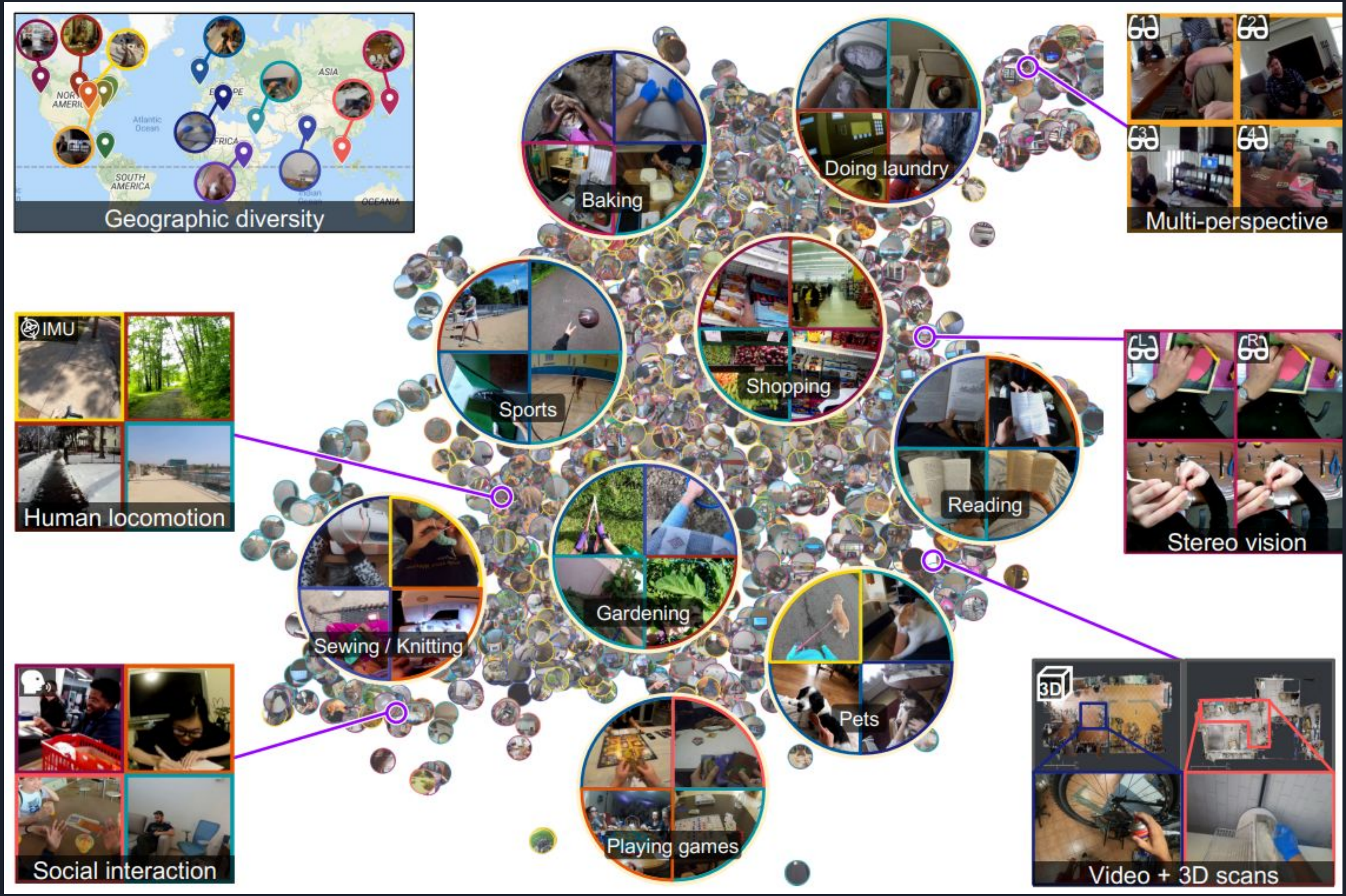
# Ego4D: Around the World in 3,000 Hours of Egocentric Video

Jeff Zhuo & Wei Shan


Geographic diversity
NORTH AMERICA
Atlantic Ocean
EUROPE
AFRICA
ASIA
SOUTH AMERICA
Indian Ocean
OCEANIA
Baking
Doing laundry
Multi-perspective
1
2
3
4
IMU
Sports
Shopping
Human locomotion
Reading
L
R
Stereo vision
Gardening
Sewing / Knitting
Pets
Social interaction
Playing games
3D
Video + 3D scans

Carnegie
Mellon
University

SICILIAE STVDIVM GENERALE 1434

Università
di Catania

NUS
National University
of Singapore

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

東京大学
THE UNIVERSITY OF TOKYO

University of
BRISTOL

INDIANA UNIVERSITY
BLOOMINGTON

UNIVERSITY
OF MINNESOTA

Penn
UNIVERSITY of PENNSYLVANIA

GT
Georgia Institute
of Technology

INTERNATIONAL INSTITUTE OF
INFORMATION TECHNOLOGY
HYDERABAD

Carnegie
Mellon
University
Africa

Universidad de
los Andes
Colombia

FACEBOOK AI

# Motivation

- Lack of large first-person video datasets
- Fuel progress in video understanding
  - Specifically for egocentric videos

# Related Work

3rd Person Video Dataset

- Kinetics
- AVA
- UCF
- ActivityNet
- HowTo100M

*Video Sampled from ActivityNet*

# Related Work

## Egocentric Video Dataset

- EPIC-Kitchens
- UT Ego
- ADL
- Charades-Ego
- EGTEA

*Demo from EPIC-Kitchens*

# Comparison

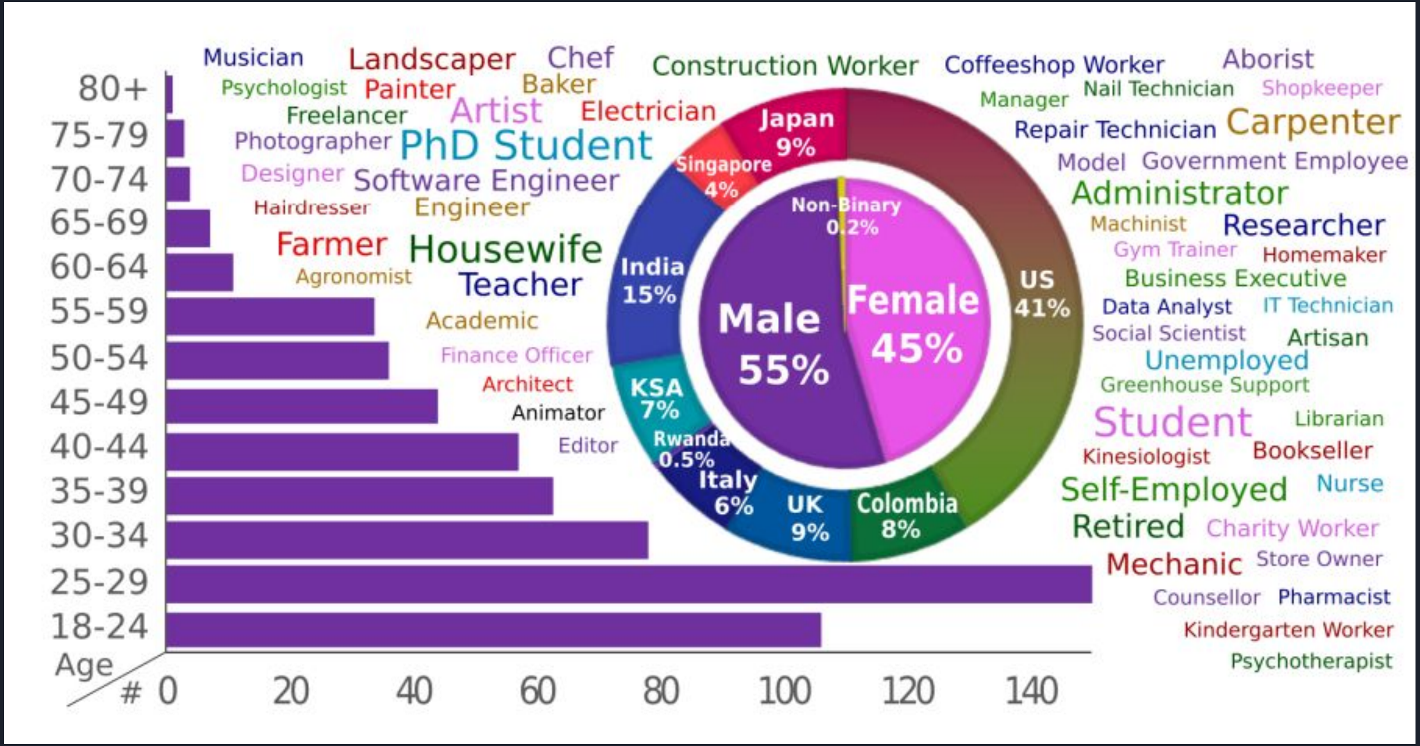
## Ego4D

- 3670 hours
- 931 unique camera wearers
- Hundreds of different environment
- 74 cities worldwide

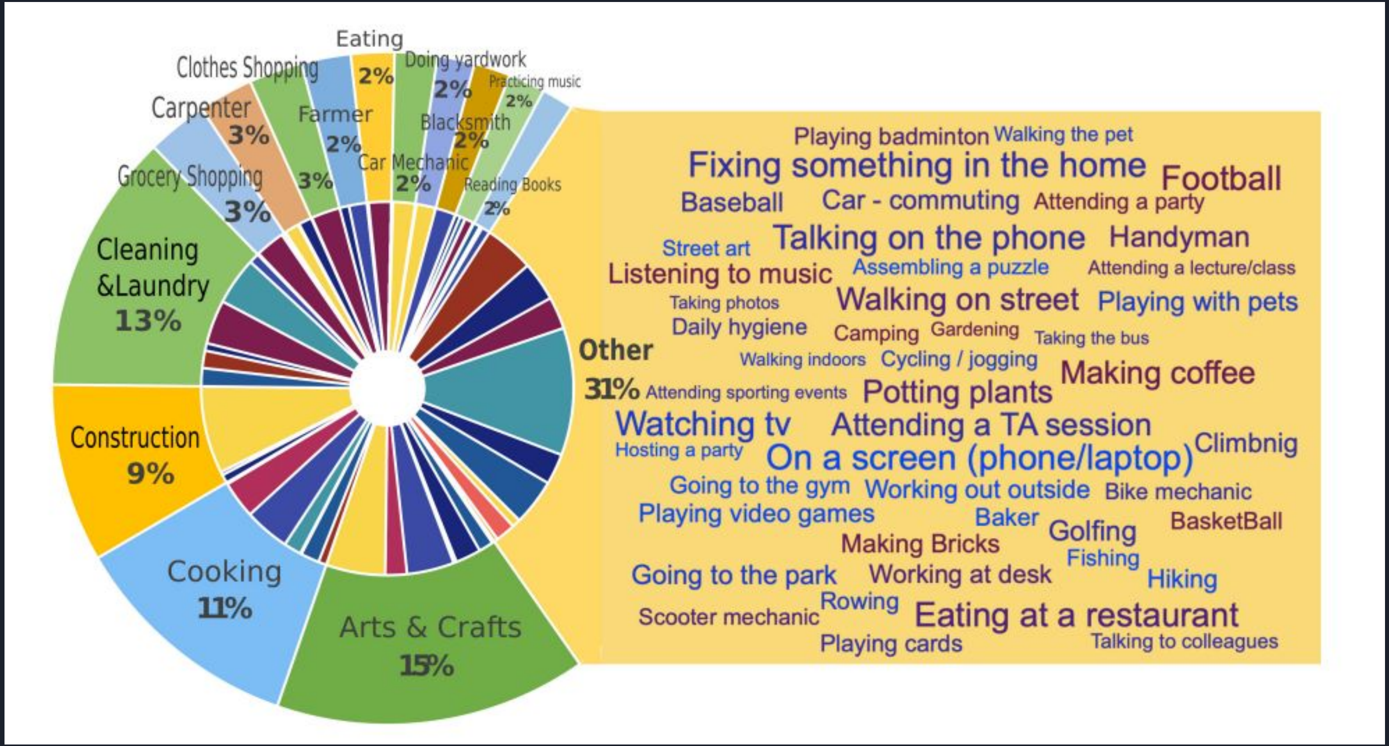
## Other Egocentric Datasets

- 100 hours
- 71 unique camera wearers
- One or dozen different environments
- One or few cities

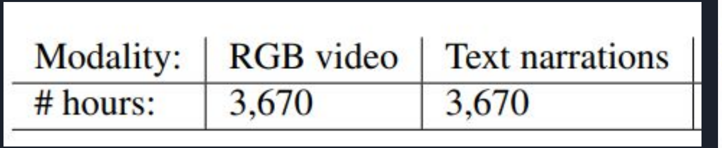
# Collection Diversity



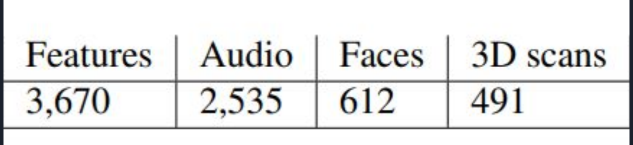
# Scenario Composition



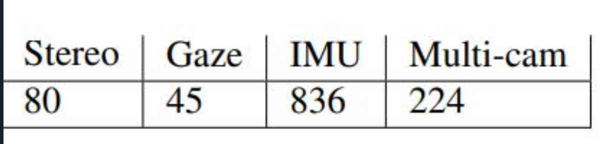
# Cameras and Modality



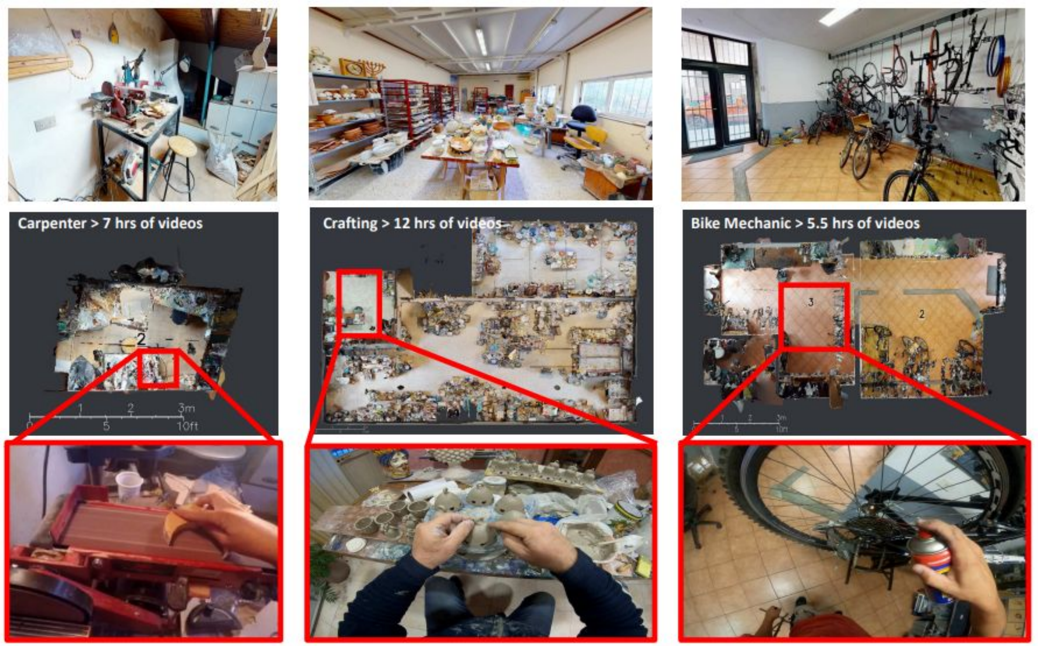
| Modality: | RGB video | Text narrations |
|---|---|---|
| # hours: | 3,670 | 3,670 |



| Features | Audio | Faces | 3D scans |
|---|---|---|---|
| 3,670 | 2,535 | 612 | 491 |



| Stereo | Gaze | IMU | Multi-cam |
|---|---|---|---|
| 80 | 45 | 836 | 224 |



Carpenter > 7 hrs of videos

Crafting > 12 hrs of videos

Bike Mechanic > 5.5 hrs of videos

# Cameras and Modality

# Potential Biases

- 74 Locations worldwide
- More urban and college towns
- COVID-19
- Battery Life: active footage
- Annotation Bias

# Privacy and Ethics

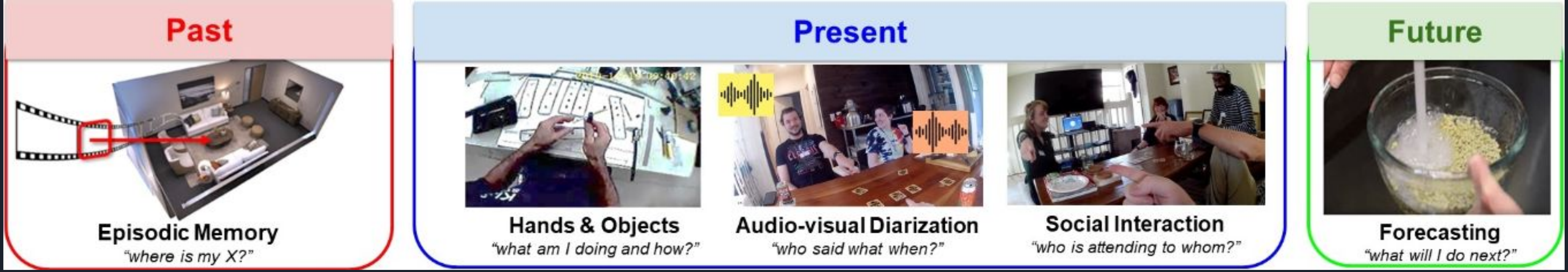
Privacy and ethics policy vary by partner, but all must include the following:

| | |
|---|---|
| University Research Standard | Informed Consent |
| Respect the rights of others | De-identification |

# Accessibility

- Precomputed features from SlowFast w. ResNet 101 backbone
- Mini-set to download

# Benchmark Suite



# Benchmark Suite(Episodic Memory)

- Motivation
- Task definition
    - Visual Query
    - Language Query
    - Moments Query
- Annotation

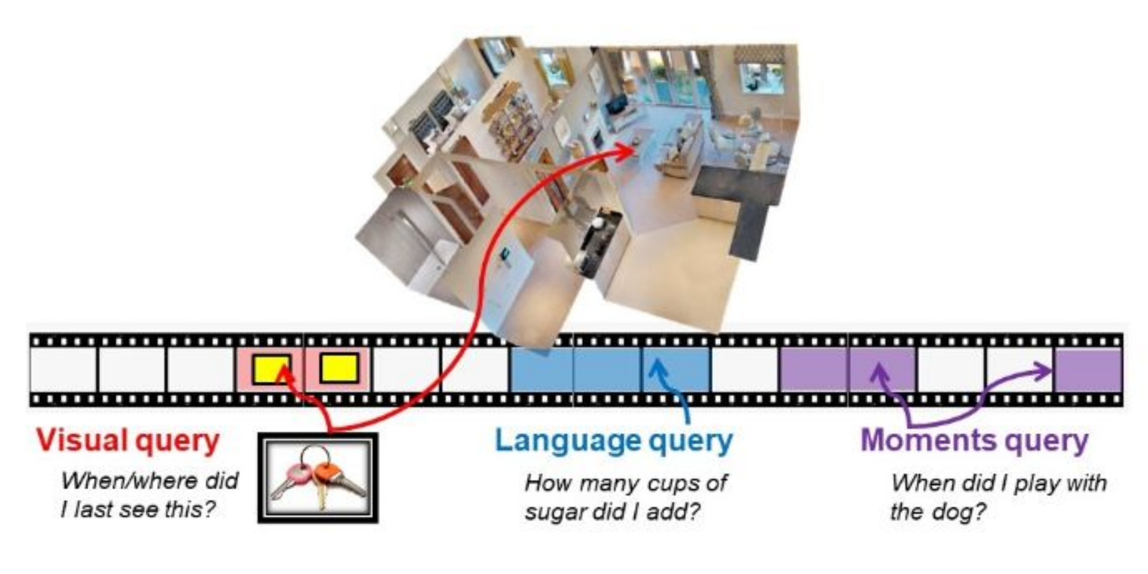
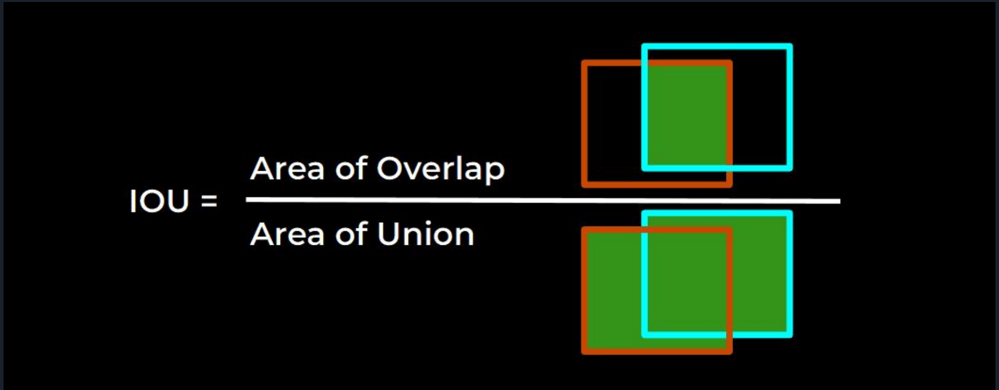

Visual query
When/where did I last see this?
Language query
How many cups of sugar did I add?
Moments query
When did I play with the dog?

Figure 7. Episodic Memory's three query types

# Benchmark Suite(Episodic Memory)

- Evaluation
  - Natural Language Query
    - top-k recall at a certain temporal intersection over union (tIoU) threshold
    - AKA The percentage of times at least one of the top k predicted candidates have an intersection-over-union (IoU) of at least m.



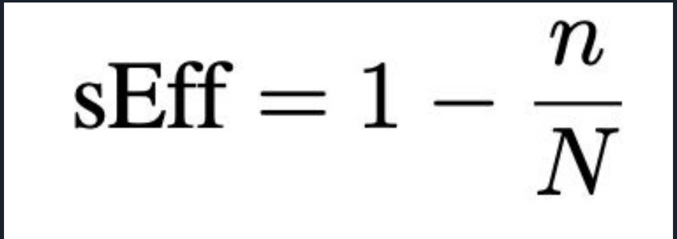
# Benchmark Suite(Episodic Memory)

- Evaluation
  - Moments Query
    - mAP at multiple tIoU thresholds, as well as top-kx recall

$$mAP = \frac{1}{k}\sum_{i}^{k} AP_i$$

# Benchmark Suite(Episodic Memory)

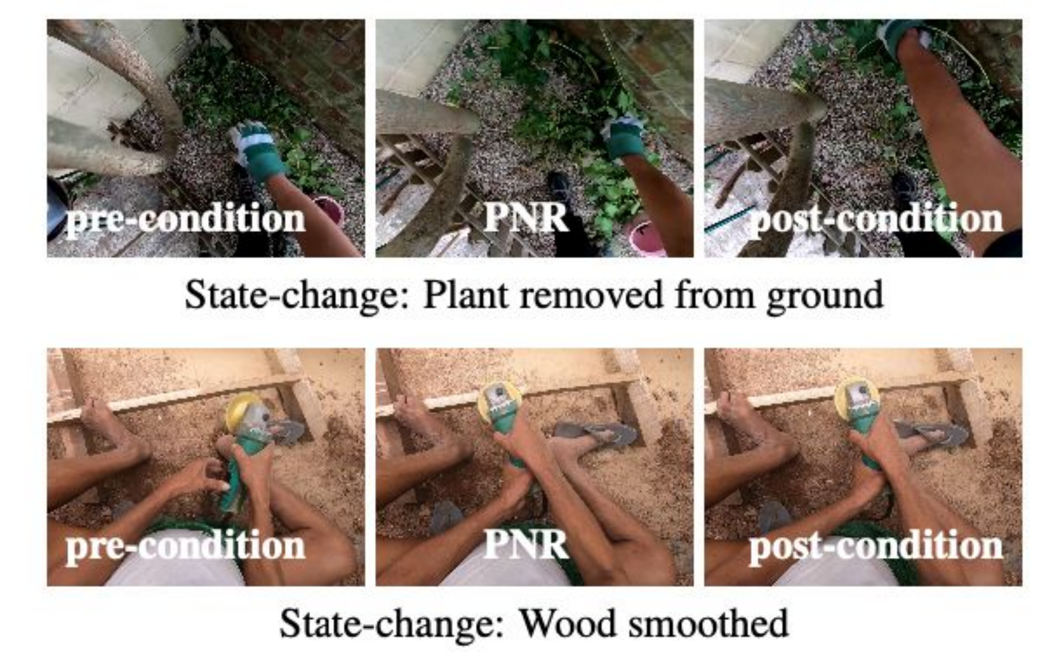
- Evaluation
  - Visual Query
    - temporal and spatio-temporal localization metrics as well as timeliness metrics that encourage speedy searches



$$\mathrm{sEff} = 1 - \frac{n}{N}$$

# Benchmark Suite(Hands and Objects)

- Motivation
- Task definition
    - Point-of-no-return temporal localization
    - State change object detection
    - Object state change classification
- Annotation

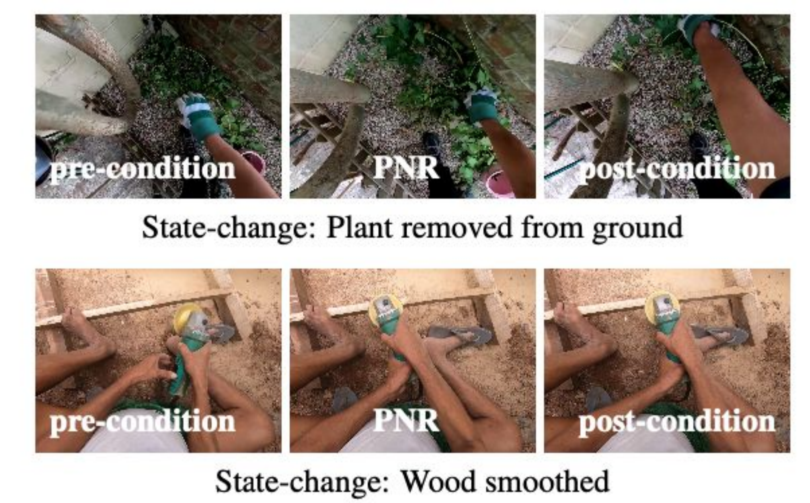


State-change: Plant removed from ground

State-change: Wood smoothed

# Benchmark Suite(Hands and Objects)

- Evaluation
    - Point-of-no-return temporal localization
        - Absolute temporal error (s)
    - State change object detection
        - AP
    - Object state change classification
        - classification accuracy



State-change: Plant removed from ground

State-change: Wood smoothed

# Benchmark Suite(Audio-Visual Diarization)

- Motivation
- Task definition
  - Localization and tracking
  - Active speaker detection
  - Diarization
  - Transcription
- Annotation

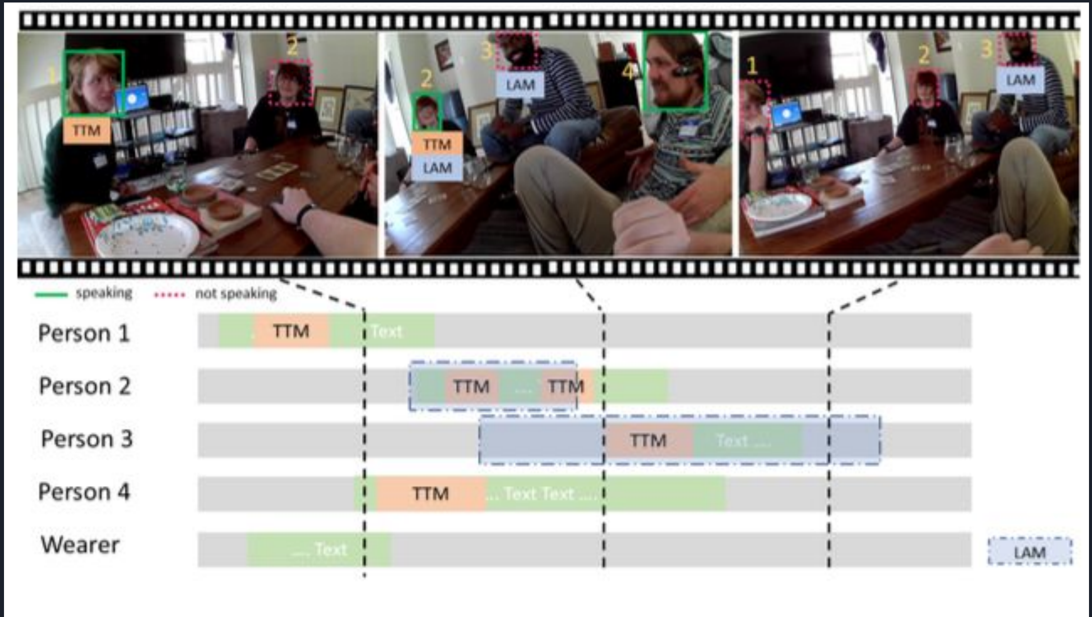


Figure 9. Audio-Visual and Social benchmark annotations

# Benchmark Suite(Audio-Visual Diarization)

- Evaluation
  - Localization and tracking
    - MOTA
    - MOTP

$$MOTA = 1 - \frac{\sum_t FN_t + FP_t + IDS_t}{\sum_t GT_t}$$

$$\mathrm{MOTP} = \frac{\sum_{i,t} d_t^i}{\sum_t c_t}.$$

# Benchmark Suite(Audio-Visual Diarization)

- Evaluation
  - Localization and tracking
    - MOT metrics
  - Active speaker detection
    - mAP
  - Diarization

$$\text{DER }(\%) = (E_{miss} + E_{fa} + E_{spk}) \times 100,$$

  - Transcription

$$\text{WER }(\%) = \frac{S + D + I}{N_w} \times 100.$$

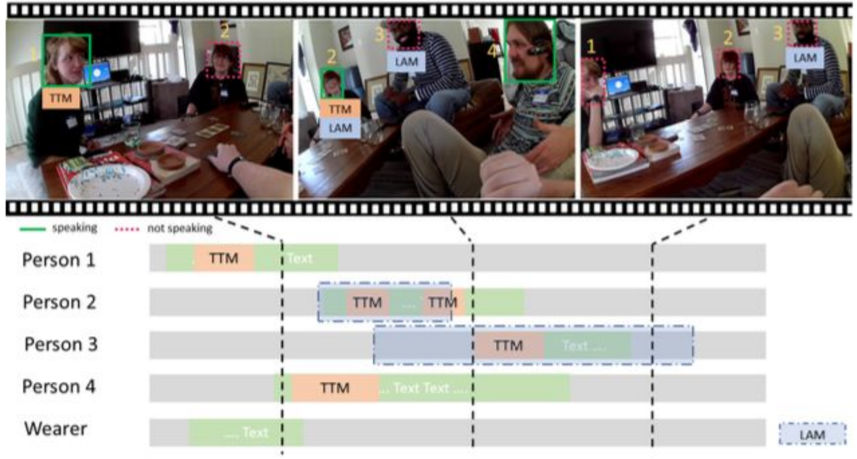


Figure 9. Audio-Visual and Social benchmark annotations

# Benchmark Suite(Social Interactions)

- Motivation
- Task definition
    - LAM
    - TTM
- Annotation
- Evaluation
    - mAP

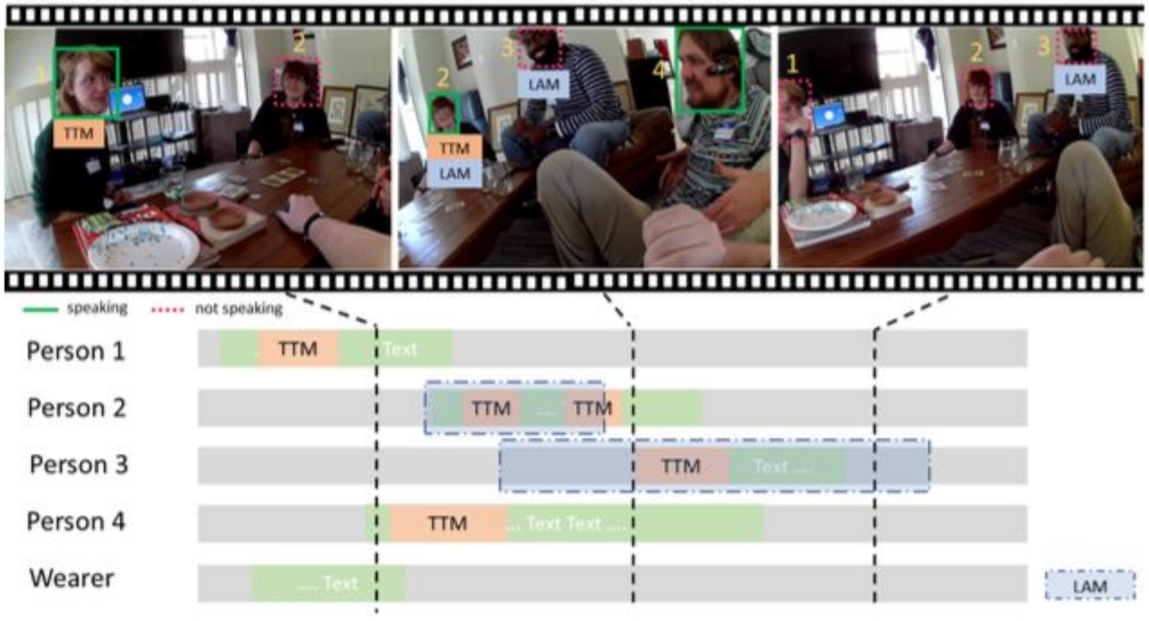


Figure 9. Audio-Visual and Social benchmark annotations

# Benchmark Suite(Forecasting)

- Motivation
- Task definition
  - Locomotion Movements
  - Hands Movements
  - {Short, Long} -Term Anticipation
- Annotation

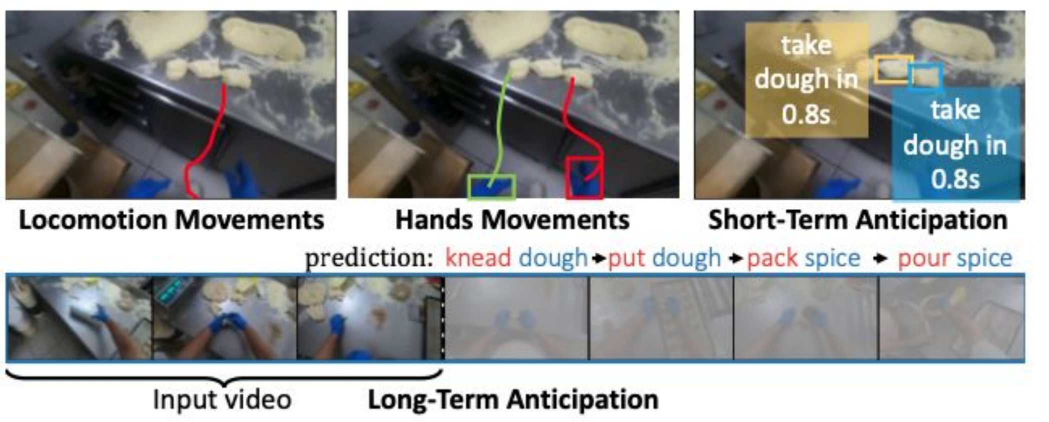


Figure 10. The Forecasting benchmark aims to predict future locomotion, movement of hands, next object interactions, and sequences of future actions.

# Benchmark Suite(Forecasting)

- Evaluation
    - Locomotion Movements

$$\mathrm{K-MTE} = \operatorname*{argmin}_{\{\mathcal{X}_k\}_{k=1}^{K}} \frac{1}{\sum_t v_t} \sum_t v_t \|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|,$$

$$\mathrm{PCT}\epsilon = \frac{1}{K}\delta\left(\frac{1}{\sum_t v_t} \sum_t v_t \|\mathbf{x}_t - \widehat{\mathbf{x}}_t\| < \epsilon\right)$$

- 
    - Hands Movements

$$D_m = \frac{1}{n} \sum_{i \in H_t} \|h_i - \hat{h}_i\|$$

$$D_c = \|h_c - \hat{h}_c\|$$

- 
    - Short-Term Anticipation
        - mAP
    - Long-Term Anticipation
        - Edit Distance