realism of tasks.
Most recent egocentric datasets (Ego4D, Ego-Exo4D, HoloAssist, IndustReal, MECCANO) rely on **wearables and AR headsets** with integrated gaze tracking.

External Devices (Screen-based Eye Trackers)

Technology: Infrared (IR) light + high-speed cameras (pupil + corneal reflections).

Examples:

- EyeLink 1000 Plus.
- Tobii bar(desktop mode).

Pros: Sub-degree accuracy, stable in lab, good for controlled experiments.

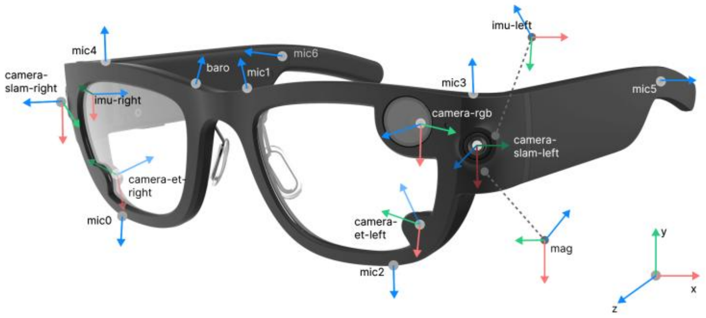
Cons: Not portable, sensitive to head movement, limited to screen-based setups.

Wearable Eye Trackers

Technology: Inward-facing IR cameras in glasses; scene camera records environment.
Examples:

- Pupil Labs Invisible → used in Ego4D (80h of video).
- Tobii Pro Glasses 3 → used in HCI, egonomy.
- ARIA Glass
- SMI Eye-Tracking Glasses

Pros: Natural mobility, supports egocentric datasets, records gaze in real-world contexts.
Cons: Less precise than lab trackers, prone to drift, battery limits.



Mixed Reality & AR/VR Headsets

Technology: Eye cameras embedded in AR/VR headsets; often combined with IMU, SLAM, depth sensors.

Examples:

- HoloLens 2.
- Meta Aria Glasses.
- Meta Quest Pro.

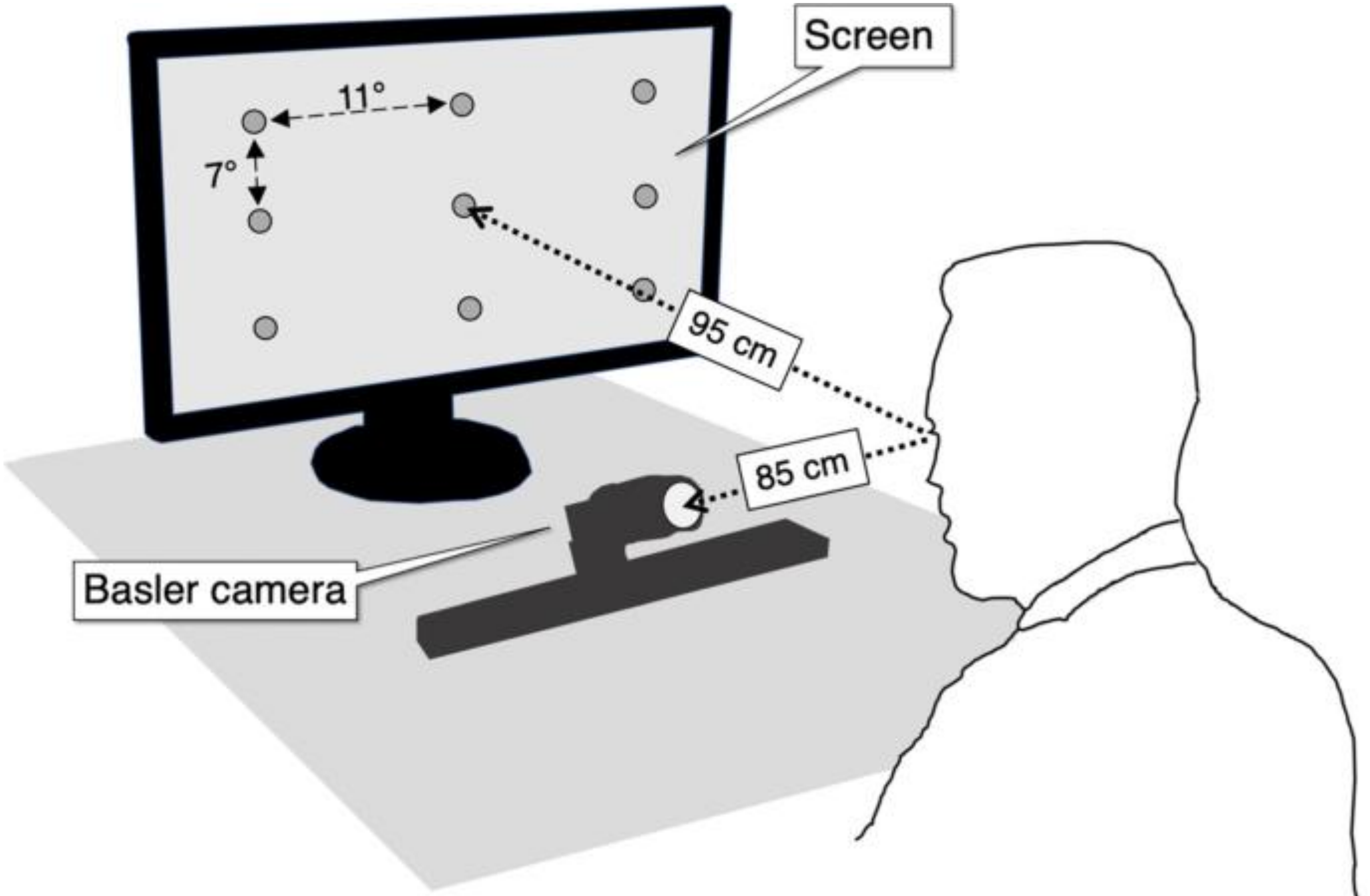
Pros: Rich multimodal data (video, IMU, depth, gaze), natural for egocentric vision.

Cons: Drift, calibration issues with glasses wearers, heavier hardware.

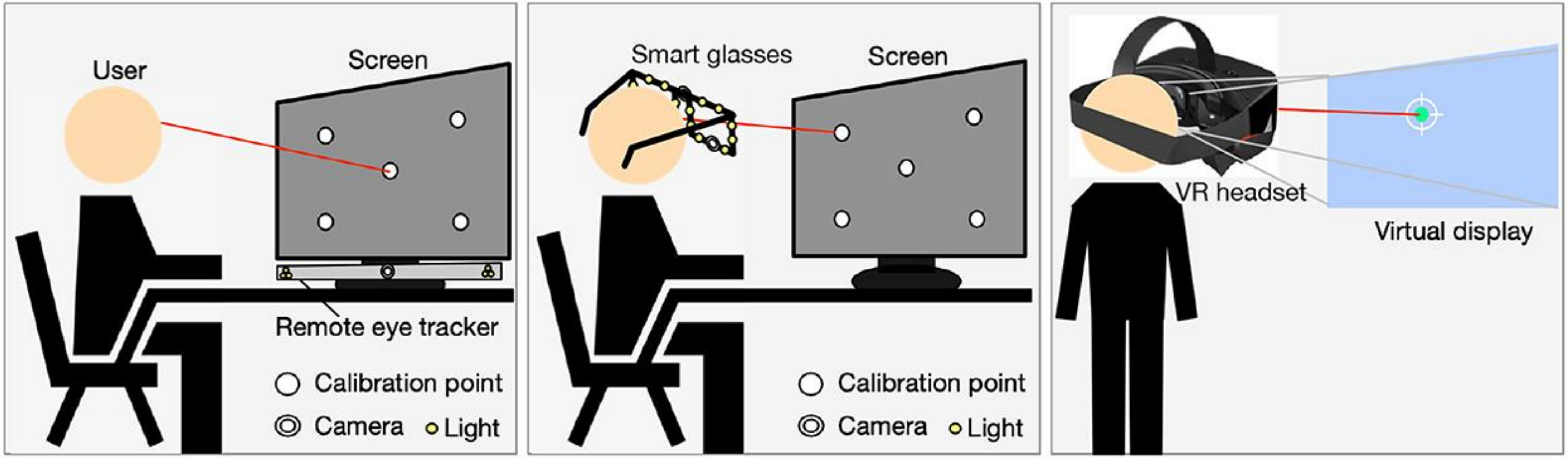
Calibration

All systems require **initial calibration**, where the user looks at known target points.
Calibration compensates for individual differences (eye shape, distance to camera, headset alignment).

Gaze calibration works by having a user follow moving targets on a screen, allowing the eye-tracking system to measure and map individual eye characteristics and the specific way light reflects from their eye



Calibration



A review on personal calibration issues for video-oculographic-based gaze tracking

Gaze-Based Datasets in Egocentric Vision

Gaze-Based Datasets in Egocentric Vision

- Gaze datasets provide ground-truth annotations aligned with egocentric video.
- They are essential for training and evaluating gaze estimation and gaze prediction models.
- Several benchmark datasets have been proposed, often tied to specific tasks (e.g., cooking, daily activities, long-form interaction).

2012 – GTEA Gaze

- 14 subjects performing 17 sequences of meal prep tasks.
- Actions annotated with verb-noun pairs (e.g., “pour milk into cup”).
- Gaze recorded using Tobii glasses, 15 fps extraction.
- Frame-level action annotations for train/test splits.

GTEA Gaze

A. Fathi, Y. Li, J.M. Rehg: “Learning to Recognize Daily Actions Using Gaze.” ECCV, 2012.

2015 – GTEA Gaze+

- 26 subjects performing 7 meal-prep activities, 37 videos.
- HD video (24 fps) and gaze (30 Hz) recorded with SMI glasses.
- Actions annotated using ELAN; audio available on request.
- Supports fine-grained action recognition and gaze analysis.

GTEA Gaze+

Y. Li, Z. Ye, J.M. Rehg: "Delving into Egocentric Actions." CVPR, 2015.

2018 – EGTEA Gaze+

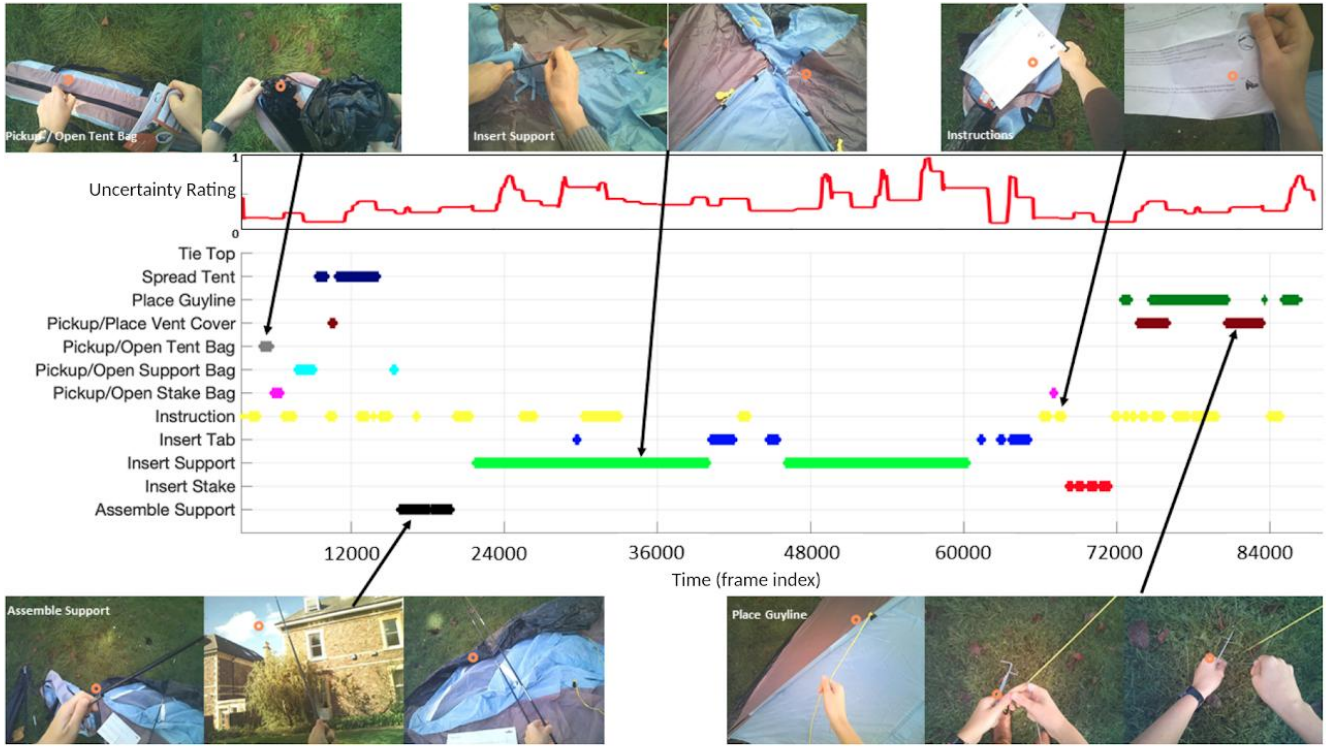
- 32 subjects performing 86 cooking sessions, 28 hours of video.
- HD video (1280×960), gaze tracking (30 Hz), and audio.
- 10,325 action instances and 15,176 hand masks annotated.
- Supports large-scale gaze-action modeling and skill assessment.

EGTEA Gaze+

Y. Li, M. Liu, J.M. Rehg: "In the Eye of the Beholder: Joint Learning of Gaze and Actions in First-Person Video." ECCV, 2018.

2019 – EPIC-Tent

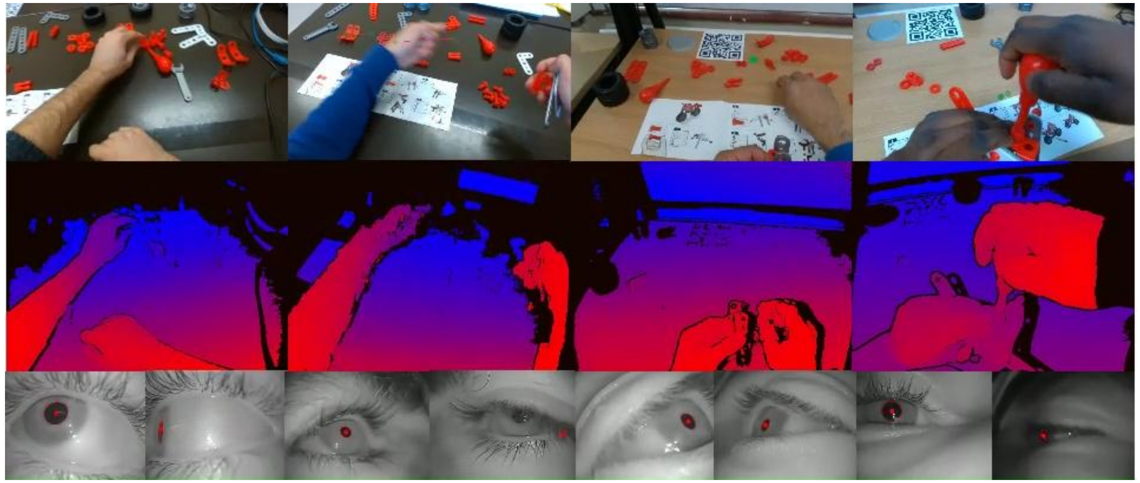
- 5.4h egocentric video from 24 participants assembling a camping tent
- Dual cameras: GoPro + SMI eye tracker (gaze at 30–60Hz)
- Annotated with 38 sub-tasks, 12 main tasks, 8 error types, and uncertainty
- Participants with varying skill levels performing non-rigid object tasks

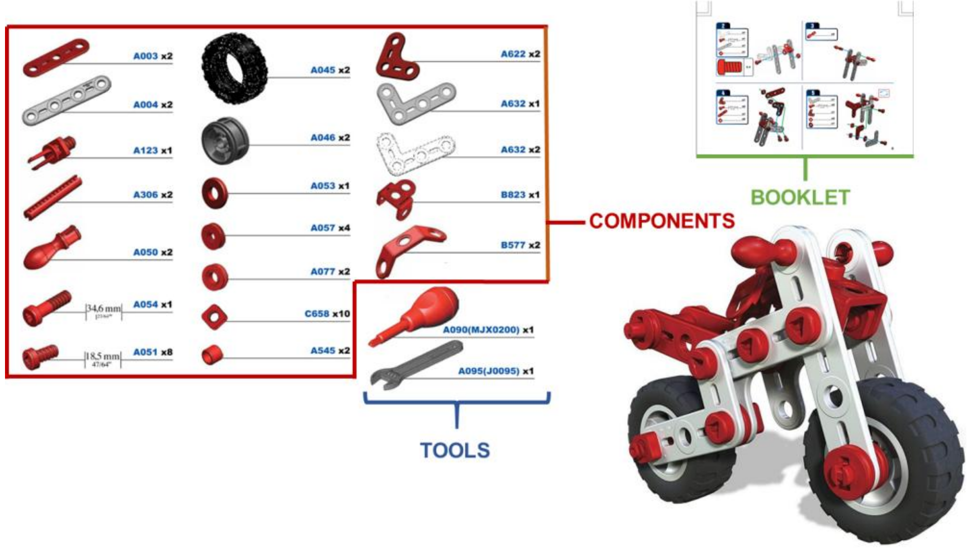


Y. Jang, B. Sullivan, C. Ludwig, I.D. Gilchrist, D. Damen, W. Mayol-Cuevas: "EPIC-Tent: An Egocentric Video Dataset for Camping Tent Assembly." ICCVW, 2019.

2021/2022 – MECCANO

- Multimodal egocentric dataset for industrial-like human behavior understanding.
- **Modalities & Scale:** RGB (1920×1080, 12 fps), depth (640×480, 12 fps), and high-frequency gaze (200 Hz) from 20 participants, 299K frames.
- **Tasks & Annotations:** 61 action classes, 20 object classes, EHOI, gaze estimation, action anticipation, next-active object detection. 8857 action segments and 64K active objects annotated.
- **Gaze in Benchmarks:** Used in action recognition (RGB/Depth/Gaze combinations) and action anticipation (RULSTM with gaze branch) to capture attention and predict future actions.





F. Ragusa, A. Furnari, S. Livatino, G.M. Farinella: "The MECCANO Dataset: Understanding Human-Object Interactions From Egocentric Videos in an Industrial-Like Domain." WACV, 2021.

2022 – EGO4D

A large egocentric video dataset with **3,670 hours** of daily-life activities from **931 wearers** across **74 locations in 9 countries**. It covers diverse scenarios (home, work, outdoor, leisure) with strong privacy safeguards. Additionally, **80 hours** include eye-gaze data collected using Pupil Labs wearable trackers.

Gaze data is integral to the Social Interaction Benchmark, particularly the "Looking at Me" (LAM) task, which classifies if social partners are looking at the camera wearer. It also underpins future tasks like Egocentric Attention Prediction (EAP) and Social Gaze Prediction (SGP), expanding research on eye contact and social gaze

K. Grauman et al.: "Ego4D: Around the World in 3,000 Hours of Egocentric Video." CVPR, 2022.

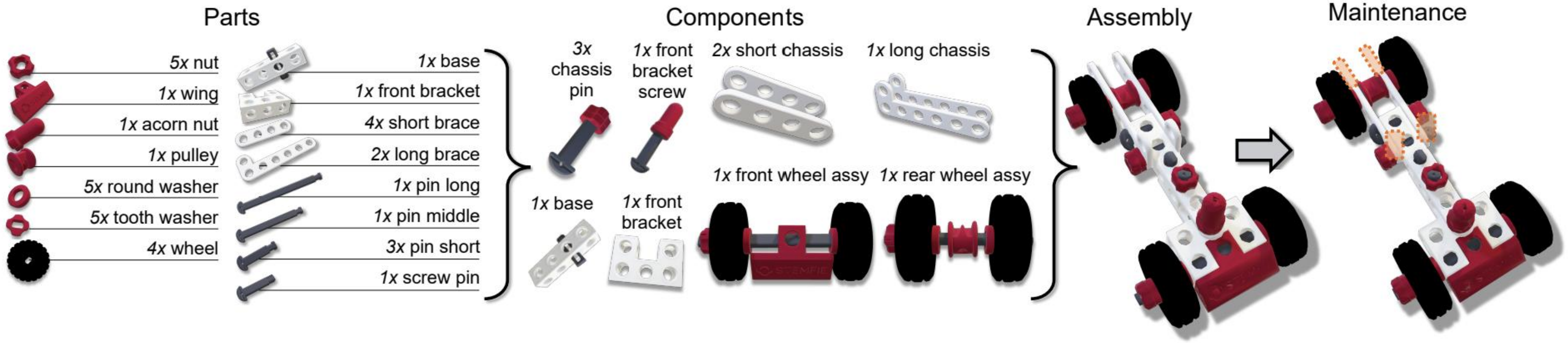
2023 - HoloAssist

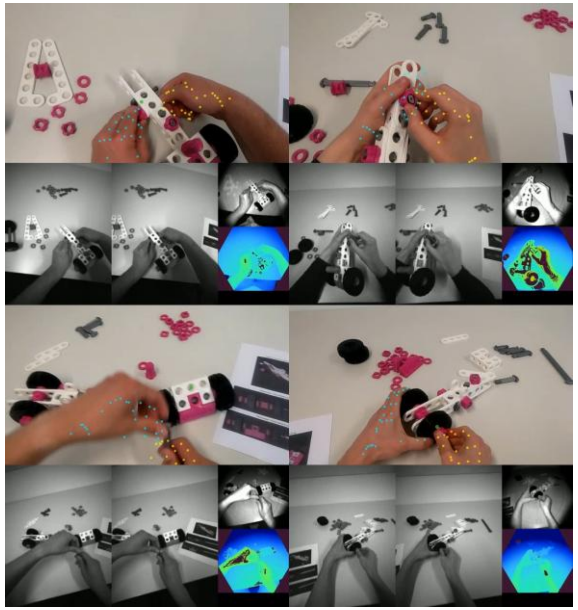
- 166 hours of data, 222 participants, 350 instructor–performer pairs
- 7 synchronized modalities: RGB, depth, head pose, 3D hand pose, eye gaze, audio, IMU
- 20 object-centric manipulation tasks with third-person annotations (actions, mistakes, interventions)
- Eye gaze crucial for predicting intentions and anticipating actions; provides largest boost for intervention prediction

X. Wang, T. Kwon, M. Rad, B. Pan, I. Chakraborty, S. Andrist, D. Bohus, A. Feniello, B. Tekin, F.V. Frujeri, N. Joshi, M. Pollefeys: "HoloAssist: An Egocentric Human Interaction Dataset for Interactive AI Assistants in the Real World." ICCV, 2023.

2024 - IndustReal

- Build on MECCANO
- Novel Task: Introduces Procedure Step Recognition (PSR) to track correct completion and order of procedural steps, complementing AR & ASD.
- Dataset Stats: 27 participants, egocentric multi-modal (HoloLens 2), 48 flexible execution orders, rich error annotations, open-source 3D parts.
- Benchmarks: Baselines for AR, ASD, PSR using SlowFast, MViTv2, YOLOv8-m.
- Gaze: Recorded as a modality but not used in current experiments.

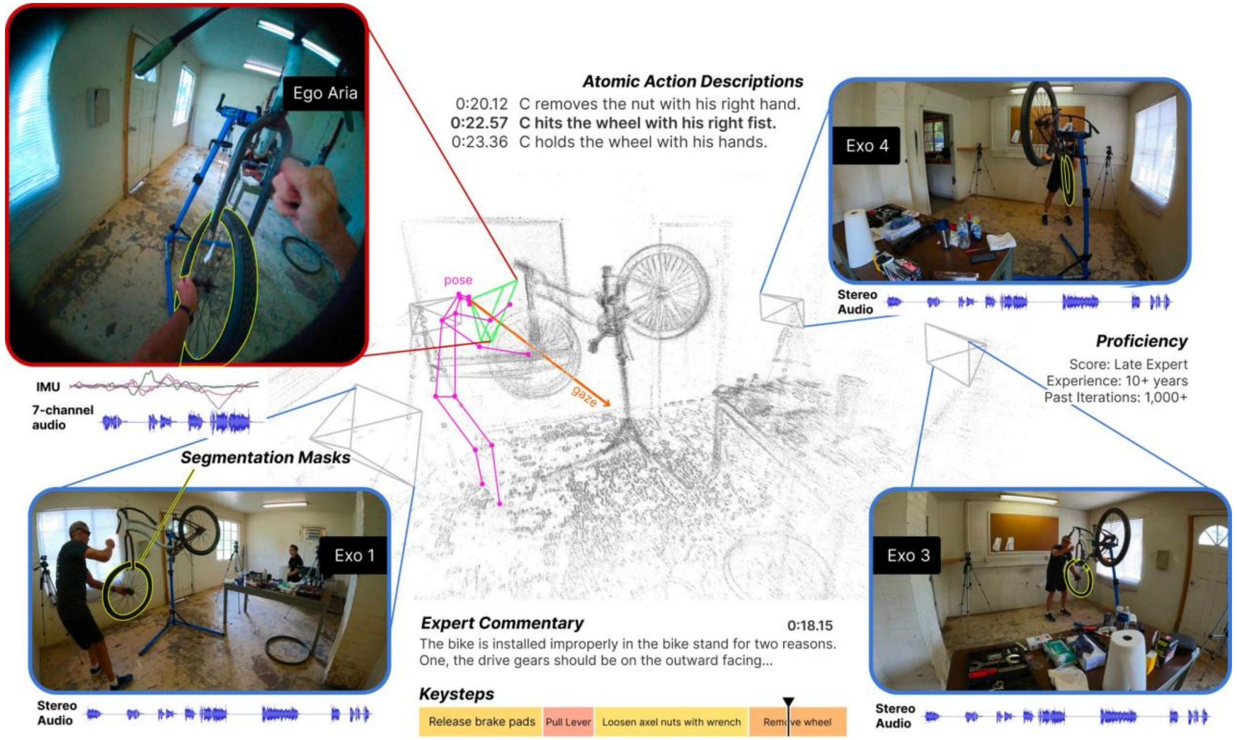




T.J. Schoonbeek, T. Houben, H. Onvlee, F. van der Sommen et al.: "IndustReal: A Dataset for Procedure Step Recognition Handling Execution Errors in Egocentric Videos in an Industrial-Like Setting." WACV, 2024.

2024 – Ego-Exo4D

- Large-scale, multimodal, multiview: 1,286 hours, 740 participants, 13 cities, 123 scenes.
- Sensors via Aria Glasses: video, multichannel audio, IMU, 3D point clouds, camera poses, and eye gaze.
- Gaze Capture: Two monochrome eye-tracking cameras at 10 fps (320×240), with pre-computed 2D gaze points and optional calibration.
- Benchmark Role: Gaze is captured extensively but excluded from inference for core tasks (keystep recognition, proficiency, ego pose), though it can be leveraged for training.



K. Grauman et al.: "Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives." arXiv, 2023.

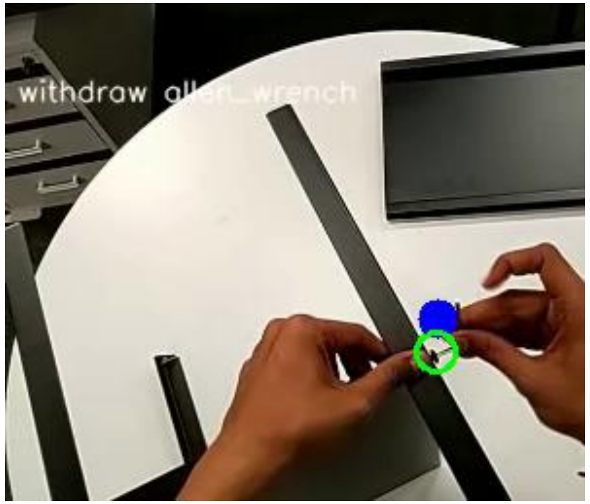
Gaze signal in computer vision

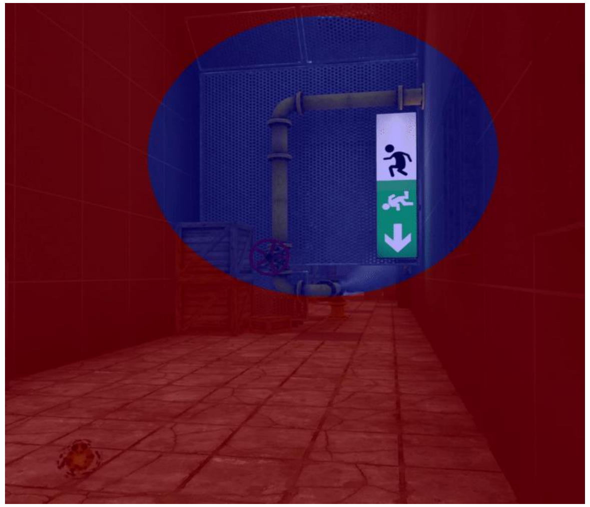
Gaze signal in computer vision

The gaze signal represents the direction and focus of a person's visual attention.
In computer vision, analyzing gaze allows us to understand what, when, and how a person observes a scene or object.

Applications include:

- Gaze Prediction
- Object Referring and Attended Object Detection – identifying the objects being focused on
- Gaze-based Mistake Detection – spotting errors or anomalies in interaction
- Foveation resolution representation

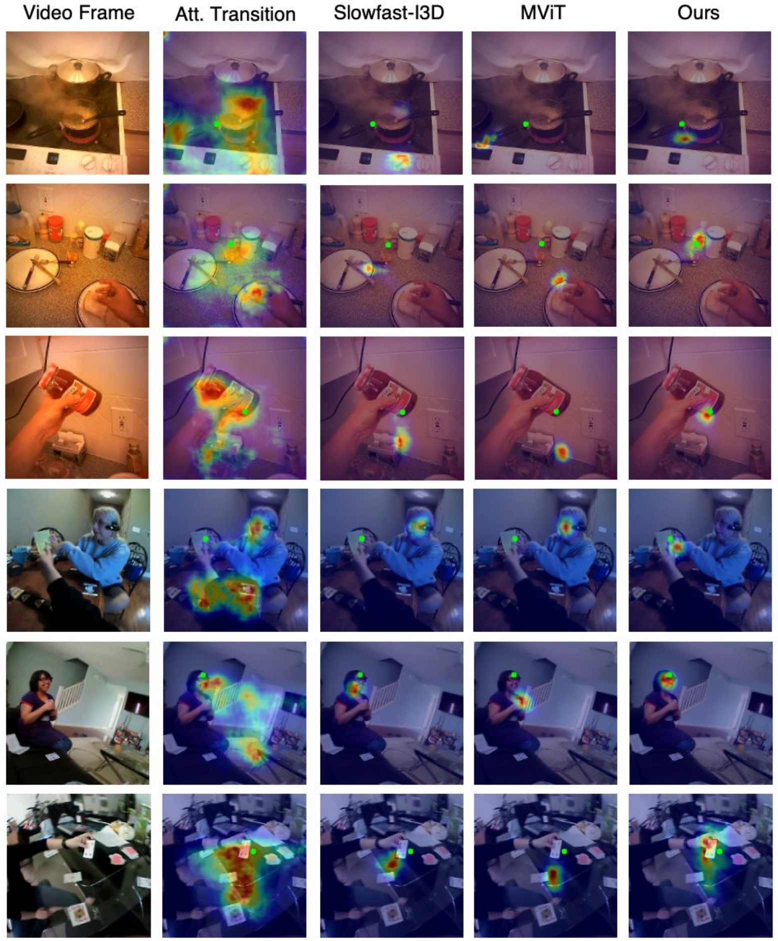




Gaze - Prediction

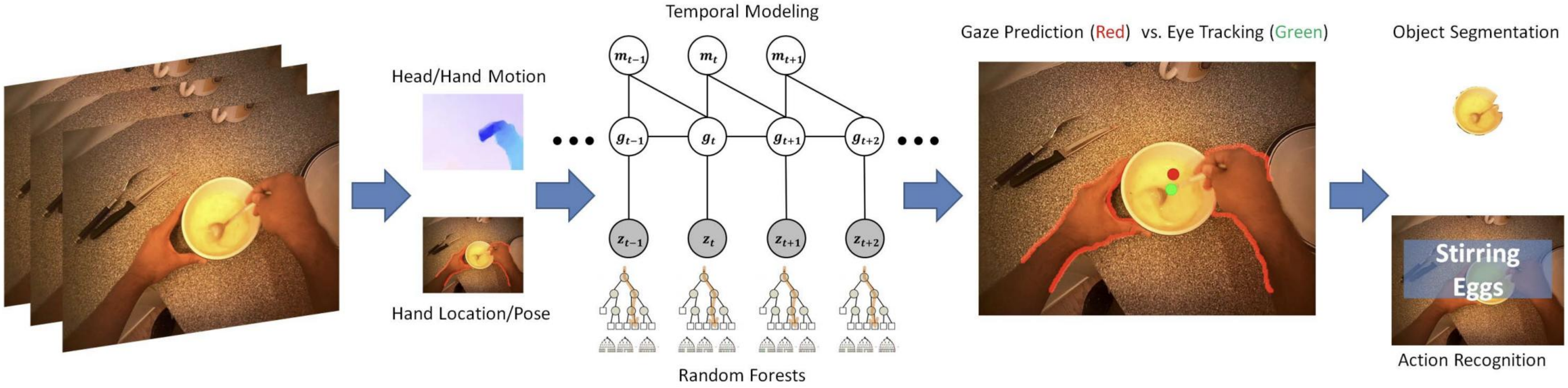
Exocentric Gaze estimation

Egocentric Gaze estimation



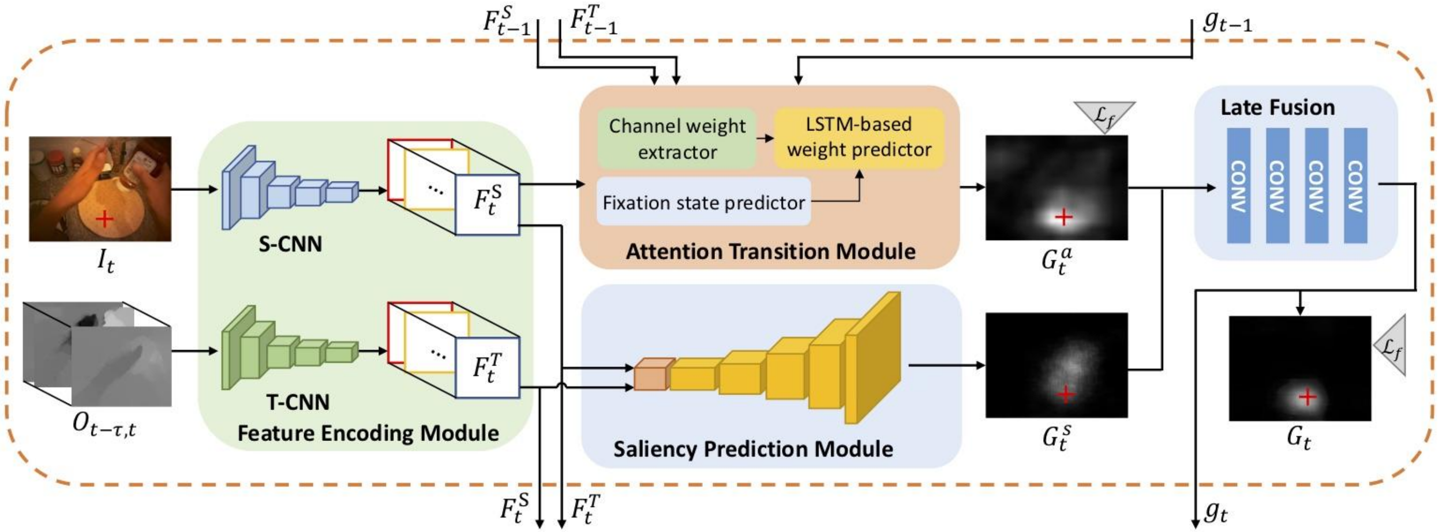
Gaze – Estimation - Egocentric

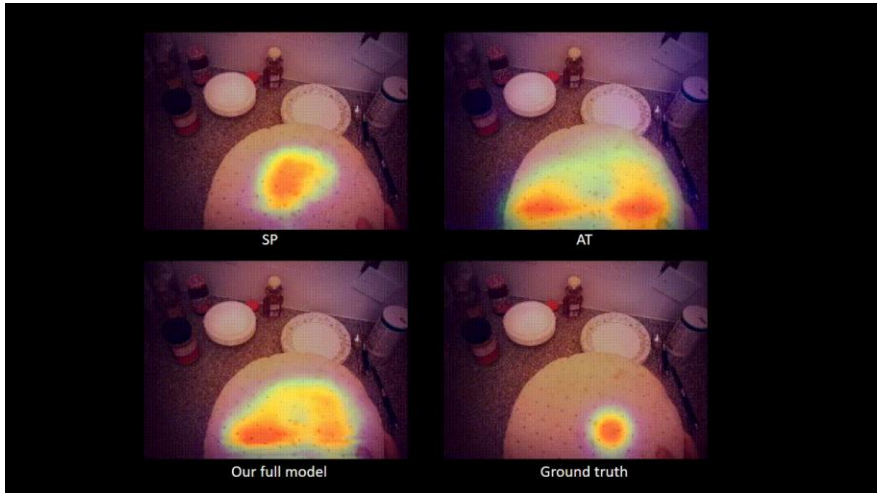
Li et al. (2013) predict egocentric gaze by modeling the coordination of head, hand, and eye movements using egocentric video alone, combining single-frame random forest predictions with temporal fixation modeling for accurate gaze estimation.



Yin Li, Alireza Fathi, James M. Rehg, Learning to Predict Gaze in Egocentric Video ICCV 2013

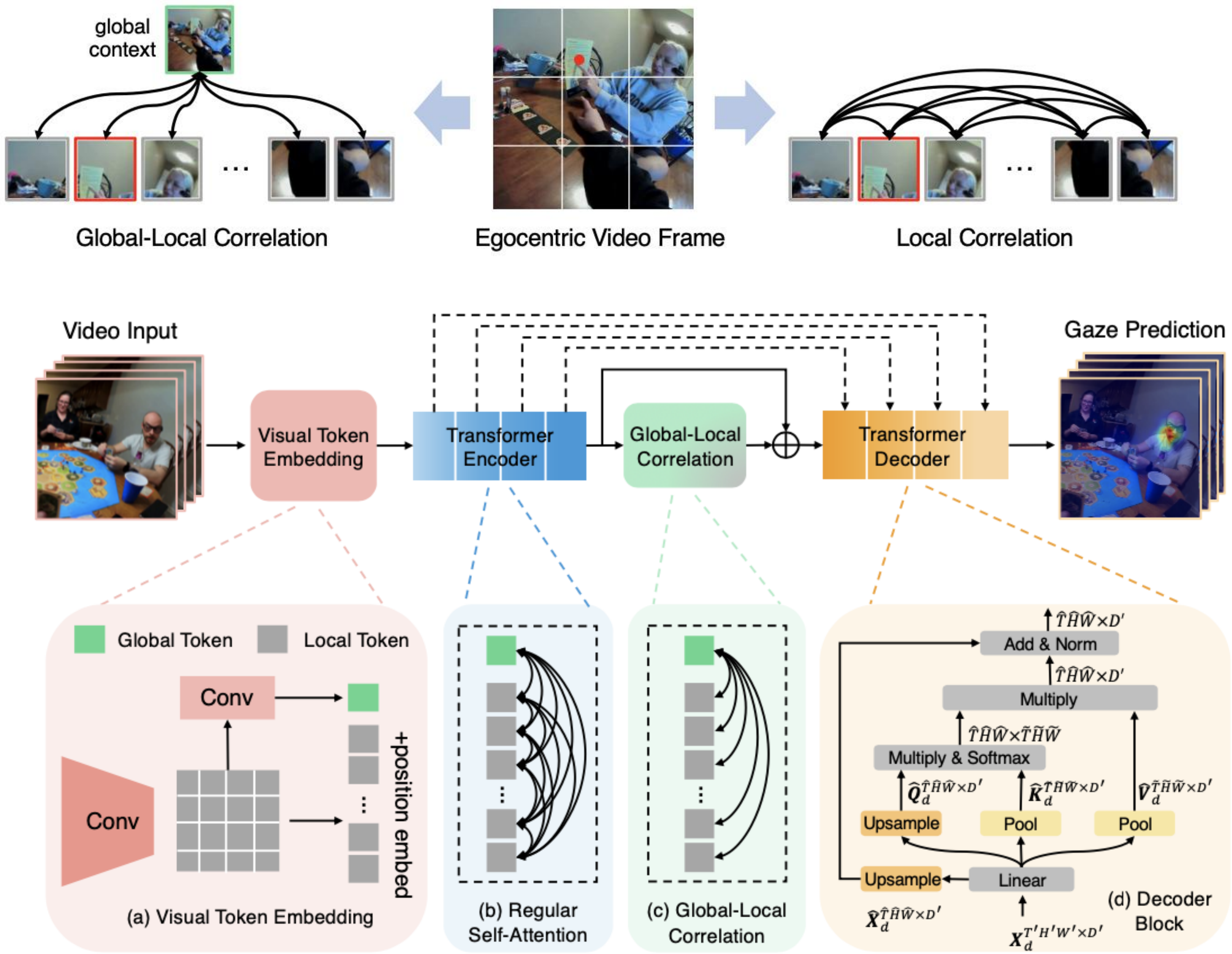
Gaze – Estimation - Egocentric





Y. Huang, M. Cai, Z. Li and Y. Sato, "Predicting Gaze in Egocentric Video by Learning Task-dependent Attention Transition," European Conference on Computer Vision (ECCV), 2018.

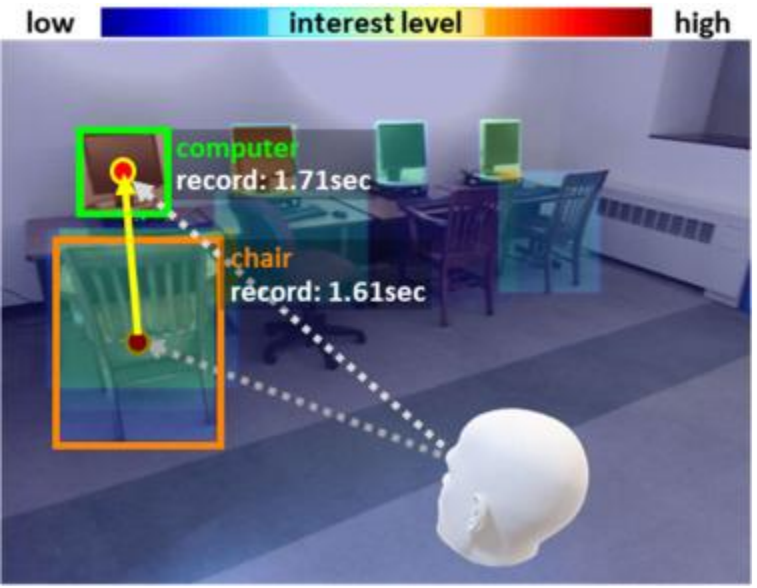
Gaze – Estimation - Egocentric

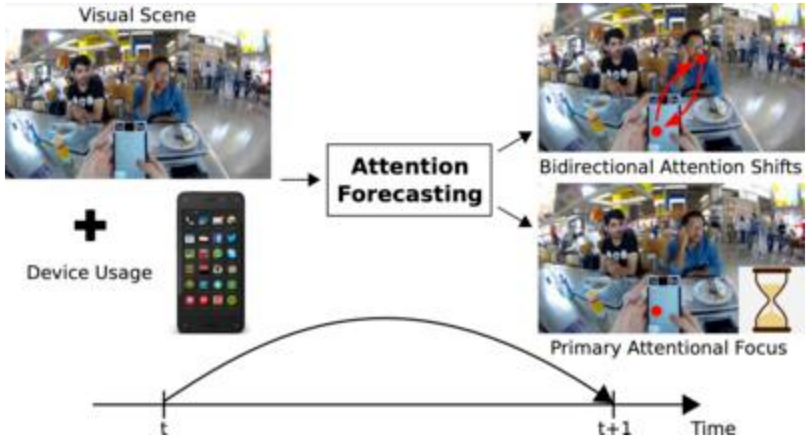


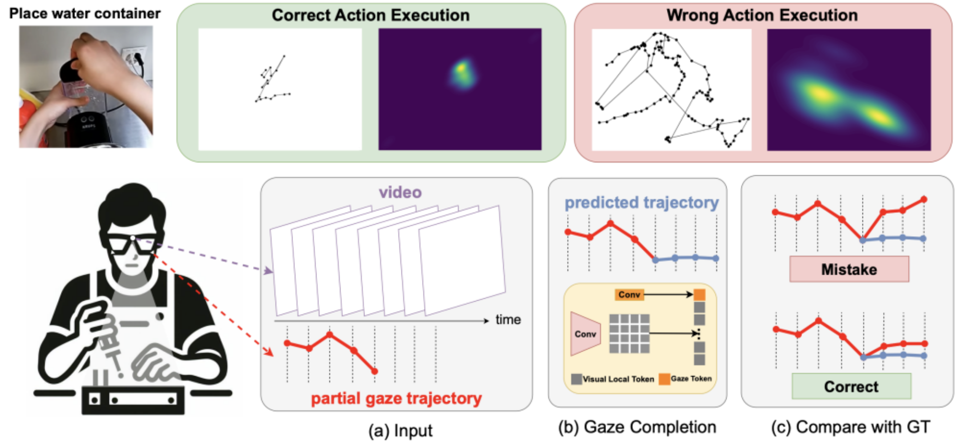
In the Eye of Transformer: Global-Local Correlation for Egocentric Gaze Estimation. Bolin Lai, Miao Liu, Fiona Ryan, James M. Rehg. BMVC, 2022 (Spotlight, Best Student Paper)

From Estimation to Application

Once gaze can be reliably estimated, it is no longer just a prediction target.
Gaze becomes a **powerful signal** of attention, intention, and interaction.
This enables a wide range of **downstream tasks** in egocentric vision and beyond.







One-Stage Object Referring with Gaze Estimation

Object Referring (ObjRef): A multi-modal task to localise objects in an image based on natural language descriptions
The Challenge: Real-world language descriptions can be ambiguous or incomplete
Proposed Solution: A novel gaze-assisted one-stage object referring framework

Key Advantages of One-Stage Gaze-Assisted Approach:

- Simplifies state-of-the-art systems by requiring fewer input signals.
- Improves inference efficiency by implicitly considering all object candidates.
- Resolves the "candidate proposal dilemma" of two-stage solutions, avoiding high computational costs or missing referred objects

J. Chen et al., "One-Stage Object Referring with Gaze Estimation," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 2022, pp. 5017-5026, doi: 10.1109/CVPRW56347.2022.00550.

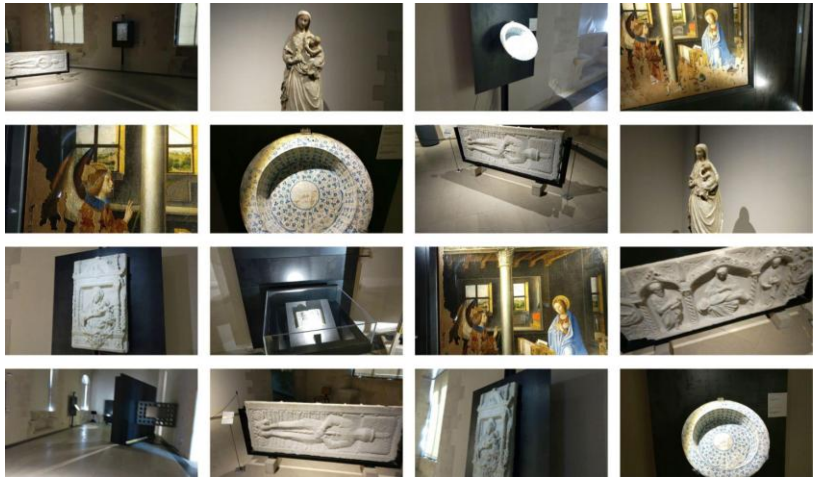
The Attended object detection task

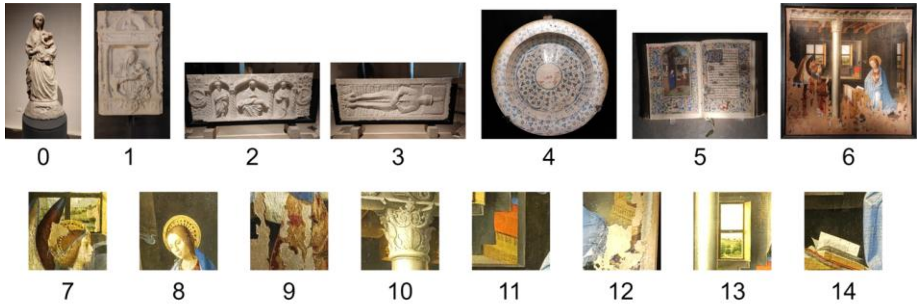
M. Mazzamuto, F. Ragusa, A. Furnari, G.M. Farinella: "Weakly Supervised Attended Object Detection Using Gaze Data as Annotations". In: International Conference on Image Analysis and Processing (ICIAP), 2022, pp.263–274;

The EGO-CH-GAZE Dataset

Details related to the dataset:

- 7 subjects (aged between 24 and 40)
- Video Acquisition: 2272×1278 pixels at 30 fps
- 11 training videos and 3 validation/test videos
- 178977 frames with object of interest annotated with bounding boxes
- 15 objects of interest (8 of the considered objects of interest represent details of the artwork “Annunziazione”).

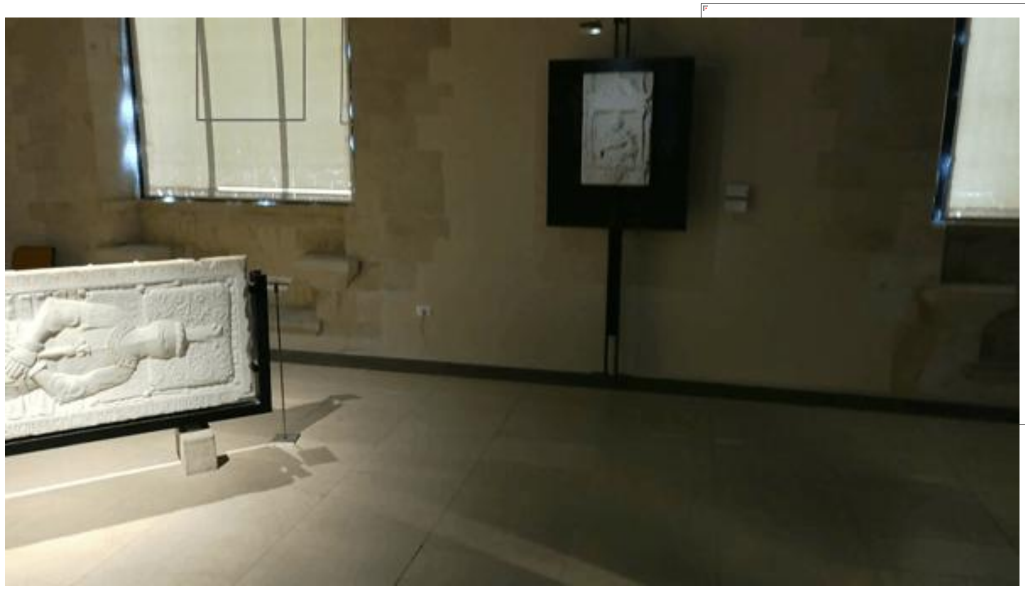


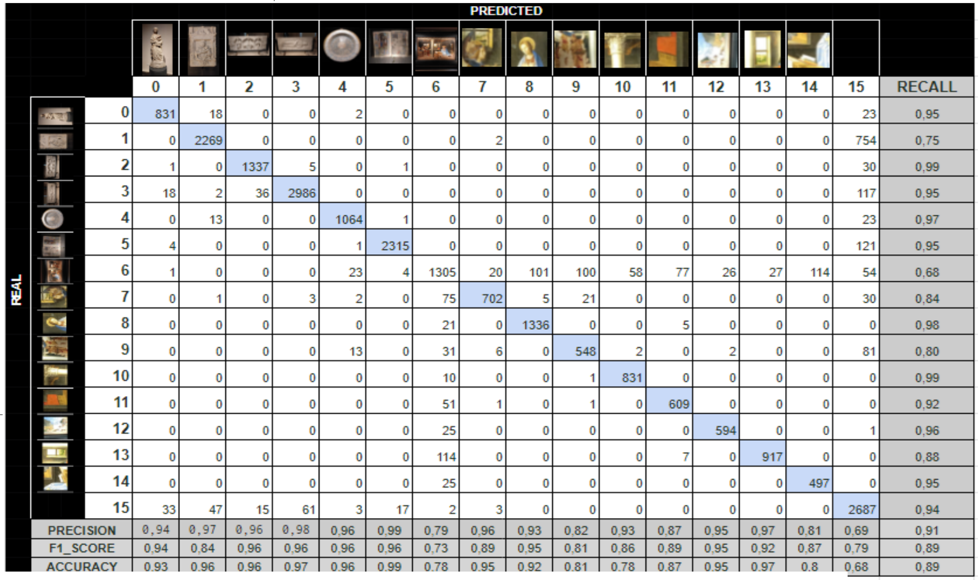


M. Mazzamuto, F. Ragusa, A. Furnari, G.M. Farinella: “Weakly Supervised Attended Object Detection Using Gaze Data as Annotations”. In: International Conference on Image Analysis and Processing (ICIAP), 2022, pp.263–274;

Sliding Window approach

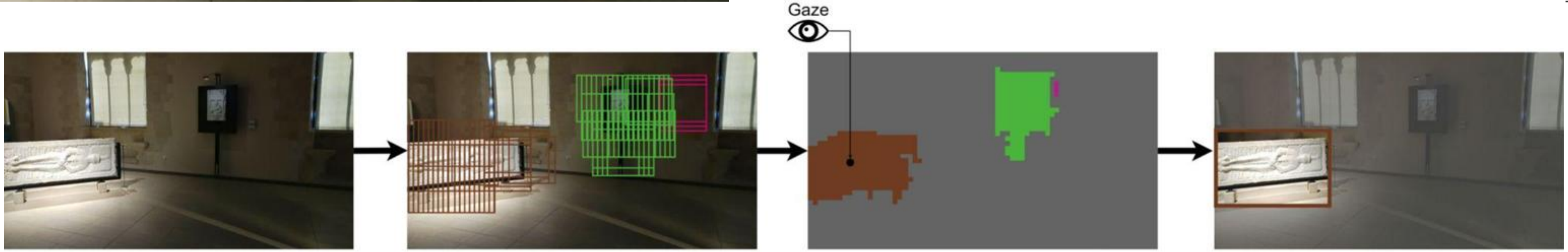
We train a CNN to classify image patches around gaze points using frame-level annotations. This classifier is then used for semantic segmentation by applying a sliding window at test time, producing a segmentation mask. Finally, the gaze is used to extract the attended object's connected component.





PREDICTED

REAL	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	RECALL
0	831	18	0	0	2	0	0	0	0	0	0	0	0	0	0	23	0,95
1	0	2269	0	0	0	0	0	2	0	0	0	0	0	0	0	754	0,75
2	1	0	1337	5	0	1	0	0	0	0	0	0	0	0	0	30	0,99
3	18	2	36	2986	0	0	0	0	0	0	0	0	0	0	0	117	0,95
4	0	13	0	0	1064	1	0	0	0	0	0	0	0	0	0	23	0,97
5	4	0	0	0	1	2315	0	0	0	0	0	0	0	0	0	121	0,95
6	1	0	0	0	23	4	1305	20	101	100	58	77	26	27	114	54	0,68
7	0	1	0	3	2	0	75	702	5	21	0	0	0	0	0	30	0,84
8	0	0	0	0	0	0	21	0	1336	0	0	5	0	0	0	0	0,98
9	0	0	0	0	13	0	31	6	0	548	2	0	2	0	0	81	0,80
10	0	0	0	0	0	0	10	0	0	1	831	0	0	0	0	0	0,99
11	0	0	0	0	0	0	51	1	0	1	0	609	0	0	0	0	0,92
12	0	0	0	0	0	0	25	0	0	0	0	0	594	0	0	1	0,96
13	0	0	0	0	0	0	114	0	0	0	0	7	0	917	0	0	0,88
14	0	0	0	0	0	0	25	0	0	0	0	0	0	0	497	0	0,95
15	33	47	15	61	3	17	2	3	0	0	0	0	0	0	0	2687	0,94
PRECISION	0,94	0,97	0,96	0,98	0,96	0,99	0,79	0,96	0,93	0,82	0,93	0,87	0,95	0,97	0,81	0,69	0,91
F1_SCORE	0,94	0,84	0,96	0,96	0,96	0,96	0,73	0,89	0,95	0,81	0,86	0,89	0,95	0,92	0,87	0,79	0,89
ACCURACY	0.93	0.96	0.96	0.97	0.96	0.99	0.78	0.95	0.92	0.81	0.78	0.87	0.95	0.97	0.8	0,68	0,89



(a) (b) (c) (d)

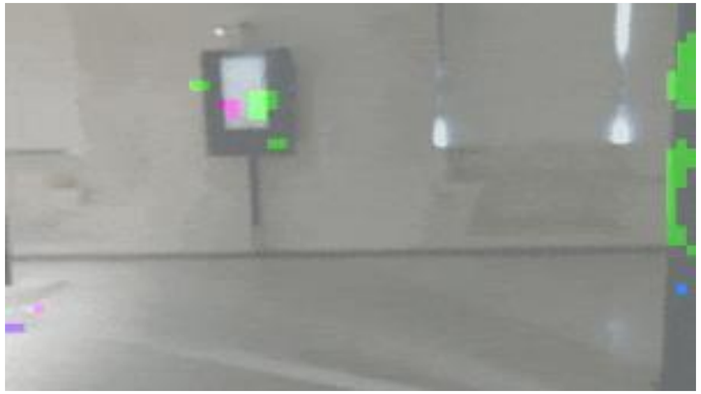
M. Mazzamuto, F. Ragusa, A. Furnari, G.M. Farinella: "Weakly Supervised Attended Object Detection Using Gaze Data as Annotations". In: International Conference on Image Analysis and Processing (ICIAP), 2022, pp.263–274;

Fully Convolutional attended object detection

Sliding Window approach has the main drawback of being slow. (Processing an image at full resolution takes up to 168 seconds on a Tesla-K80 GPU. To speed up the approach, we modify the trained ResNet by removing the Global Average Pooling operation and replacing it with a fully connected classifier with a 1×1 convolutional layer.

This allows the network to predict a semantic segmentation mask of the whole image in a single step.

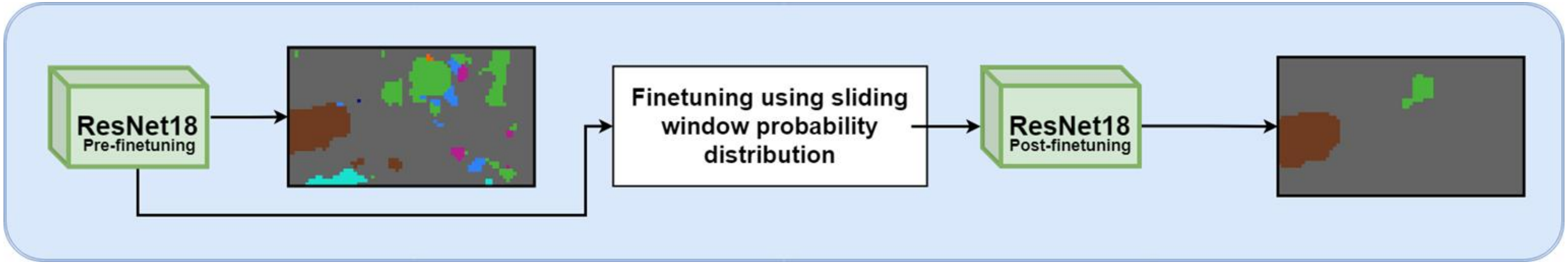
Given an input frame, the model outputs the class probability distributions for each pixel.

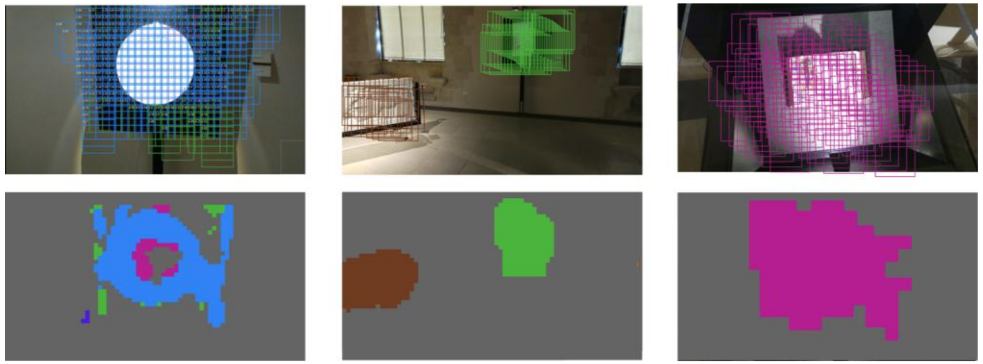


M. Mazzamuto, F. Ragusa, A. Furnari, G.M. Farinella: "Weakly Supervised Attended Object Detection Using Gaze Data as Annotations". In: International Conference on Image Analysis and Processing (ICIAP), 2022, pp.263–274;

Finetuning

The fully convolutional approach is faster but less accurate than the sliding window method. To improve its performance, we fine-tune it using coarse segmentation masks from the sliding window approach. We apply the Kullback-Leibler Divergence loss to align the pixel-wise probability distributions of the two models.





M. Mazzamuto, F. Ragusa, A. Furnari, G.M. Farinella: "Weakly Supervised Attended Object Detection Using Gaze Data as Annotations". In: International Conference on Image Analysis and Processing (ICIAP), 2022, pp.263–274;

Results comparison

Fine-tuning the fully convolutional model with the proposed optimization procedure allows to achieve a performance similar to one obtained with the sliding window approach, with an mAP of 0.19 and an mAP50 of 0.41, while retaining the reduced inference time of 0.31 seconds per image.

Model	Supervision	Inference time (seconds)	mAP	mAP 50
Sliding window	class	168	0.19	0.43
Fully convolutional	class	0.31	0.18	0.34
Fully convolutional + fine-tuning	class	0.31	0.19	0.41
Faster-RCNN (baseline)	bbox	0.80	0.42	0.60

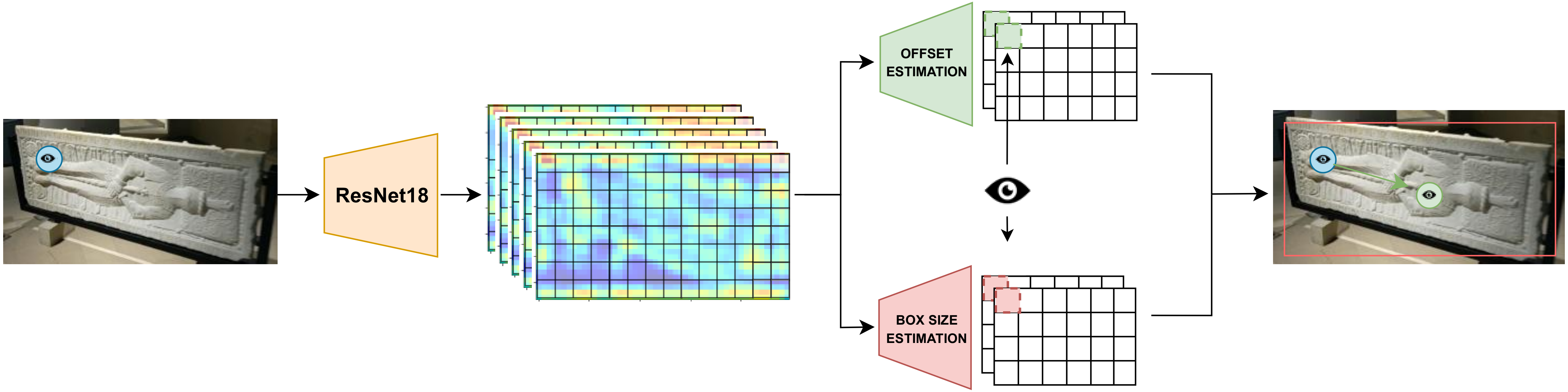
M. Mazzamuto, F. Ragusa, A. Furnari, G.M. Farinella: "Weakly Supervised Attended Object Detection Using Gaze Data as Annotations". In: International Conference on Image Analysis and Processing (ICIAP), 2022, pp.263–274;

Box coordinates regressor

We extended our conference paper to a journal by introducing a method that regresses the attended object's bounding box from gaze coordinates.

The proposed approach extracts spatial features from the image and processes them through two convolutional modules. These estimate the bounding box center and dimensions for each grid point, generating multiple predictions.

By explicitly leveraging gaze location, the method focuses on detecting a single object "prompted" by gaze.



M. Mazzamuto, F. Ragusa, A. Furnari, G.M. Farinella: "Learning to Detect Attended Objects in Cultural Sites with Gaze Signals and Weak Object Supervision." In: ACM Journal on Computing and Cultural Heritage (JOCCH), 2024, Volume 17, Issue 3

Results comparison

Results obtained by the compared approaches, when all classes are considered, for each level of supervision (from the highest to the lowest). In bold the best results by supervision group.

METHOD	GAZE	CLASS	ATTENDED BOX	ALL BOXES	PRE-TRAIN	mAP50	mAP	Inference (ms)
1) Faster-RCNN	✓	✓	✓	✓	COCO	0,60	**0,42**	67,8
2) RetinaNet	✓	✓	✓	✓	COCO	**0,63**	0,41	**48,6**
3) Faster-RCNN	✓	✓	✓	✗	COCO	0,53	0,36	67,8
4) RetinaNet	✓	✓	✓	✗	COCO	**0,57**	**0,39**	48,6
5) Our (gaze conditioned box regressor)	✓	✓	✓	✗	ImageNet	0,54	0,35	**21**
6) Our (gaze conditioned box regressor)	✓	✓	✓	✗	COCO	**0,57**	0,37	**21**
7) Sliding window	✓	✓	✗	✗	COCO	**0,43**	**0,19**	9000
8) Our (FC)	✓	✓	✗	✗	COCO	0,34	0,18	**26**
9) Our (FC+Finetuning)	✓	✓	✗	✗	COCO	0,41	**0,19**	**26**
10) InSPyReNet	✓	✗	✗	✗	DUTS-TR	**0,1**	**0,06**	370
11) U^2-Net	✓	✗	✗	✗	DUTS-TR	0,09	**0,06**	370
12) Faster-RCNN	✓	✗	✗	✗	COCO	0,02	0,005	67,8
13) RetinaNet	✓	✗	✗	✗	COCO	0,024	0,007	48,6
14) Our (gaze conditioned box regressor)	✓	✗	✗	✗	COCO	**0,1**	0,008	**21**

M. Mazzamuto, F. Ragusa, A. Furnari, G.M. Farinella: "Learning to Detect Attended Objects in Cultural Sites with Gaze Signals and Weak Object Supervision." In: ACM Journal on Computing and Cultural Heritage (JOCCH), 2024, Volume 17, Issue 3

Results comparison

The fully supervised approaches, Faster-RCNN and RetinaNet, show strong performance in object detection when provided with gaze, class, attended box, and all box information as supervision.

METHOD	GAZE	CLASS	ATTENDED BOX	ALL BOXES	PRE-TRAIN	mAP50	mAP	Inference (ms)
1) Faster-RCNN	✓	✓	✓	✓	COCO	0,60	**0,42**	67,8
2) RetinaNet	✓	✓	✓	✓	COCO	**0,63**	0,41	**48,6**
3) Faster-RCNN	✓	✓	✓	✗	COCO	0,53	0,36	67,8
4) RetinaNet	✓	✓	✓	✗	COCO	**0,57**	**0,39**	48,6
5) Our (gaze conditioned box regressor)	✓	✓	✓	✗	ImageNet	0,54	0,35	**21**
6) Our (gaze conditioned box regressor)	✓	✓	✓	✗	COCO	**0,57**	0,37	**21**
7) Sliding window	✓	✓	✗	✗	COCO	**0,43**	**0,19**	9000
8) Our (FC)	✓	✓	✗	✗	COCO	0,34	0,18	**26**
9) Our (FC+Finetuning)	✓	✓	✗	✗	COCO	0,41	**0,19**	**26**
10) InSPyReNet	✓	✗	✗	✗	DUTS-TR	**0,1**	**0,06**	370
11) U^2-Net	✓	✗	✗	✗	DUTS-TR	0,09	**0,06**	370
12) Faster-RCNN	✓	✗	✗	✗	COCO	0,02	0,005	67,8
13) RetinaNet	✓	✗	✗	✗	COCO	0,024	0,007	48,6
14) Our (gaze conditioned box regressor)	✓	✗	✗	✗	COCO	**0,1**	0,008	**21**

M. Mazzamuto, F. Ragusa, A. Furnari, G.M. Farinella: "Learning to Detect Attended Objects in Cultural Sites with Gaze Signals and Weak Object Supervision." In: ACM Journal on Computing and Cultural Heritage (JOCCH), 2024, Volume 17, Issue 3

Results comparison

Among the weakly supervised approaches, using gaze, class, and the attended object box, RetinaNet (row 4) and our proposed gaze-conditioned box regressor (row 6) achieved the highest results. Our proposed approach performs with a reduced inference time of 21 ms.

METHOD	GAZE	CLASS	ATTENDED BOX	ALL BOXES	PRE-TRAIN	mAP50	mAP	Inference (ms)
1) Faster-RCNN	✓	✓	✓	✓	COCO	0,60	**0,42**	67,8
2) RetinaNet	✓	✓	✓	✓	COCO	**0,63**	0,41	**48,6**
3) Faster-RCNN	✓	✓	✓	✗	COCO	0,53	0,36	67,8
4) RetinaNet	✓	✓	✓	✗	COCO	**0,57**	**0,39**	48,6
5) Our (gaze conditioned box regressor)	✓	✓	✓	✗	ImageNet	0,54	0,35	**21**
6) Our (gaze conditioned box regressor)	✓	✓	✓	✗	COCO	**0,57**	0,37	**21**
7) Sliding window	✓	✓	✗	✗	COCO	**0,43**	**0,19**	9000
8) Our (FC)	✓	✓	✗	✗	COCO	0,34	0,18	**26**
9) Our (FC+Finetuning)	✓	✓	✗	✗	COCO	0,41	**0,19**	**26**
10) InSPyReNet	✓	✗	✗	✗	DUTS-TR	**0,1**	**0,06**	370
11) U^2-Net	✓	✗	✗	✗	DUTS-TR	0,09	**0,06**	370
12) Faster-RCNN	✓	✗	✗	✗	COCO	0,02	0,005	67,8
13) RetinaNet	✓	✗	✗	✗	COCO	0,024	0,007	48,6
14) Our (gaze conditioned box regressor)	✓	✗	✗	✗	COCO	**0,1**	0,008	**21**

M. Mazzamuto, F. Ragusa, A. Furnari, G.M. Farinella: "Learning to Detect Attended Objects in Cultural Sites with Gaze Signals and Weak Object Supervision." In: ACM Journal on Computing and Cultural Heritage (JOCCH), 2024, Volume 17, Issue 3

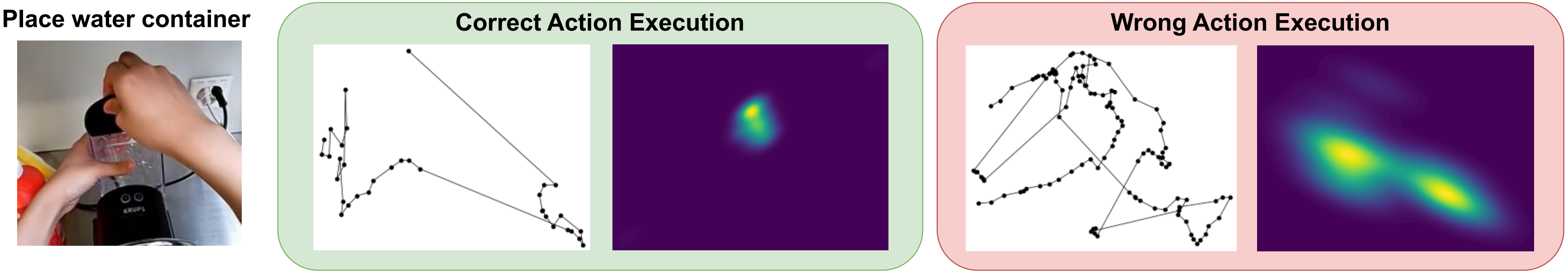
Foveation resolution representation

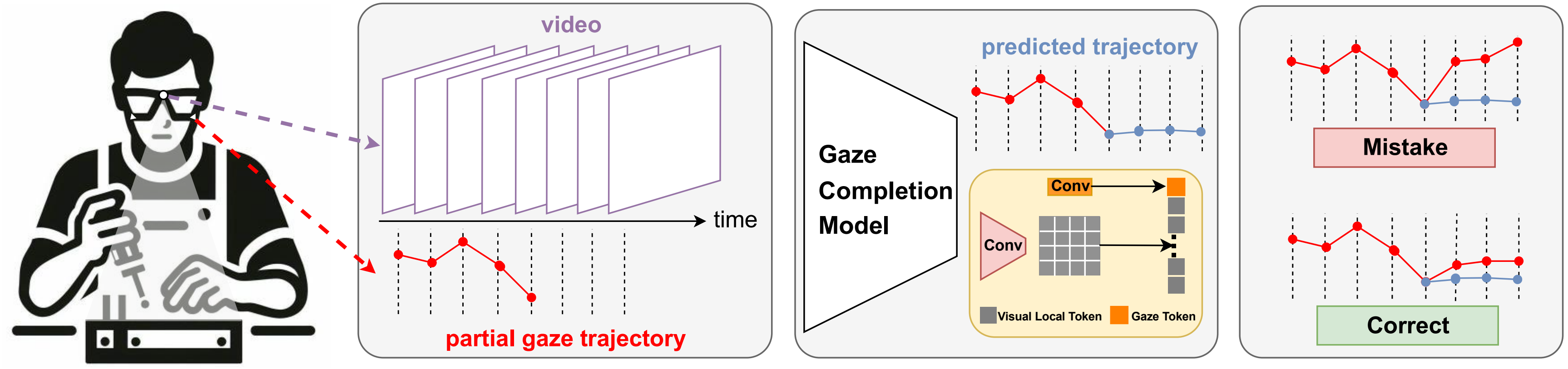
DeepFovea Overview

- DeepFovea: Neural reconstruction for foveated rendering & video compression
- Uses GANs to reconstruct peripheral vision from a small fraction of pixels
- Improves computational efficiency without noticeable quality loss
- Supports real-time gaze-contingent AR/VR displays

A.S. Kaplanyan, A. Sochenov, T. Leimkühler, M. Okunev, T. Goodall, G. Rufo: "DeepFovea: Neural Reconstruction for Foveated Rendering and Video Compression Using Learned Statistics of Natural Videos." ACM Transactions on Graphics (TOG), 2019.

Gazing Into Missteps





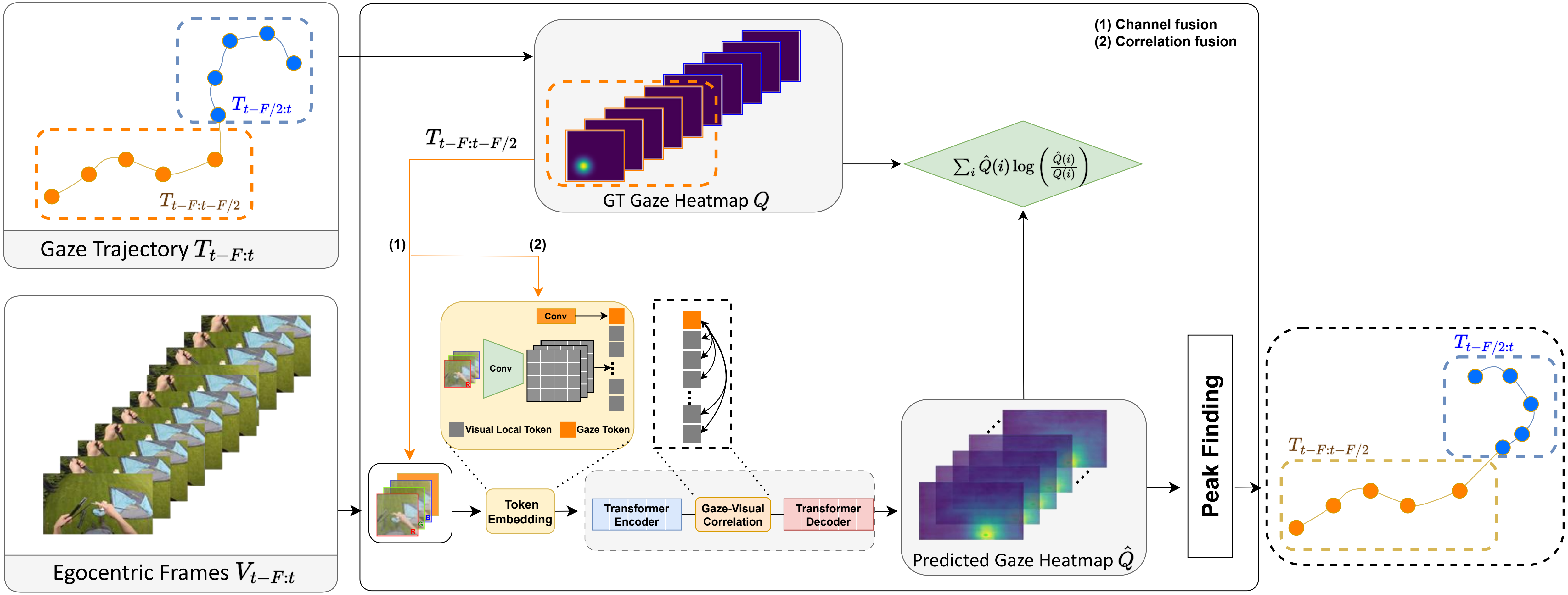
(a) Input (b) Gaze Completion (c) Compare with GT

M. Mazzamuto, A. Furnari, Y. Sato, G.M. Farinella: "Gazing Into Missteps: Leveraging Eye-Gaze for Unsupervised Mistake Detection in Egocentric Videos of Skilled Human Activities by Detecting Unpredictable Gaze." CVPR, 2025.

Gaze behaeviour during a mistake

M. Mazzamuto, A. Furnari, Y. Sato, G.M. Farinella: "Gazing Into Missteps: Leveraging Eye-Gaze for Unsupervised Mistake Detection in Egocentric Videos of Skilled Human Activities by Detecting Unpredictable Gaze." CVPR, 2025.

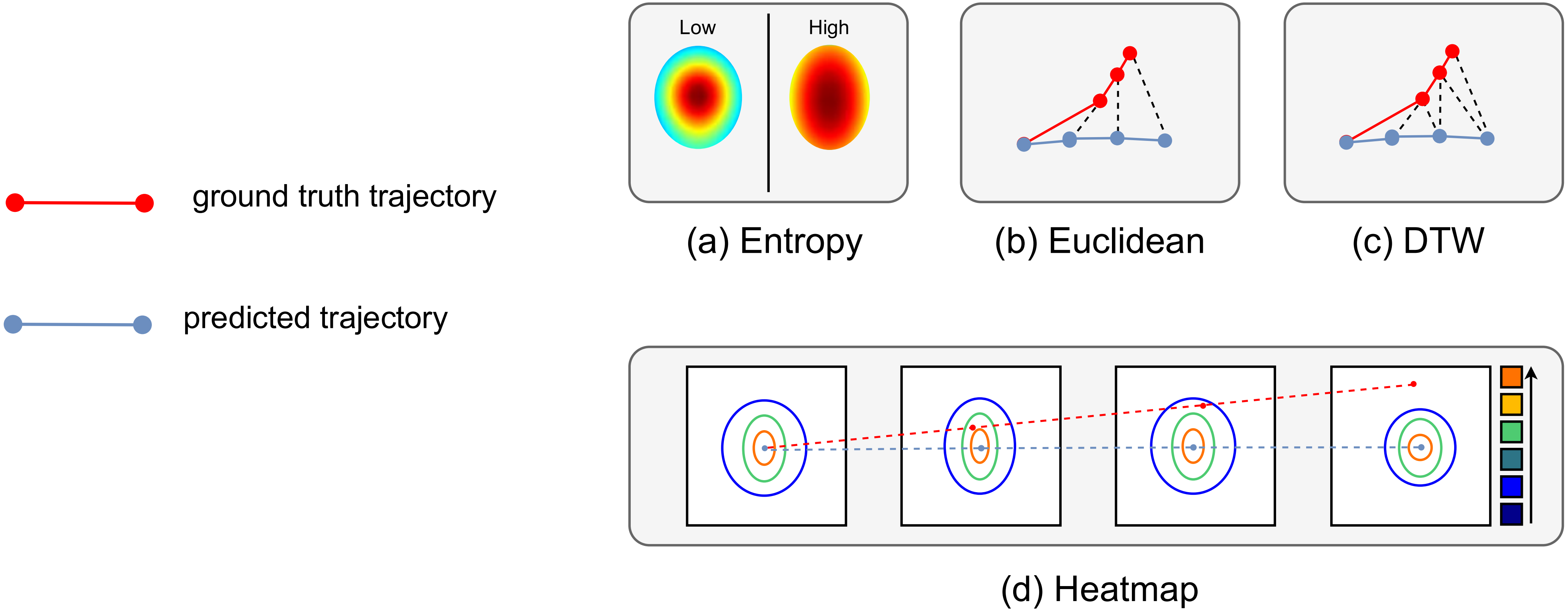
Gaze completion Module



M. Mazzamuto, A. Furnari, Y. Sato, G.M. Farinella: "Gazing Into Missteps: Leveraging Eye-Gaze for Unsupervised Mistake Detection in Egocentric Videos of Skilled Human Activities by Detecting Unpredictable Gaze." CVPR, 2025.

Scoring

To compare the predicted trajectory with the ground truth (GT) and check if a mistake is occurring or not, we introduce different scoring functions.



M. Mazzamuto, A. Furnari, Y. Sato, G.M. Farinella: "Gazing Into Missteps: Leveraging Eye-Gaze for Unsupervised Mistake Detection in Egocentric Videos of Skilled Human Activities by Detecting Unpredictable Gaze." CVPR, 2025.

Results

Ablation Study

	Scoring	Fusion	F1	Precision	Recall	AUC
1	Random	//	0.36	0.29	0.42	0.51
2	Entropy	//	0.41	0.27	0.62	0.51
3	Euclidean	//	0.42	0.29	0.60	0.55
4	DTW	//	0.44	0.31	0.68	0.56
5	Heatmap	//	0.45	0.32	0.70	0.57
6	Heatmap	CH	0.45	0.32	0.74	0.63
7	Heatmap	CORR	0.50	0.36	0.82	0.65
8	Heatmap	CH + CORR	**0.51**	**0.36**	**0.85**	**0.69**

Mistake detection result on EPIC-Tent.

Method	Sup. Level	F1	Precision	Recall	AUC
Random	//	0.36	0.29	0.42	0.51
TimeSformer [2]	Fully Supervised	0.49	0.35	0.80	0.67
C2F [35]	Fully Supervised	**0.58**	**0.44**	**0.85**	**0.72**
TrajREC (G) [39]	One-Class	0.40	0.26	0.88	0.51
MoCoDAD (G) [8]	One-Class	0.43	0.27	**0.91**	0.50
TrajREC (H) [39]	One-Class	0.44	0.31	0.76	0.55
MoCoDAD (H) [8]	One-Class	0.46	0.33	0.79	0.60
TrajREC (H+G) [39]	One-Class	0.42	0.29	0.75	0.53
MoCoDAD (H+G) [8]	One-Class	0.43	0.30	0.77	0.56
TrajREC (H+G)* [39]	One-Class	0.47	0.34	0.77	0.63
MoCoDAD (H+G)* [8]	One-Class	0.49	0.35	0.81	0.65
GLC [19]	One-Class	0.46	0.37	0.62	0.66
Ours	One-Class	0.52	0.37	0.85	0.69
Ours + MoCoDAD (H)*	One-Class	**0.54**	**0.41**	0.86	**0.72**
TrajREC (G) [39]	Unsupervised	0.27	0.16	**0.94**	0.50
MoCoDAD (G) [8]	Unsupervised	0.33	0.21	0.88	0.51
TrajREC (H) [39]	Unsupervised	0.40	0.27	0.79	0.58
MoCoDAD (H) [8]	Unsupervised	0.41	0.27	0.86	0.60
MoCoDAD (H+G)* [8]	Unsupervised	0.41	0.27	0.88	0.60
GLC [19]	Unsupervised	0.44	0.33	0.70	0.61
Ours	Unsupervised	0.51	0.36	0.85	0.69
Ours + MoCoDAD (H)*	Unsupervised	**0.52**	**0.37**	0.88	**0.70**

* Late fusion

M. Mazzamuto, A. Furnari, Y. Sato, G.M. Farinella: "Gazing Into Missteps: Leveraging Eye-Gaze for Unsupervised Mistake Detection in Egocentric Videos of Skilled Human Activities by Detecting Unpredictable Gaze." CVPR, 2025.

Results

Without fusion or trajectory conditioning, entropy-based scoring performed only slightly better than random

Ablation Study

	Scoring	Fusion	F1	Precision	Recall	AUC
1	Random	//	0.36	0.29	0.42	0.51
2	Entropy	//	0.41	0.27	0.62	0.51
3	Euclidean	//	0.42	0.29	0.60	0.55
4	DTW	//	0.44	0.31	0.68	0.56
5	Heatmap	//	0.45	0.32	0.70	0.57
6	Heatmap	CH	0.45	0.32	0.74	0.63
7	Heatmap	CORR	0.50	0.36	0.82	0.65
8	Heatmap	CH + CORR	**0.51**	**0.36**	**0.85**	**0.69**

Mistake detection result on EPIC-Tent.

Method	Sup. Level	F1	Precision	Recall	AUC
Random	//	0.36	0.29	0.42	0.51
TimeSformer [2]	Fully Supervised	0.49	0.35	0.80	0.67
C2F [35]	Fully Supervised	**0.58**	**0.44**	**0.85**	**0.72**
TrajREC (G) [39]	One-Class	0.40	0.26	0.88	0.51
MoCoDAD (G) [8]	One-Class	0.43	0.27	**0.91**	0.50
TrajREC (H) [39]	One-Class	0.44	0.31	0.76	0.55
MoCoDAD (H) [8]	One-Class	0.46	0.33	0.79	0.60
TrajREC (H+G) [39]	One-Class	0.42	0.29	0.75	0.53
MoCoDAD (H+G) [8]	One-Class	0.43	0.30	0.77	0.56
TrajREC (H+G)* [39]	One-Class	0.47	0.34	0.77	0.63
MoCoDAD (H+G)* [8]	One-Class	0.49	0.35	0.81	0.65
GLC [19]	One-Class	0.46	0.37	0.62	0.66
Ours	One-Class	0.52	0.37	0.85	0.69
Ours + MoCoDAD (H)*	One-Class	**0.54**	**0.41**	0.86	**0.72**
TrajREC (G) [39]	Unsupervised	0.27	0.16	**0.94**	0.50
MoCoDAD (G) [8]	Unsupervised	0.33	0.21	0.88	0.51
TrajREC (H) [39]	Unsupervised	0.40	0.27	0.79	0.58
MoCoDAD (H) [8]	Unsupervised	0.41	0.27	0.86	0.60
MoCoDAD (H+G)* [8]	Unsupervised	0.41	0.27	0.88	0.60
GLC [19]	Unsupervised	0.44	0.33	0.70	0.61
Ours	Unsupervised	0.51	0.36	0.85	0.69
Ours + MoCoDAD (H)*	Unsupervised	**0.52**	**0.37**	0.88	**0.70**

* Late fusion

M. Mazzamuto, A. Furnari, Y. Sato, G.M. Farinella: "Gazing Into Missteps: Leveraging Eye-Gaze for Unsupervised Mistake Detection in Egocentric Videos of Skilled Human Activities by Detecting Unpredictable Gaze." CVPR, 2025.

Results

Euclidean and DTW scoring functions improved the results.

Ablation Study

	Scoring	Fusion	F1	Precision	Recall	AUC
1	Random	//	0.36	0.29	0.42	0.51
2	Entropy	//	0.41	0.27	0.62	0.51
3	Euclidean	//	0.42	0.29	0.60	0.55
4	DTW	//	0.44	0.31	0.68	0.56
5	Heatmap	//	0.45	0.32	0.70	0.57
6	Heatmap	CH	0.45	0.32	0.74	0.63
7	Heatmap	CORR	0.50	0.36	0.82	0.65
8	Heatmap	CH + CORR	**0.51**	**0.36**	**0.85**	**0.69**

Mistake detection result on EPIC-Tent.

Method	Sup. Level	F1	Precision	Recall	AUC
Random	//	0.36	0.29	0.42	0.51
TimeSformer [2]	Fully Supervised	0.49	0.35	0.80	0.67
C2F [35]	Fully Supervised	**0.58**	**0.44**	**0.85**	**0.72**
TrajREC (G) [39]	One-Class	0.40	0.26	0.88	0.51
MoCoDAD (G) [8]	One-Class	0.43	0.27	**0.91**	0.50
TrajREC (H) [39]	One-Class	0.44	0.31	0.76	0.55
MoCoDAD (H) [8]	One-Class	0.46	0.33	0.79	0.60
TrajREC (H+G) [39]	One-Class	0.42	0.29	0.75	0.53
MoCoDAD (H+G) [8]	One-Class	0.43	0.30	0.77	0.56
TrajREC (H+G)* [39]	One-Class	0.47	0.34	0.77	0.63
MoCoDAD (H+G)* [8]	One-Class	0.49	0.35	0.81	0.65
GLC [19]	One-Class	0.46	0.37	0.62	0.66
Ours	One-Class	0.52	0.37	0.85	0.69
Ours + MoCoDAD (H)*	One-Class	**0.54**	**0.41**	0.86	**0.72**
TrajREC (G) [39]	Unsupervised	0.27	0.16	**0.94**	0.50
MoCoDAD (G) [8]	Unsupervised	0.33	0.21	0.88	0.51
TrajREC (H) [39]	Unsupervised	0.40	0.27	0.79	0.58
MoCoDAD (H) [8]	Unsupervised	0.41	0.27	0.86	0.60
MoCoDAD (H+G)* [8]	Unsupervised	0.41	0.27	0.88	0.60
GLC [19]	Unsupervised	0.44	0.33	0.70	0.61
Ours	Unsupervised	0.51	0.36	0.85	0.69
Ours + MoCoDAD (H)*	Unsupervised	**0.52**	**0.37**	0.88	**0.70**

* Late fusion

M. Mazzamuto, A. Furnari, Y. Sato, G.M. Farinella: "Gazing Into Missteps: Leveraging Eye-Gaze for Unsupervised Mistake Detection in Egocentric Videos of Skilled Human Activities by Detecting Unpredictable Gaze." CVPR, 2025.

Results

While the **heatmap-based** scoring function achieved the best performance, making it the chosen method for the rest of the experiments.

Ablation Study

	Scoring	Fusion	F1	Precision	Recall	AUC
1	Random	//	0.36	0.29	0.42	0.51
2	Entropy	//	0.41	0.27	0.62	0.51
3	Euclidean	//	0.42	0.29	0.60	0.55
4	DTW	//	0.44	0.31	0.68	0.56
5	Heatmap	//	0.45	0.32	0.70	0.57
6	Heatmap	CH	0.45	0.32	0.74	0.63
7	Heatmap	CORR	0.50	0.36	0.82	0.65
8	Heatmap	CH + CORR	**0.51**	**0.36**	**0.85**	**0.69**

Mistake detection result on EPIC-Tent.

Method	Sup. Level	F1	Precision	Recall	AUC
Random	//	0.36	0.29	0.42	0.51
TimeSformer [2]	Fully Supervised	0.49	0.35	0.80	0.67
C2F [35]	Fully Supervised	**0.58**	**0.44**	**0.85**	**0.72**
TrajREC (G) [39]	One-Class	0.40	0.26	0.88	0.51
MoCoDAD (G) [8]	One-Class	0.43	0.27	**0.91**	0.50
TrajREC (H) [39]	One-Class	0.44	0.31	0.76	0.55
MoCoDAD (H) [8]	One-Class	0.46	0.33	0.79	0.60
TrajREC (H+G) [39]	One-Class	0.42	0.29	0.75	0.53
MoCoDAD (H+G) [8]	One-Class	0.43	0.30	0.77	0.56
TrajREC (H+G)* [39]	One-Class	0.47	0.34	0.77	0.63
MoCoDAD (H+G)* [8]	One-Class	0.49	0.35	0.81	0.65
GLC [19]	One-Class	0.46	0.37	0.62	0.66
Ours	One-Class	0.52	0.37	0.85	0.69
Ours + MoCoDAD (H)*	One-Class	**0.54**	**0.41**	0.86	**0.72**
TrajREC (G) [39]	Unsupervised	0.27	0.16	**0.94**	0.50
MoCoDAD (G) [8]	Unsupervised	0.33	0.21	0.88	0.51
TrajREC (H) [39]	Unsupervised	0.40	0.27	0.79	0.58
MoCoDAD (H) [8]	Unsupervised	0.41	0.27	0.86	0.60
MoCoDAD (H+G)* [8]	Unsupervised	0.41	0.27	0.88	0.60
GLC [19]	Unsupervised	0.44	0.33	0.70	0.61
Ours	Unsupervised	0.51	0.36	0.85	0.69
Ours + MoCoDAD (H)*	Unsupervised	**0.52**	**0.37**	0.88	**0.70**

* Late fusion

M. Mazzamuto, A. Furnari, Y. Sato, G.M. Farinella: "Gazing Into Missteps: Leveraging Eye-Gaze for Unsupervised Mistake Detection in Egocentric Videos of Skilled Human Activities by Detecting Unpredictable Gaze." CVPR, 2025.

Results

Both fusion strategies enhanced performance, with correlation-based fusion outperforming channel fusion.

Ablation Study

	Scoring	Fusion	F1	Precision	Recall	AUC
1	Random	//	0.36	0.29	0.42	0.51
2	Entropy	//	0.41	0.27	0.62	0.51
3	Euclidean	//	0.42	0.29	0.60	0.55
4	DTW	//	0.44	0.31	0.68	0.56
5	Heatmap	//	0.45	0.32	0.70	0.57
6	Heatmap	CH	0.45	0.32	0.74	0.63
7	Heatmap	CORR	0.50	0.36	0.82	0.65
8	Heatmap	CH + CORR	**0.51**	**0.36**	**0.85**	**0.69**

Mistake detection result on EPIC-Tent.

Method	Sup. Level	F1	Precision	Recall	AUC
Random	//	0.36	0.29	0.42	0.51
TimeSformer [2]	Fully Supervised	0.49	0.35	0.80	0.67
C2F [35]	Fully Supervised	**0.58**	**0.44**	**0.85**	**0.72**
TrajREC (G) [39]	One-Class	0.40	0.26	0.88	0.51
MoCoDAD (G) [8]	One-Class	0.43	0.27	**0.91**	0.50
TrajREC (H) [39]	One-Class	0.44	0.31	0.76	0.55
MoCoDAD (H) [8]	One-Class	0.46	0.33	0.79	0.60
TrajREC (H+G) [39]	One-Class	0.42	0.29	0.75	0.53
MoCoDAD (H+G) [8]	One-Class	0.43	0.30	0.77	0.56
TrajREC (H+G)* [39]	One-Class	0.47	0.34	0.77	0.63
MoCoDAD (H+G)* [8]	One-Class	0.49	0.35	0.81	0.65
GLC [19]	One-Class	0.46	0.37	0.62	0.66
Ours	One-Class	0.52	0.37	0.85	0.69
Ours + MoCoDAD (H)*	One-Class	**0.54**	**0.41**	0.86	**0.72**
TrajREC (G) [39]	Unsupervised	0.27	0.16	**0.94**	0.50
MoCoDAD (G) [8]	Unsupervised	0.33	0.21	0.88	0.51
TrajREC (H) [39]	Unsupervised	0.40	0.27	0.79	0.58
MoCoDAD (H) [8]	Unsupervised	0.41	0.27	0.86	0.60
MoCoDAD (H+G)* [8]	Unsupervised	0.41	0.27	0.88	0.60
GLC [19]	Unsupervised	0.44	0.33	0.70	0.61
Ours	Unsupervised	0.51	0.36	0.85	0.69
Ours + MoCoDAD (H)*	Unsupervised	**0.52**	**0.37**	0.88	**0.70**

* Late fusion

M. Mazzamuto, A. Furnari, Y. Sato, G.M. Farinella: "Gazing Into Missteps: Leveraging Eye-Gaze for Unsupervised Mistake Detection in Egocentric Videos of Skilled Human Activities by Detecting Unpredictable Gaze." CVPR, 2025.

Results

Combining both fusion strategies achieves the best results. This is the final setup we will use for all the other experiments.

Ablation Study

	Scoring	Fusion	F1	Precision	Recall	AUC
1	Random	//	0.36	0.29	0.42	0.51
2	Entropy	//	0.41	0.27	0.62	0.51
3	Euclidean	//	0.42	0.29	0.60	0.55
4	DTW	//	0.44	0.31	0.68	0.56
5	Heatmap	//	<u>0.45</u>	<u>0.32</u>	<u>0.70</u>	<u>0.57</u>
6	Heatmap	CH	0.45	0.32	0.74	0.63
7	Heatmap	CORR	0.50	0.36	0.82	0.65
8	Heatmap	CH + CORR	**<u>0.51</u>**	**<u>0.36</u>**	**<u>0.85</u>**	**<u>0.69</u>**

Mistake detection result on EPIC-Tent.

Method	Sup. Level	F1	Precision	Recall	AUC
Random	//	0.36	0.29	0.42	0.51
TimeSformer [2]	Fully Supervised	<u>0.49</u>	<u>0.35</u>	<u>0.80</u>	<u>0.67</u>
C2F [35]	Fully Supervised	**0.58**	**0.44**	**0.85**	**0.72**
TrajREC (G) [39]	One-Class	0.40	0.26	<u>0.88</u>	0.51
MoCoDAD (G) [8]	One-Class	0.43	0.27	**0.91**	0.50
TrajREC (H) [39]	One-Class	0.44	0.31	0.76	0.55
MoCoDAD (H) [8]	One-Class	0.46	0.33	0.79	0.60
TrajREC (H+G) [39]	One-Class	0.42	0.29	0.75	0.53
MoCoDAD (H+G) [8]	One-Class	0.43	0.30	0.77	0.56
TrajREC (H+G)* [39]	One-Class	0.47	0.34	0.77	0.63
MoCoDAD (H+G)* [8]	One-Class	0.49	0.35	0.81	0.65
GLC [19]	One-Class	0.46	<u>0.37</u>	0.62	0.66
Ours	One-Class	<u>0.52</u>	<u>0.37</u>	0.85	<u>0.69</u>
Ours + MoCoDAD (H)*	One-Class	**0.54**	**0.41**	0.86	**0.72**
TrajREC (G) [39]	Unsupervised	0.27	0.16	**0.94**	0.50
MoCoDAD (G) [8]	Unsupervised	0.33	0.21	<u>0.88</u>	0.51
TrajREC (H) [39]	Unsupervised	0.40	0.27	0.79	0.58
MoCoDAD (H) [8]	Unsupervised	0.41	0.27	0.86	0.60
MoCoDAD (H+G)* [8]	Unsupervised	0.41	0.27	<u>0.88</u>	0.60
GLC [19]	Unsupervised	0.44	0.33	0.70	0.61
Ours	Unsupervised	<u>0.51</u>	<u>0.36</u>	0.85	<u>0.69</u>
Ours + MoCoDAD (H)*	Unsupervised	**0.52**	**0.37**	<u>0.88</u>	**0.70**

* Late fusion

M. Mazzamuto, A. Furnari, Y. Sato, G.M. Farinella: "Gazing Into Missteps: Leveraging Eye-Gaze for Unsupervised Mistake Detection in Egocentric Videos of Skilled Human Activities by Detecting Unpredictable Gaze." CVPR, 2025.

Results

C2F outperforms TimeSformer across all metrics, especially in F1 score, highlighting its superior temporal reasoning for dynamic activity mistake detection.

Ablation Study

	Scoring	Fusion	F1	Precision	Recall	AUC
1	Random	//	0.36	0.29	0.42	0.51
2	Entropy	//	0.41	0.27	0.62	0.51
3	Euclidean	//	0.42	0.29	0.60	0.55
4	DTW	//	0.44	0.31	0.68	0.56
5	Heatmap	//	0.45	0.32	0.70	0.57
6	Heatmap	CH	0.45	0.32	0.74	0.63
7	Heatmap	CORR	0.50	0.36	0.82	0.65
8	Heatmap	CH + CORR	**0.51**	**0.36**	**0.85**	**0.69**

Mistake detection result on EPIC-Tent.

Method	Sup. Level	F1	Precision	Recall	AUC
Random	//	0.36	0.29	0.42	0.51
TimeSformer [2]	Fully Supervised	0.49	0.35	0.80	0.67
C2F [35]	Fully Supervised	**0.58**	**0.44**	**0.85**	**0.72**
TrajREC (G) [39]	One-Class	0.40	0.26	0.88	0.51
MoCoDAD (G) [8]	One-Class	0.43	0.27	**0.91**	0.50
TrajREC (H) [39]	One-Class	0.44	0.31	0.76	0.55
MoCoDAD (H) [8]	One-Class	0.46	0.33	0.79	0.60
TrajREC (H+G) [39]	One-Class	0.42	0.29	0.75	0.53
MoCoDAD (H+G) [8]	One-Class	0.43	0.30	0.77	0.56
TrajREC (H+G)* [39]	One-Class	0.47	0.34	0.77	0.63
MoCoDAD (H+G)* [8]	One-Class	0.49	0.35	0.81	0.65
GLC [19]	One-Class	0.46	0.37	0.62	0.66
Ours	One-Class	0.52	0.37	0.85	0.69
Ours + MoCoDAD (H)*	One-Class	**0.54**	**0.41**	0.86	**0.72**
TrajREC (G) [39]	Unsupervised	0.27	0.16	**0.94**	0.50
MoCoDAD (G) [8]	Unsupervised	0.33	0.21	0.88	0.51
TrajREC (H) [39]	Unsupervised	0.40	0.27	0.79	0.58
MoCoDAD (H) [8]	Unsupervised	0.41	0.27	0.86	0.60
MoCoDAD (H+G)* [8]	Unsupervised	0.41	0.27	0.88	0.60
GLC [19]	Unsupervised	0.44	0.33	0.70	0.61
Ours	Unsupervised	0.51	0.36	0.85	0.69
Ours + MoCoDAD (H)*	Unsupervised	**0.52**	**0.37**	0.88	**0.70**

* Late fusion

M. Mazzamuto, A. Furnari, Y. Sato, G.M. Farinella: "Gazing Into Missteps: Leveraging Eye-Gaze for Unsupervised Mistake Detection in Egocentric Videos of Skilled Human Activities by Detecting Unpredictable Gaze." CVPR, 2025.

Results

Then we adapted popular baseline for anomaly detection that used body keypoins , like TrajREC and MoCoDAD, to the mistake detection task considering both gaze and hand trajectory for a fair comparison.

Ablation Study

	Scoring	Fusion	F1	Precision	Recall	AUC
1	Random	//	0.36	0.29	0.42	0.51
2	Entropy	//	0.41	0.27	0.62	0.51
3	Euclidean	//	0.42	0.29	0.60	0.55
4	DTW	//	0.44	0.31	0.68	0.56
5	Heatmap	//	0.45	0.32	0.70	0.57
6	Heatmap	CH	0.45	0.32	0.74	0.63
7	Heatmap	CORR	0.50	0.36	0.82	0.65
8	Heatmap	CH + CORR	**0.51**	**0.36**	**0.85**	**0.69**

Mistake detection result on EPIC-Tent.

Method	Sup. Level	F1	Precision	Recall	AUC
Random	//	0.36	0.29	0.42	0.51
TimeSformer [2]	Fully Supervised	0.49	0.35	0.80	0.67
C2F [35]	Fully Supervised	**0.58**	**0.44**	**0.85**	**0.72**
TrajREC (G) [39]	One-Class	0.40	0.26	0.88	0.51
MoCoDAD (G) [8]	One-Class	0.43	0.27	**0.91**	0.50
TrajREC (H) [39]	One-Class	0.44	0.31	0.76	0.55
MoCoDAD (H) [8]	One-Class	0.46	0.33	0.79	0.60
TrajREC (H+G) [39]	One-Class	0.42	0.29	0.75	0.53
MoCoDAD (H+G) [8]	One-Class	0.43	0.30	0.77	0.56
TrajREC (H+G)* [39]	One-Class	0.47	0.34	0.77	0.63
MoCoDAD (H+G)* [8]	One-Class	0.49	0.35	0.81	0.65
GLC [19]	One-Class	0.46	0.37	0.62	0.66
Ours	One-Class	0.52	0.37	0.85	0.69
Ours + MoCoDAD (H)*	One-Class	**0.54**	**0.41**	0.86	**0.72**
TrajREC (G) [39]	Unsupervised	0.27	0.16	**0.94**	0.50
MoCoDAD (G) [8]	Unsupervised	0.33	0.21	0.88	0.51
TrajREC (H) [39]	Unsupervised	0.40	0.27	0.79	0.58
MoCoDAD (H) [8]	Unsupervised	0.41	0.27	0.86	0.60
MoCoDAD (H+G)* [8]	Unsupervised	0.41	0.27	0.88	0.60
GLC [19]	Unsupervised	0.44	0.33	0.70	0.61
Ours	Unsupervised	0.51	0.36	0.85	0.69
Ours + MoCoDAD (H)*	Unsupervised	**0.52**	**0.37**	0.88	**0.70**

* Late fusion

M. Mazzamuto, A. Furnari, Y. Sato, G.M. Farinella: "Gazing Into Missteps: Leveraging Eye-Gaze for Unsupervised Mistake Detection in Egocentric Videos of Skilled Human Activities by Detecting Unpredictable Gaze." CVPR, 2025.

Results

We then tested a state-of-the-art gaze prediction approach called GLC. Our proposed gaze completion method achieved the best performance.

Ablation Study

	Scoring	Fusion	F1	Precision	Recall	AUC
1	Random	//	0.36	0.29	0.42	0.51
2	Entropy	//	0.41	0.27	0.62	0.51
3	Euclidean	//	0.42	0.29	0.60	0.55
4	DTW	//	0.44	0.31	0.68	0.56
5	Heatmap	//	0.45	0.32	0.70	0.57
6	Heatmap	CH	0.45	0.32	0.74	0.63
7	Heatmap	CORR	0.50	0.36	0.82	0.65
8	Heatmap	CH + CORR	**0.51**	**0.36**	**0.85**	**0.69**

Mistake detection result on EPIC-Tent.

Method	Sup. Level	F1	Precision	Recall	AUC
Random	//	0.36	0.29	0.42	0.51
TimeSformer [2]	Fully Supervised	0.49	0.35	0.80	0.67
C2F [35]	Fully Supervised	**0.58**	**0.44**	**0.85**	**0.72**
TrajREC (G) [39]	One-Class	0.40	0.26	0.88	0.51
MoCoDAD (G) [8]	One-Class	0.43	0.27	**0.91**	0.50
TrajREC (H) [39]	One-Class	0.44	0.31	0.76	0.55
MoCoDAD (H) [8]	One-Class	0.46	0.33	0.79	0.60
TrajREC (H+G) [39]	One-Class	0.42	0.29	0.75	0.53
MoCoDAD (H+G) [8]	One-Class	0.43	0.30	0.77	0.56
TrajREC (H+G)* [39]	One-Class	0.47	0.34	0.77	0.63
MoCoDAD (H+G)* [8]	One-Class	0.49	0.35	0.81	0.65
GLC [19]	One-Class	0.46	0.37	0.62	0.66
Ours	One-Class	0.52	0.37	0.85	0.69
Ours + MoCoDAD (H)*	One-Class	**0.54**	**0.41**	0.86	**0.72**
TrajREC (G) [39]	Unsupervised	0.27	0.16	**0.94**	0.50
MoCoDAD (G) [8]	Unsupervised	0.33	0.21	0.88	0.51
TrajREC (H) [39]	Unsupervised	0.40	0.27	0.79	0.58
MoCoDAD (H) [8]	Unsupervised	0.41	0.27	0.86	0.60
MoCoDAD (H+G)* [8]	Unsupervised	0.41	0.27	0.88	0.60
GLC [19]	Unsupervised	0.44	0.33	0.70	0.61
Ours	Unsupervised	0.51	0.36	0.85	0.69
Ours + MoCoDAD (H)*	Unsupervised	**0.52**	**0.37**	0.88	**0.70**

* Late fusion

M. Mazzamuto, A. Furnari, Y. Sato, G.M. Farinella: "Gazing Into Missteps: Leveraging Eye-Gaze for Unsupervised Mistake Detection in Egocentric Videos of Skilled Human Activities by Detecting Unpredictable Gaze." CVPR, 2025.

Results

The trend in unsupervised settings is similar, but performance is slightly lower due to the inclusion of mistake clips during the training phase.

Ablation Study

	Scoring	Fusion	F1	Precision	Recall	AUC
1	Random	//	0.36	0.29	0.42	0.51
2	Entropy	//	0.41	0.27	0.62	0.51
3	Euclidean	//	0.42	0.29	0.60	0.55
4	DTW	//	0.44	0.31	0.68	0.56
5	Heatmap	//	0.45	0.32	0.70	0.57
6	Heatmap	CH	0.45	0.32	0.74	0.63
7	Heatmap	CORR	0.50	0.36	0.82	0.65
8	Heatmap	CH + CORR	**0.51**	**0.36**	**0.85**	**0.69**

Mistake detection result on EPIC-Tent.

Method	Sup. Level	F1	Precision	Recall	AUC
Random	//	0.36	0.29	0.42	0.51
TimeSformer [2]	Fully Supervised	0.49	0.35	0.80	0.67
C2F [35]	Fully Supervised	**0.58**	**0.44**	**0.85**	**0.72**
TrajREC (G) [39]	One-Class	0.40	0.26	0.88	0.51
MoCoDAD (G) [8]	One-Class	0.43	0.27	**0.91**	0.50
TrajREC (H) [39]	One-Class	0.44	0.31	0.76	0.55
MoCoDAD (H) [8]	One-Class	0.46	0.33	0.79	0.60
TrajREC (H+G) [39]	One-Class	0.42	0.29	0.75	0.53
MoCoDAD (H+G) [8]	One-Class	0.43	0.30	0.77	0.56
TrajREC (H+G)* [39]	One-Class	0.47	0.34	0.77	0.63
MoCoDAD (H+G)* [8]	One-Class	0.49	0.35	0.81	0.65
GLC [19]	One-Class	0.46	0.37	0.62	0.66
Ours	One-Class	0.52	0.37	0.85	0.69
Ours + MoCoDAD (H)*	One-Class	**0.54**	**0.41**	0.86	**0.72**
TrajREC (G) [39]	Unsupervised	0.27	0.16	**0.94**	0.50
MoCoDAD (G) [8]	Unsupervised	0.33	0.21	0.88	0.51
TrajREC (H) [39]	Unsupervised	0.40	0.27	0.79	0.58
MoCoDAD (H) [8]	Unsupervised	0.41	0.27	0.86	0.60
MoCoDAD (H+G)* [8]	Unsupervised	0.41	0.27	0.88	0.60
GLC [19]	Unsupervised	0.44	0.33	0.70	0.61
Ours	Unsupervised	0.51	0.36	0.85	0.69
Ours + MoCoDAD (H)*	Unsupervised	**0.52**	**0.37**	0.88	**0.70**

* Late fusion

M. Mazzamuto, A. Furnari, Y. Sato, G.M. Farinella: "Gazing Into Missteps: Leveraging Eye-Gaze for Unsupervised Mistake Detection in Egocentric Videos of Skilled Human Activities by Detecting Unpredictable Gaze." CVPR, 2025.

Results

Mistake detection result on HoloAssist.

Method	Sup. Level	F1	Precision	Recall	AUC
Random	//	0.04	0.02	0.39	0.50
TimeSformer [2]	Fully Supervised	0.21	0.35	0.13	0.58
C2F [35]	Fully Supervised	**0.38**	**0.37**	**0.40**	**0.65**
TrajREC (G) [39]	One-Class	0.09	0.04	**0.96**	0.50
MoCoDAD (G) [8]	One-Class	0.11	0.06	0.94	0.51
TrajREC (H) [39]	One-Class	0.19	0.11	0.72	0.56
MoCoDAD (H) [8]	One-Class	0.17	0.10	0.71	0.55
TrajREC (H+G) [39]	One-Class	0.13	0.07	0.68	0.52
MoCoDAD (H+G) [8]	One-Class	0.14	0.08	0.62	0.52
TrajREC (H+G)* [39]	One-Class	0.20	0.12	0.71	0.56
MoCoDAD (H+G)* [8]	One-Class	0.21	0.12	0.75	0.57
GLC [19]	One-Class	0.19	0.11	0.56	0.60
Ours	One-Class	0.22	0.14	0.59	0.61
Ours + MoCoDAD (H)*	One-Class	**0.26**	**0.16**	0.73	**0.63**
TrajREC (G) [39]	Unsupervised	0.05	0.03	**0.92**	0.50
MoCoDAD (G) [8]	Unsupervised	0.07	0.04	**0.92**	0.50
TrajREC (H) [39]	Unsupervised	0.11	0.07	0.32	0.56
MoCoDAD (H) [8]	Unsupervised	0.14	0.10	0.25	0.55
MoCoDAD (H+G)* [8]	Unsupervised	0.15	0.11	0.25	0.56
GLC [19]	Unsupervised	0.10	0.06	0.34	0.54
Ours	Unsupervised	0.18	0.12	0.40	0.59
Ours + MoCoDAD (H)*	Unsupervised	**0.21**	**0.15**	0.40	**0.60**

* Late fusion

Mistake detection result on IndustReal.

Method	Sup. Level	F1	Precision	Recall	AUC
Random	//	0.12	0.06	**0.62**	0.51
TimeSformer [2]	Fully Supervised	0.20	0.12	0.35	0.58
C2F [35]	Fully Supervised	**0.31**	**0.29**	0.31	**0.67**
TrajRE(G) [39]	One-Class	0.17	0.09	0.90	0.53
MoCoDAD(G) [8]	One-Class	0.18	0.10	**0.91**	0.55
TrajREC(H) [39]	One-Class	0.21	0.12	0.88	0.57
MoCoDAD(H) [8]	One-Class	0.22	0.13	0.81	0.60
TrajREC(H+G) [39]	One-Class	0.18	0.10	0.86	0.55
MoCoDAD(H+G) [8]	One-Class	0.19	0.11	0.79	0.58
TrajREC(H+G)* [39]	One-Class	0.21	0.12	0.88	0.58
MoCoDAD(H+G)* [8]	One-Class	0.22	0.13	0.82	0.61
GLC [19]	One-Class	0.21	0.15	0.33	0.60
Ours	One-Class	0.24	**0.18**	0.35	0.63
Ours + MoCoDAD (H)*	One-Class	**0.26**	0.17	0.60	**0.65**
TrajREC (G) [39]	Unsupervised	0.11	0.06	**0.92**	0.51
MoCoDAD (G) [8]	Unsupervised	0.11	0.06	**0.92**	0.51
TrajREC (H) [39]	Unsupervised	0.15	0.11	0.28	0.55
MoCoDAD (H) [8]	Unsupervised	0.16	0.12	0.29	0.57
MoCoDAD (H+G)* [8]	Unsupervised	0.17	0.12	0.30	0.57
GLC [19]	Unsupervised	**0.21**	0.15	0.33	0.58
Ours	Unsupervised	**0.21**	**0.16**	0.33	**0.62**
Ours + MoCoDAD (H)*	Unsupervised	0.20	0.15	0.32	0.61

* Late fusion

M. Mazzamuto, A. Furnari, Y. Sato, G.M. Farinella: "Gazing Into Missteps: Leveraging Eye-Gaze for Unsupervised Mistake Detection in Egocentric Videos of Skilled Human Activities by Detecting Unpredictable Gaze." CVPR, 2025.

Results

Mistake detection result on HoloAssist.

Method	Sup. Level	F1	Precision	Recall	AUC
Random	//	0.04	0.02	0.39	0.50
TimeSformer [2]	Fully Supervised	0.21	0.33	0.13	0.58
C2F [35]	Fully Supervised	**0.38**	**0.37**	**0.40**	**0.65**
TrajREC (G) [39]	One-Class	0.09	0.04	**0.96**	0.50
MoCoDAD (G) [8]	One-Class	0.11	0.06	0.94	0.51
TrajREC (H) [39]	One-Class	0.19	0.11	0.72	0.56
MoCoDAD (H) [8]	One-Class	0.17	0.10	0.71	0.55
TrajREC (H+G) [39]	One-Class	0.13	0.07	0.68	0.52
MoCoDAD (H+G) [8]	One-Class	0.14	0.08	0.62	0.52
TrajREC (H+G)* [39]	One-Class	0.20	0.12	0.71	0.56
MoCoDAD (H+G)* [8]	One-Class	0.21	0.12	0.75	0.57
GLC [19]	One-Class	0.19	0.11	0.56	0.60
Ours	One-Class	0.22	0.14	0.59	0.61
Ours + MoCoDAD (H)*	One-Class	**0.26**	**0.16**	0.73	**0.63**
TrajREC (G) [39]	Unsupervised	0.05	0.03	**0.92**	0.50
MoCoDAD (G) [8]	Unsupervised	0.07	0.04	**0.92**	0.50
TrajREC (H) [39]	Unsupervised	0.11	0.07	0.32	0.56
MoCoDAD (H) [8]	Unsupervised	0.14	0.10	0.25	0.55
MoCoDAD (H+G)* [8]	Unsupervised	0.15	0.11	0.25	0.56
GLC [19]	Unsupervised	0.10	0.06	0.34	0.54
Ours	Unsupervised	0.18	0.12	0.40	0.59
Ours + MoCoDAD (H)*	Unsupervised	**0.21**	**0.15**	0.40	**0.60**

* Late fusion

Mistake detection result on IndustReal.

Method	Sup. Level	F1	Precision	Recall	AUC
Random	//	0.12	0.06	**0.62**	0.51
TimeSformer [2]	Fully Supervised	0.20	0.12	0.55	0.58
C2F [35]	Fully Supervised	**0.31**	**0.29**	0.31	**0.67**
TrajRE(G) [39]	One-Class	0.17	0.09	0.90	0.53
MoCoDAD(G) [8]	One-Class	0.18	0.10	**0.91**	0.55
TrajREC(H) [39]	One-Class	0.21	0.12	0.88	0.57
MoCoDAD(H) [8]	One-Class	0.22	0.13	0.81	0.60
TrajREC(H+G) [39]	One-Class	0.18	0.10	0.86	0.55
MoCoDAD(H+G) [8]	One-Class	0.19	0.11	0.79	0.58
TrajREC(H+G)* [39]	One-Class	0.21	0.12	0.88	0.58
MoCoDAD(H+G)* [8]	One-Class	0.22	0.13	0.82	0.61
GLC [19]	One-Class	0.21	0.15	0.33	0.60
Ours	One-Class	0.24	**0.18**	0.35	0.63
Ours + MoCoDAD (H)*	One-Class	**0.26**	0.17	0.60	**0.65**
TrajREC (G) [39]	Unsupervised	0.11	0.06	**0.92**	0.51
MoCoDAD (G) [8]	Unsupervised	0.11	0.06	**0.92**	0.51
TrajREC (H) [39]	Unsupervised	0.15	0.11	0.28	0.55
MoCoDAD (H) [8]	Unsupervised	0.16	0.12	0.29	0.57
MoCoDAD (H+G)* [8]	Unsupervised	0.17	0.12	0.30	0.57
GLC [19]	Unsupervised	**0.21**	0.15	0.33	0.58
Ours	Unsupervised	**0.21**	**0.16**	0.33	**0.62**
Ours + MoCoDAD (H)*	Unsupervised	0.20	0.15	0.32	0.61

* Late fusion

M. Mazzamuto, A. Furnari, Y. Sato, G.M. Farinella: "Gazing Into Missteps: Leveraging Eye-Gaze for Unsupervised Mistake Detection in Egocentric Videos of Skilled Human Activities by Detecting Unpredictable Gaze." CVPR, 2025.

Results

Mistake detection result on HoloAssist.

Method	Sup. Level	F1	Precision	Recall	AUC
Random	//	0.04	0.02	0.39	0.50
TimeSformer [2]	Fully Supervised	0.21	0.35	0.13	0.58
C2F [35]	Fully Supervised	**0.38**	**0.37**	**0.40**	**0.65**
TrajREC (G) [39]	One-Class	0.09	0.04	**0.96**	0.50
MoCoDAD (G) [8]	One-Class	0.11	0.06	0.94	0.51
TrajREC (H) [39]	One-Class	0.19	0.11	0.72	0.56
MoCoDAD (H) [8]	One-Class	0.17	0.10	0.71	0.55
TrajREC (H+G) [39]	One-Class	0.13	0.07	0.68	0.52
MoCoDAD (H+G) [8]	One-Class	0.14	0.08	0.62	0.52
TrajREC (H+G)* [39]	One-Class	0.20	0.12	0.71	0.56
MoCoDAD (H+G)* [8]	One-Class	0.21	0.12	0.75	0.57
GLC [19]	One-Class	0.19	0.11	0.56	0.60
Ours	One-Class	0.22	0.14	0.59	0.61
Ours + MoCoDAD (H)*	One-Class	**0.26**	**0.16**	0.73	**0.63**
TrajREC (G) [39]	Unsupervised	0.05	0.03	**0.92**	0.50
MoCoDAD (G) [8]	Unsupervised	0.07	0.04	**0.92**	0.50
TrajREC (H) [39]	Unsupervised	0.11	0.07	0.32	0.56
MoCoDAD (H) [8]	Unsupervised	0.14	0.10	0.25	0.55
MoCoDAD (H+G)* [8]	Unsupervised	0.15	0.11	0.25	0.56
GLC [19]	Unsupervised	0.10	0.06	0.34	0.54
Ours	Unsupervised	0.18	0.12	0.40	0.59
Ours + MoCoDAD (H)*	Unsupervised	**0.21**	**0.15**	0.40	**0.60**

* Late fusion

Mistake detection result on IndustReal.

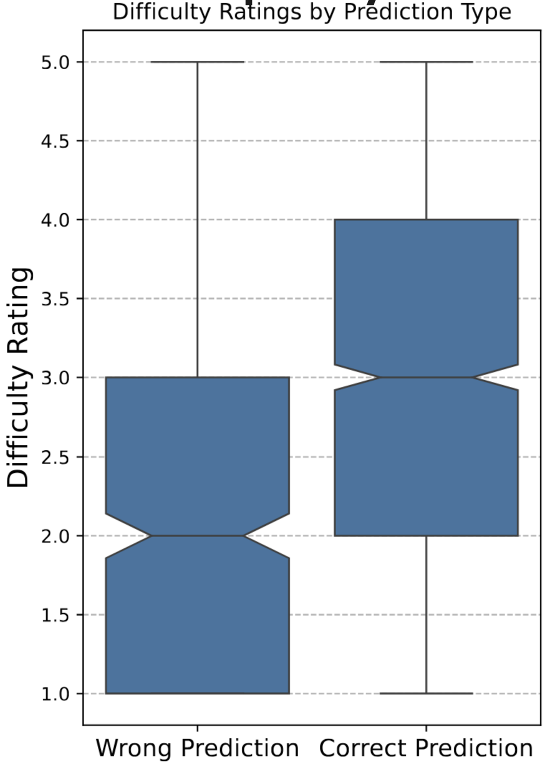
Method	Sup. Level	F1	Precision	Recall	AUC
Random	//	0.12	0.06	**0.62**	0.51
TimeSformer [2]	Fully Supervised	0.20	0.12	0.35	0.58
C2F [35]	Fully Supervised	**0.31**	**0.29**	0.31	**0.67**
TrajRE(G) [39]	One-Class	0.17	0.09	0.90	0.53
MoCoDAD(G) [8]	One-Class	0.18	0.10	**0.91**	0.55
TrajREC(H) [39]	One-Class	0.21	0.12	0.88	0.57
MoCoDAD(H) [8]	One-Class	0.22	0.13	0.81	0.60
TrajREC(H+G) [39]	One-Class	0.18	0.10	0.86	0.55
MoCoDAD(H+G) [8]	One-Class	0.19	0.11	0.79	0.58
TrajREC(H+G)* [39]	One-Class	0.21	0.12	0.88	0.58
MoCoDAD(H+G)* [8]	One-Class	0.22	0.13	0.82	0.61
GLC [19]	One-Class	0.21	0.15	0.33	0.60
Ours	One-Class	0.24	**0.18**	0.35	0.63
Ours + MoCoDAD (H)*	One-Class	**0.26**	0.17	0.60	**0.65**
TrajREC (G) [39]	Unsupervised	0.11	0.06	**0.92**	0.51
MoCoDAD (G) [8]	Unsupervised	0.11	0.06	**0.92**	0.51
TrajREC (H) [39]	Unsupervised	0.15	0.11	0.28	0.55
MoCoDAD (H) [8]	Unsupervised	0.16	0.12	0.29	0.57
MoCoDAD (H+G)* [8]	Unsupervised	0.17	0.12	0.30	0.57
GLC [19]	Unsupervised	**0.21**	0.15	0.33	0.58
Ours	Unsupervised	**0.21**	**0.16**	0.33	**0.62**
Ours + MoCoDAD (H)*	Unsupervised	0.20	0.15	0.32	0.61

* Late fusion

M. Mazzamuto, A. Furnari, Y. Sato, G.M. Farinella: "Gazing Into Missteps: Leveraging Eye-Gaze for Unsupervised Mistake Detection in Egocentric Videos of Skilled Human Activities by Detecting Unpredictable Gaze." CVPR, 2025.

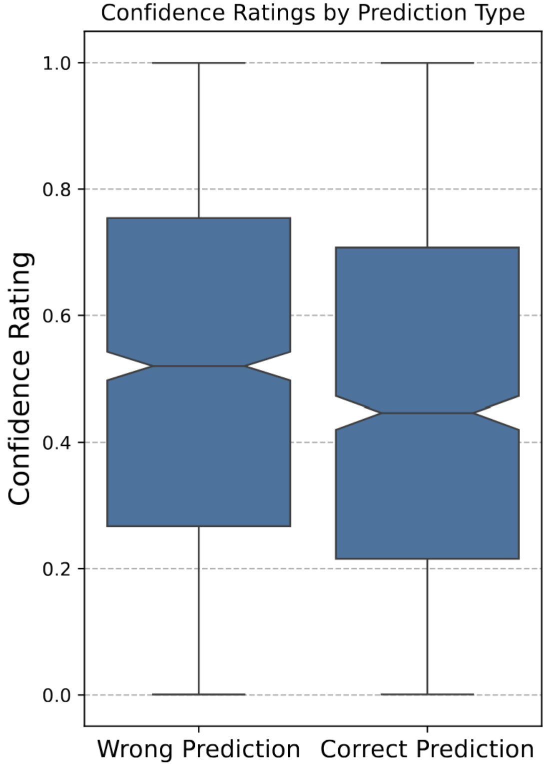
Contribution of gaze across different scenarios

Action Complexity



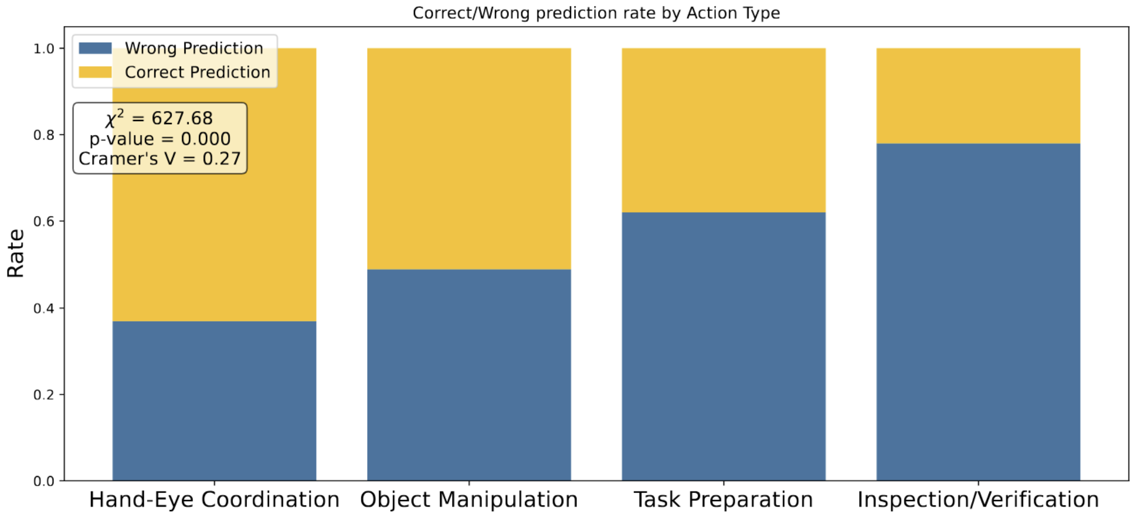
(a)

Self-Rated Confidence



(b)

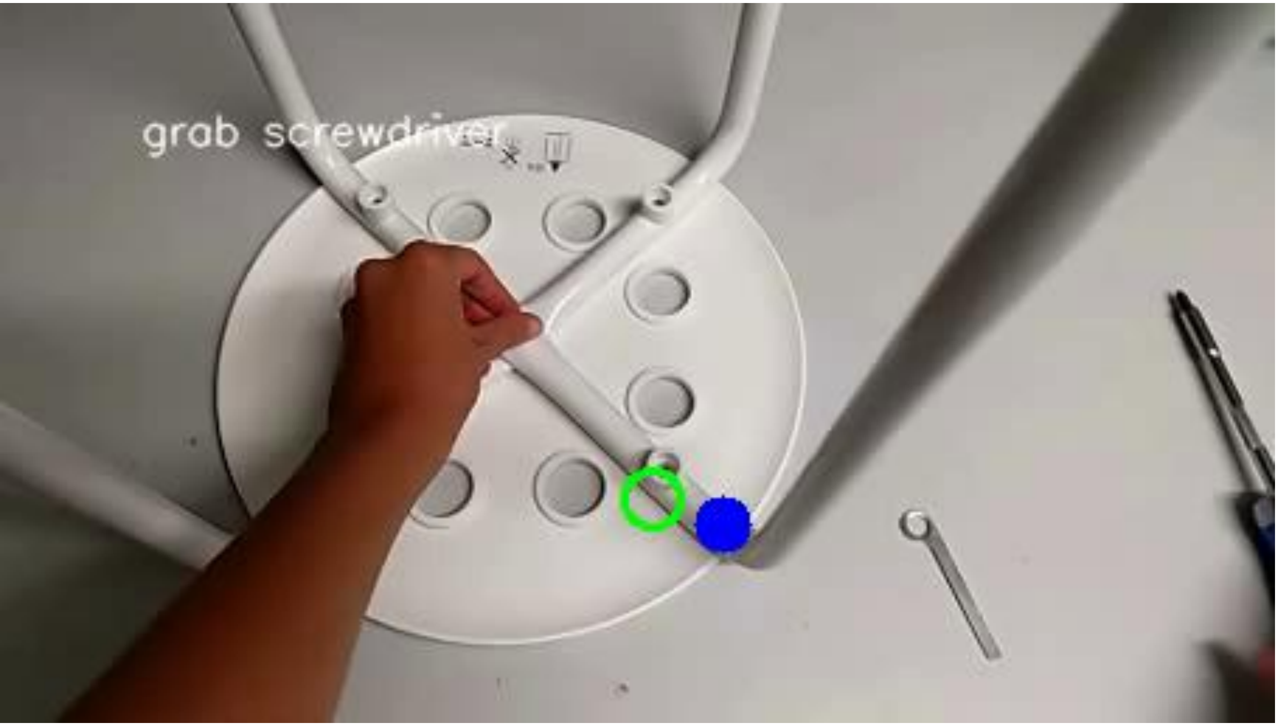
Action Type



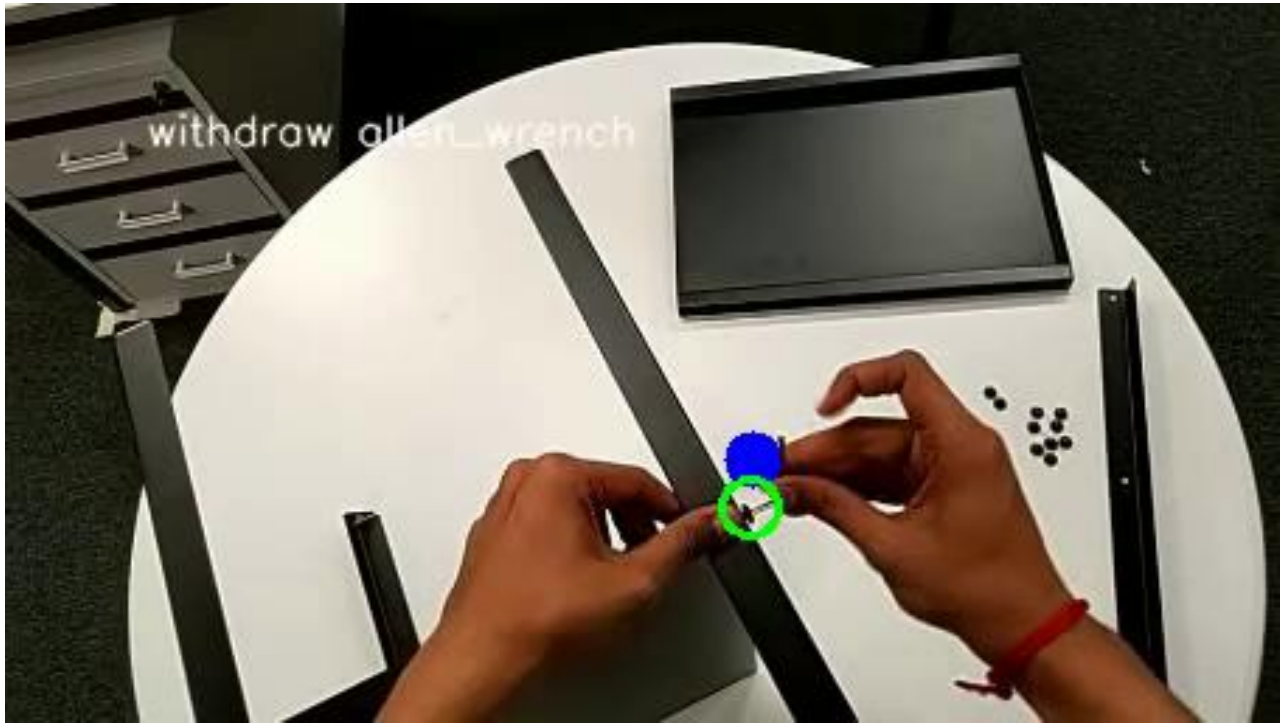
M. Mazzamuto, A. Furnari, Y. Sato, G.M. Farinella: "Gazing Into Missteps: Leveraging Eye-Gaze for Unsupervised Mistake Detection in Egocentric Videos of Skilled Human Activities by Detecting Unpredictable Gaze." CVPR, 2025.

Qualitative example

Correct



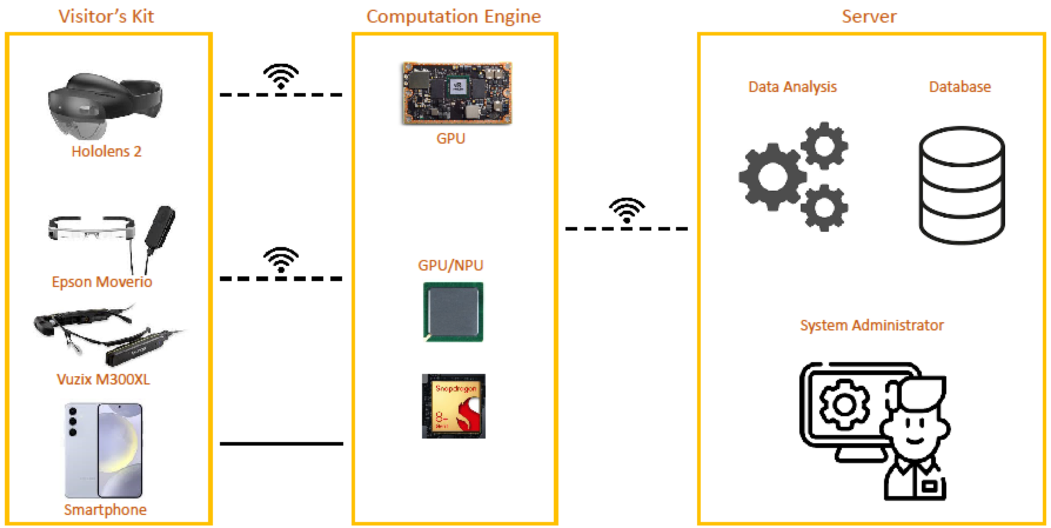
Mistake



M. Mazzamuto, A. Furnari, Y. Sato, G.M. Farinella: "Gazing Into Missteps: Leveraging Eye-Gaze for Unsupervised Mistake Detection in Egocentric Videos of Skilled Human Activities by Detecting Unpredictable Gaze." CVPR, 2025.

Building procedural assistant with VLLM

Gaze can be used to build intelligent assistants that understand what the user is focusing on. As demonstrated in the VALUE system, gaze can be integrated to support real-time human–object interactions, providing contextual information in a museum related to the observed object.



M. Mazzamuto et al., "VALUE: Visual Analysis for Location and Understanding of Environment," 2024 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE),


Ego-EXTRA: Egocentric dataset of EXpert-TRAinee assistance

Università di Catania

Every day a new MLLM is born..

Star 24
Parrot: Multilingual Visual

Star 207
Ovis: Structural Embeddin
Multimodal Large Languag

Star 90
Matryoshka Query Transfo
Language Models

Star 100
ConvLLaVA: Hierarchical Ba
Encoder for Large Multimo

Star 98
Meteor: Mamba-based Tra
Large Language and Vision

Star 1.2k
OMG-LLaVA: Bridging Image-le
Pixel-level Reasoning and Unde

Star 1.7k
Cambrian-1: A Fully Open, Visio
Exploration of Multimodal LLM

Star 293
Long Context Transfer from Lan

Star 208
Unveiling Encoder-Free Vision-

Star 128
Beyond LLaVA-HD: Diving into
Large Multimodal Models

Star 728
VideoLLaMA 2: Advancing Spat
Modeling and Audio Understan

Star 2.4k
LLaVA-OneVision: Easy

Star 12k
MiniCPM-V: A GPT-4V

VILA^2: VILA Augment

SlowFast-LLaVA: A Stron
Video Large Language

EVLM: An Efficient Visio
Visual Understanding

Star 2.5k
InternLM-XComposer-2
Language Model Suppo
Input and Output

Star 2.1k
Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution

Star 119
LongLLaVA: Scaling Multi-modal LLMs to 1000 Images Efficiently via Hybrid Architecture

Star 405
EAGLE: Exploring The Design Space for Multimodal LLMs with Mixture of Encoders

Star 2.2k
mPLUG-Owl3: Towards Long Image-Sequence Understanding in Multi-Modal Large Language Models

Star 767
VITA: Towards Open-Source Interactive Omni Multimodal LLM

Source: https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models

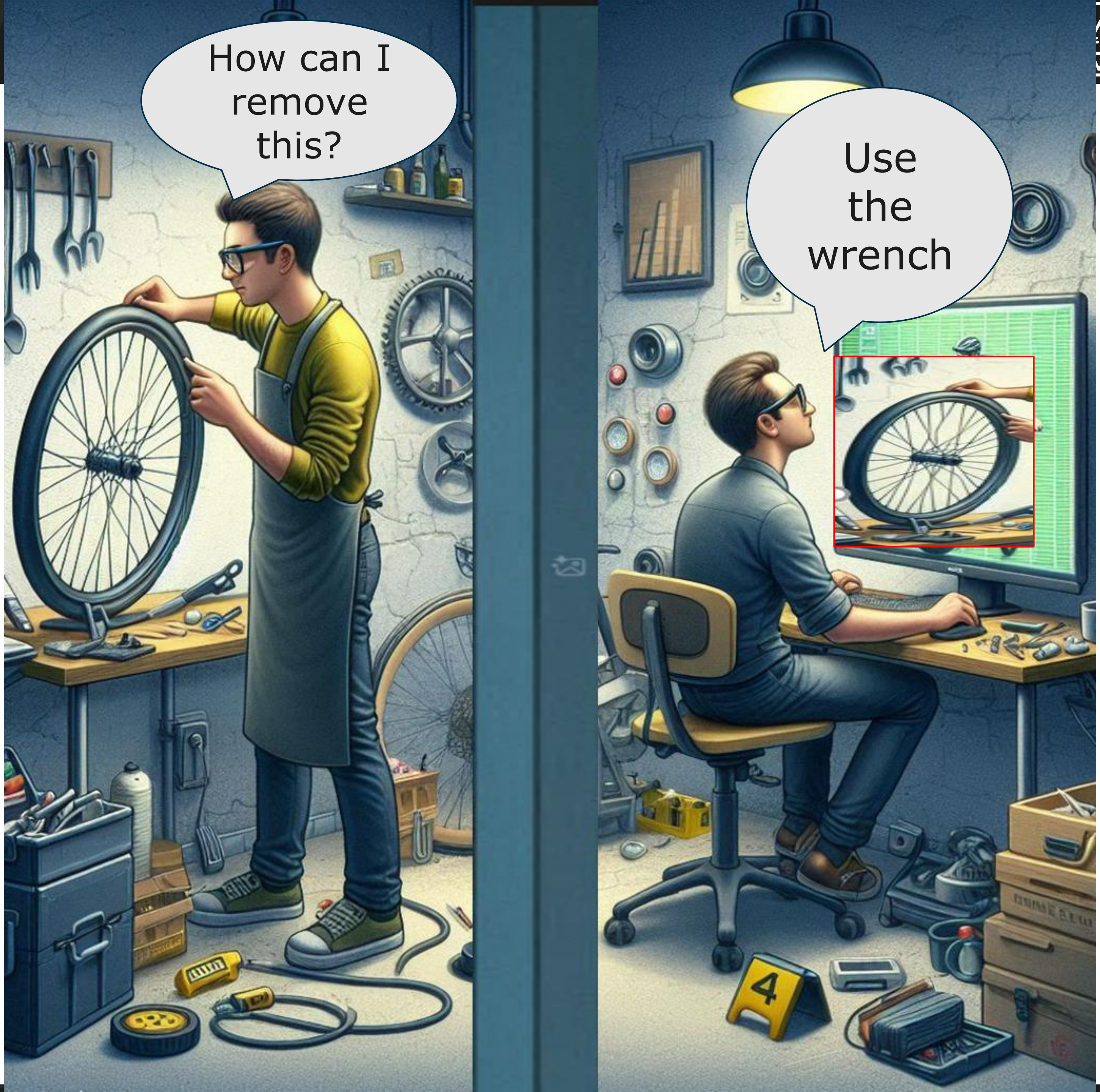
Conversations are missing

The Ideal Personal Assistant

A New Dataset

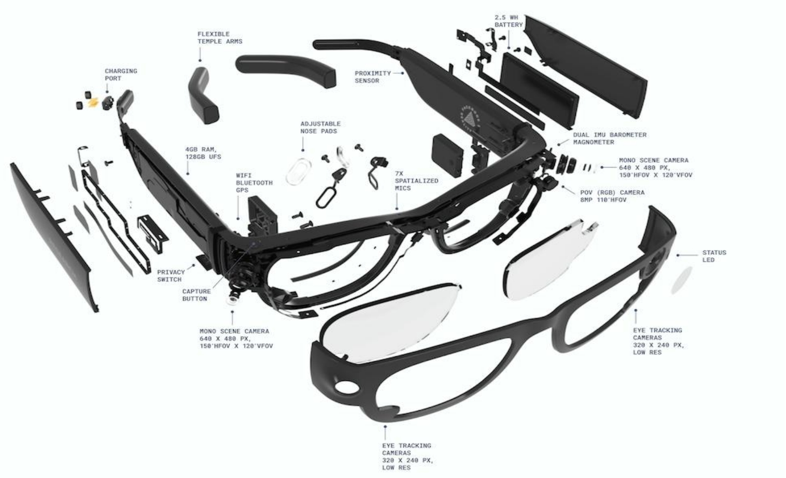
Egocentric videos of subjects engaging different procedural activities in which they are not expert or not very expert (i.e., Trainees);

Conversations between trainees and experts happen naturally during the collection.



Multimodalities

Trainee: RGB videos, Gaze, SLAM, Hand poses



Expert: Gaze

Trainee-Expert: Text (transcriptions)

Expert: Now you need to fix the electric board to the working area

Trainee: With the screwdriver?

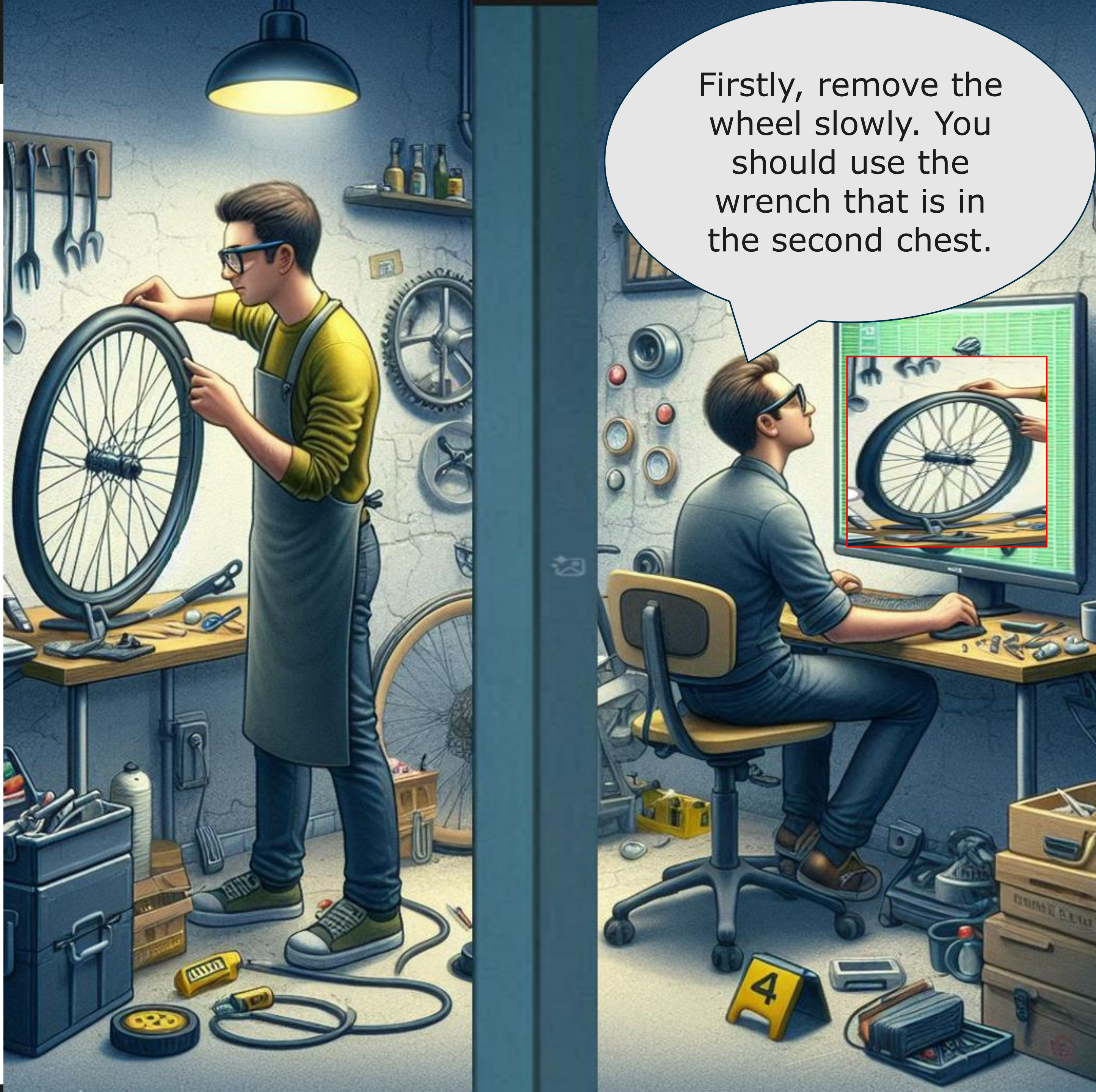
Expert: Yes

Acquisition Protocols

Pro-active:

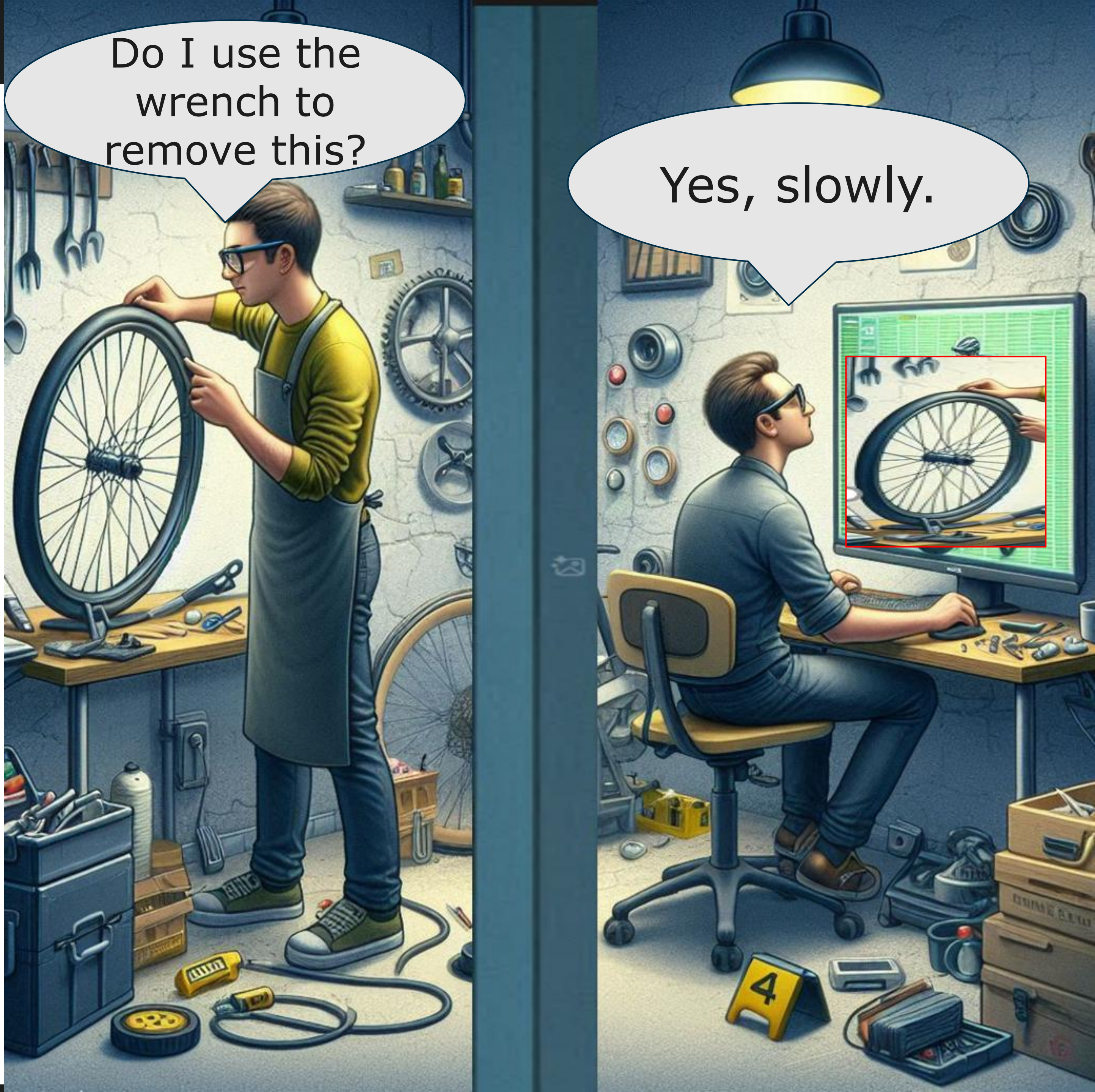
At the beginning, the trainee has not a knowledge about the environment, the objects and the procedure to perform.

The expert speaks freely with the trainee, suggesting next steps, instructions and anything that may be useful.



Non Pro-active:

The expert may answer only the trainee’s questions or alert him if a mistake is happening.

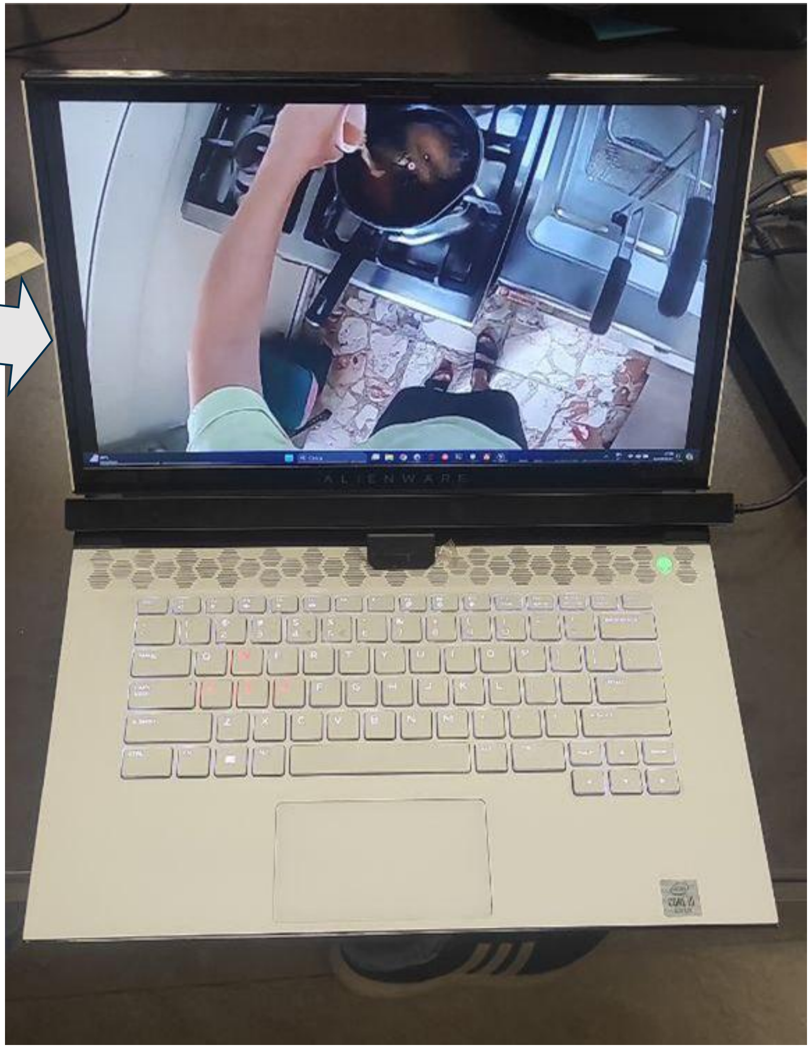


Devices

Trainees wear Aria Glasses and perform the activities

Experts observe the environment from the trainees' point of view

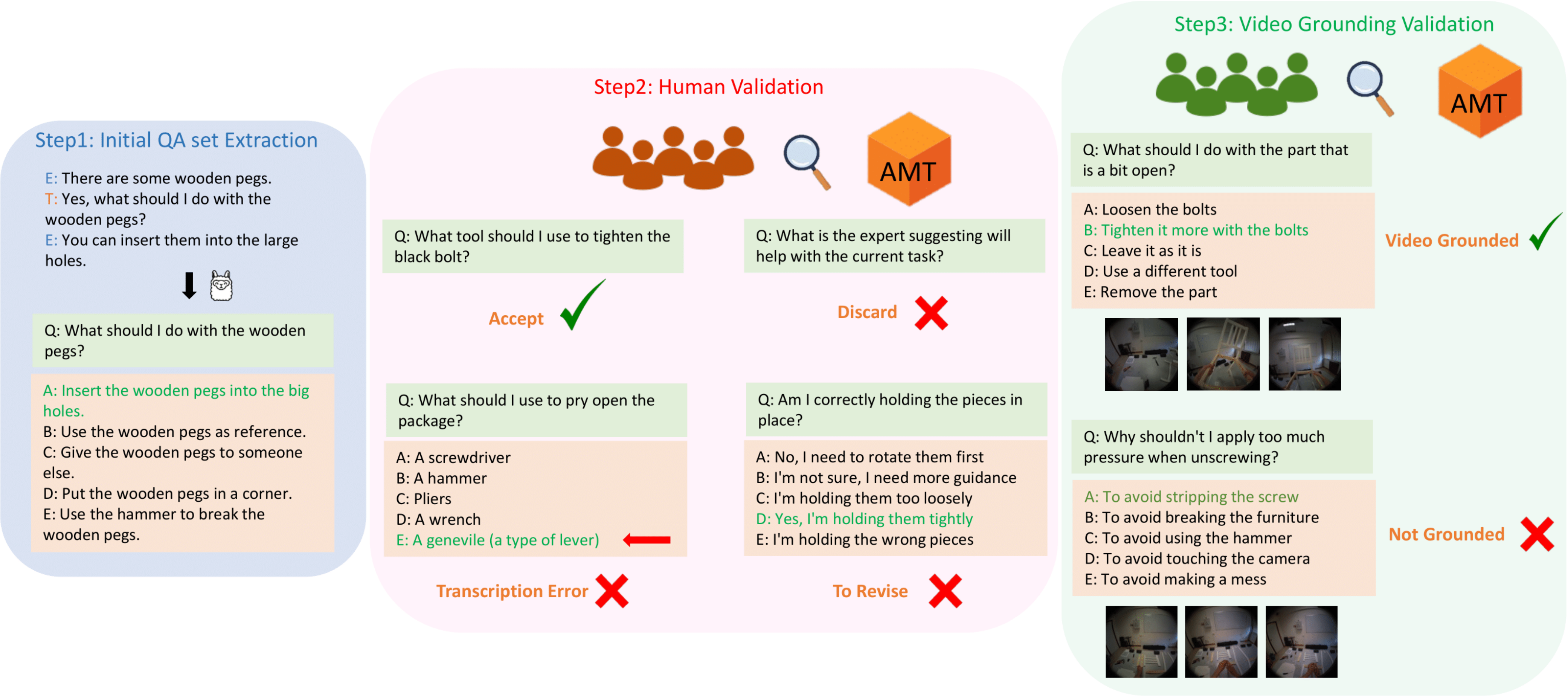
Streaming



Synchronized Gaze

Trainee-Expert Conversation

Q&A validation



F. Ragusa, M. Mazzamuto, R. Forte, I. D'Ambra, J. Fort, J. Engel, A. Furnari, G.M. Farinella: "Ego-EXTRA: Video-Language Egocentric Dataset for EXpert-TRAinee Assistance." WACV, 2026.

The Dataset

50 Hours

Expanding to 100

4 Scenarios

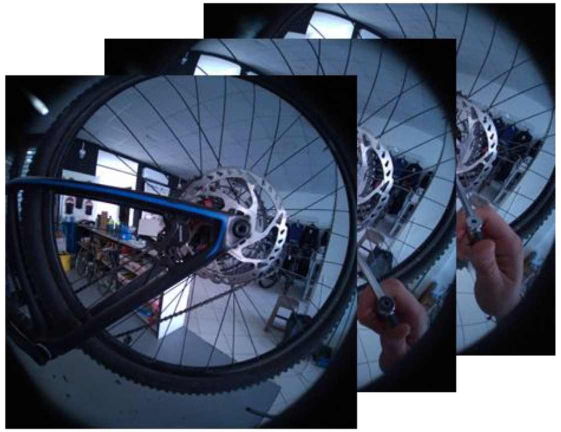
Expanding to 5

1-2 experts for each scenario

10-20 real trainees for each scenario

F. Ragusa, M. Mazzamuto, R. Forte, I. D'Ambra, J. Fort, J. Engel, A. Furnari, G.M. Farinella: "Ego-EXTRA: Video-Language Egocentric Dataset for EXpert-TRAinee Assistance." WACV, 2026.

Multiple-Choice Question Answering



Video Clip

"Do I need to worry that the wheel might fall?"

Trainee's question

Input

MiniGPT4-video → C

Video-LLama → A

Expert (GT) → A

A) "No, not at this moment. Now, hold it like that. "

B) "Maybe we should stop and secure everything again to be absolutely sure."

C) "No, but it's better to use additional supports or have someone assist you just in case."

D) "No, just let go and see if it stays in place."

F. Ragusa, M. Mazzamuto, R. Forte, I. D'Ambra, J. Fort, J. Engel, A. Furnari, G.M. Farinella: "Ego-EXTRA: Video-Language Egocentric Dataset for EXpert-TRAinee Assistance." WACV, 2026.

Human Baseline

5 What is the purpose of the wooden pins? *

- ○ To attach the seat to the chair
- ○ To hold the sticks together
- ○ To tighten the screws
- ○ To loosen the bolts

Results

	Model	Bike Workshop	Bakery	Assembly	Kitchen	Avg.
Language Only	Llama 3.1 Instruct 8B	20.20	25.71	20.11	25.00	21.30
	Llama 3.1 Instruct 70B	20.20	21.43	20.11	36.11	21.50
	Llama 3.3 Instruct Turbo	23.74	21.43	17.99	30.56	21.70
	Qwen 2.5 Instruct 72B	27.27	27.14	21.16	22.22	24.54
	DeepSeek-R1 Turbo	21.21	28.57	21.16	22.22	22.31
Video-Language	MiniGPT4-video	30.00	29.55	27.93	41.30	30.03
	LLaVa Video	27.78	38.57	29.10	27.78	29.82
	LLaVa-OneVision	**38.89**	**42.86**	**42.86**	**44.44**	**41.38**
	Qwen 2.5-VL	29.80	31.43	28.57	25.00	29.21
	Sample Human Baseline	87.50	90.91	100	81.82	89.65

Table 2. Results on the proposed VQA benchmark.

F. Ragusa, M. Mazzamuto, R. Forte, I. D'Ambra, J. Fort, J. Engel, A. Furnari, G.M. Farinella: "Ego-EXTRA: Video-Language Egocentric Dataset for EXpert-TRAinee Assistance." WACV, 2026.

Results

	Model	Bike Workshop	Bakery	Assembly	Kitchen	Avg.
Language Only	Llama 3.1 Instruct 8B	20.20	25.71	20.11	25.00	21.30
	Llama 3.1 Instruct 70B	20.20	21.43	20.11	36.11	21.50
	Llama 3.3 Instruct Turbo	23.74	21.43	17.99	30.56	21.70
	Qwen 2.5 Instruct 72B	27.27	27.14	21.16	22.22	24.54
	DeepSeek-R1 Turbo	21.21	28.57	21.16	22.22	22.31
Video-Language	MiniGPT4-video	<u>30.00</u>	29.55	27.93	<u>41.30</u>	<u>30.03</u>
	LLaVa Video	27.78	<u>38.57</u>	<u>29.10</u>	27.78	29.82
	LLaVa-OneVision	**38.89**	**42.86**	**42.86**	**44.44**	**41.38**
	Qwen 2.5-VL	29.80	31.43	28.57	25.00	29.21
	Sample Human Baseline	87.50	90.91	100	81.82	89.65

Table 2. Results on the proposed VQA benchmark.

F. Ragusa, M. Mazzamuto, R. Forte, I. D’Ambra, J. Fort, J. Engel, A. Furnari, G.M. Farinella: “Ego-EXTRA: Video-Language Egocentric Dataset for EXpert-TRAinee Assistance.” WACV, 2026.

Ego-EXTRA Data

https://fpv-iplab.github.io/Ego-EXTRA/

Ego-EXTRA

Video-Language Egocentric Dataset for EXpert-TRAinee assistance

Francesco Ragusa*[1], Michele Mazzamuto*[1], Rosario Forte[1], Irene D'Ambra[1], James Fort[2], Jakob Engel[2], Antonino Furnari[1], Giovanni Maria Farinella[1]

* Co-first authors

[1] University of Catania

[2] META Reality Labs

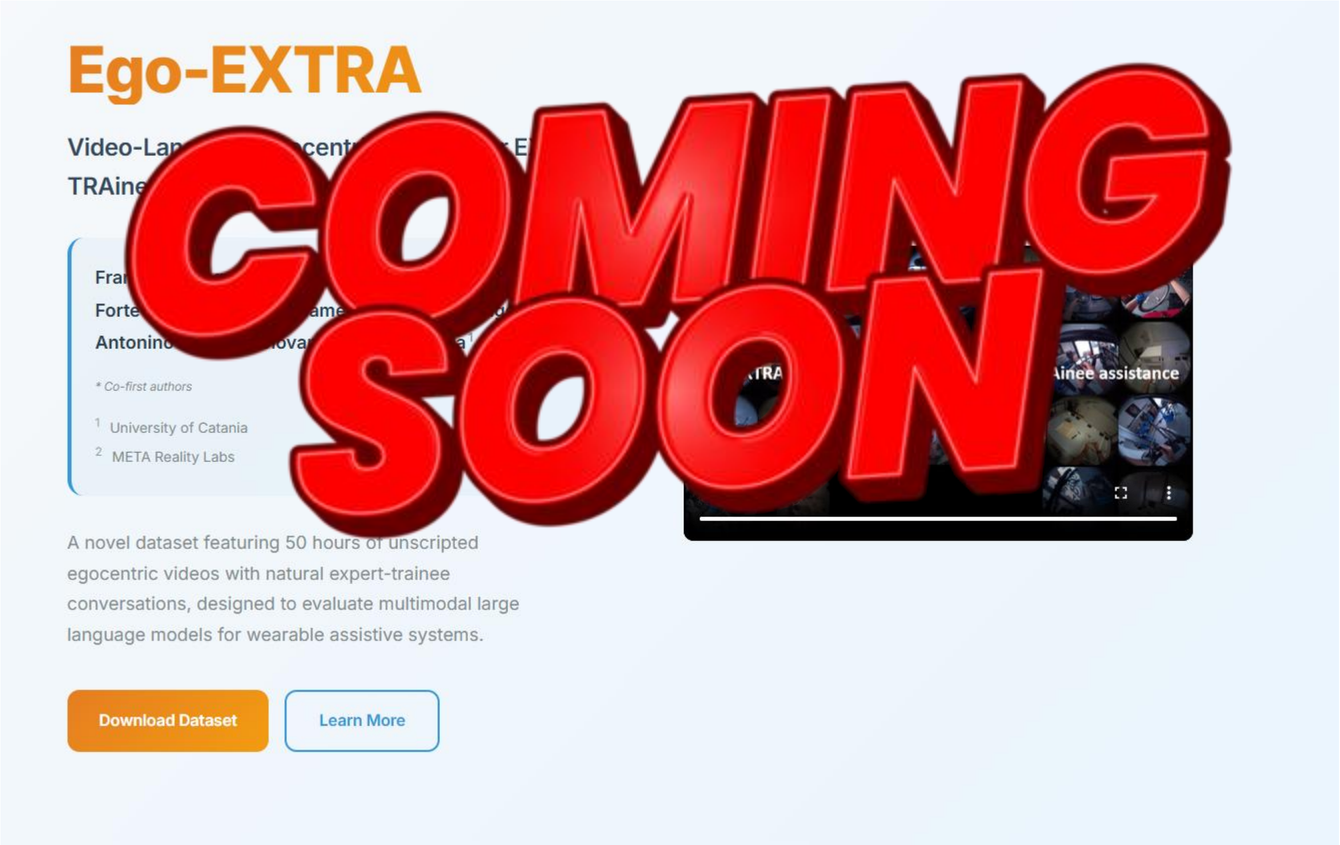
A novel dataset featuring 50 hours of unscripted egocentric videos with natural expert-trainee conversations, designed to evaluate multimodal large language models for wearable assistive systems.

Download Dataset

Learn More

F. Ragusa, M. Mazzamuto, R. Forte, I. D'Ambra, J. Fort, J. Engel, A. Furnari, G.M. Farinella: "Ego-EXTRA: Video-Language Egocentric Dataset for EXpert-TRAinee Assistance." WACV, 2026.

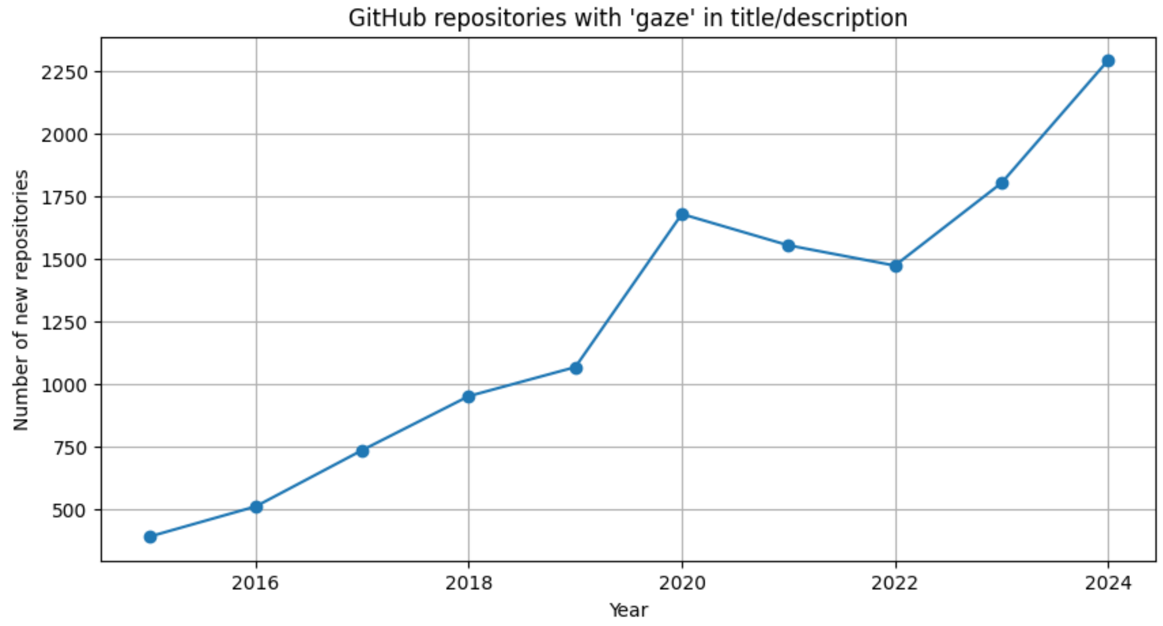
https://fpv-iplab.github.io/Ego-EXTRA/



F. Ragusa, M. Mazzamuto, R. Forte, I. D'Ambra, J. Fort, J. Engel, A. Furnari, G.M. Farinella: "Ego-EXTRA: Video-Language Egocentric Dataset for EXpert-TRAinee Assistance." WACV, 2026.

Open Challenges and Future Directions

- **Robustness & Generalization**: Making gaze estimation reliable across diverse users, tasks, and environments.
- **Calibration-Free Tracking**: Reducing or eliminating the need for explicit calibration.
- **Multimodal Integration**: Combining gaze with signals like hands, head, speech, and physiological cues.
- **Long-Term Understanding**: Modeling attention shifts and intentions over extended activities.



Thank You!

Egocentric Vision: Exploring User-Centric Perspectives

Michele Mazzamuto

LIVE Group @ UNICT - https://iplab.dmi.unict.it/live/

Next Vision - http://www.nextvisionlab.it/

Department of Mathematics and Computer Science - University of Catania

michele.mazzamuto@phd.unict.it - https://mikes95.github.io/