Stanford University

NEURAL INFORMATION PROCESSING SYSTEMS

# HourVideo

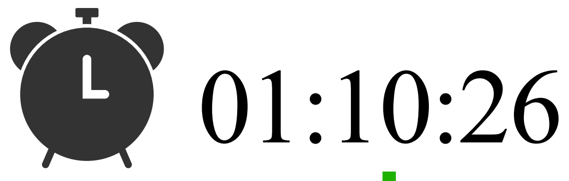
## 1-Hour Video-Language Understanding

*NeurIPS 2024 Datasets and Benchmarks*

hourvideo.stanford.edu

**Keshigeyan Chandrasegaran**, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He,
Cristobal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, Li Fei-Fei

# Long-form Video-Language Understanding


01:10:26

🔍 How can I get to the storage room from the living room?

🔍 Where did I leave my AirPods after working out?

Lack of holistic methods for evaluating hour-long video understanding capabilities.

# HourVideo at a Glance

500 egocentric videos
(381 hours)

77 everyday scenarios

- Summarization
- Perception
- Visual Reasoning
- Navigation

18 total tasks

12,976 high quality MCQs

# [HourVideo Design] Proposed Task Suite

## Summarization

- Key Events/ Objects Identification
- Temporal Sequencing
- Compare / Contrast

## Perception

- Information Retrieval
  - Factual Recall
  - Sequence Recall
  - Temporal Distance
- Tracking

## Navigation

- Room-to-Room
- Object Retrieval

## Visual Reasoning

### Spatial

- Relationship
- Proximity
- Layout

### Temporal

- Duration
- Frequency
- Pre-requisites

### Predictive

### Causal

### Counterfactual

# [HourVideo Design] Proposed Task Suite

**Summarization**
- Key Events/ Objects Identification
- Temporal Sequencing
- Compare / Contrast

**Perception**
- Information Retrieval
  - Factual Recall
  - Sequence Recall
  - Temporal Distance
- Tracking

**Navigation**
- Room-to-Room
- Object Retrieval

We manually design question prototypes for each task/ sub-task.

**Visual Reasoning**

**Spatial**
- Relationship
- Proximity
- Layout

**Temporal**
- Duration
- Frequency
- Pre-requisites

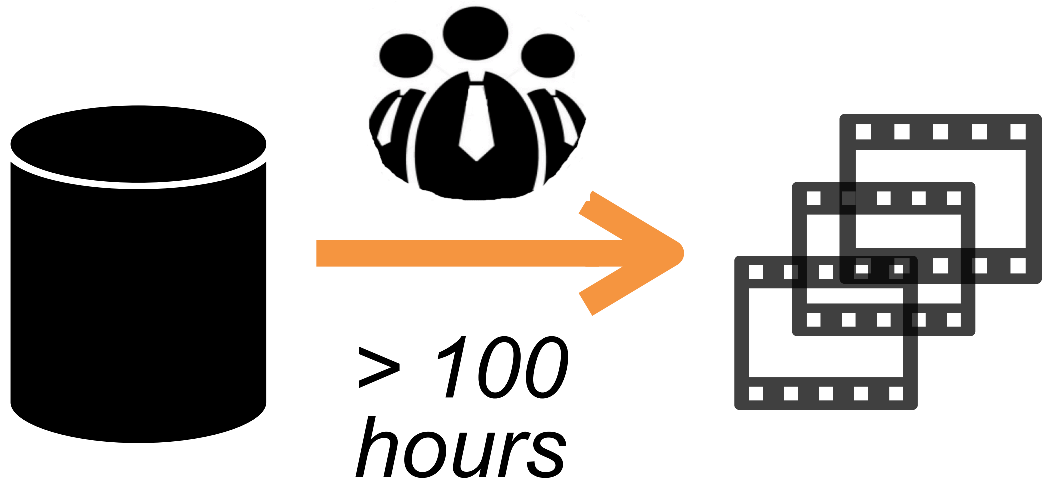
**Predictive**

**Causal**

**Counterfactual**

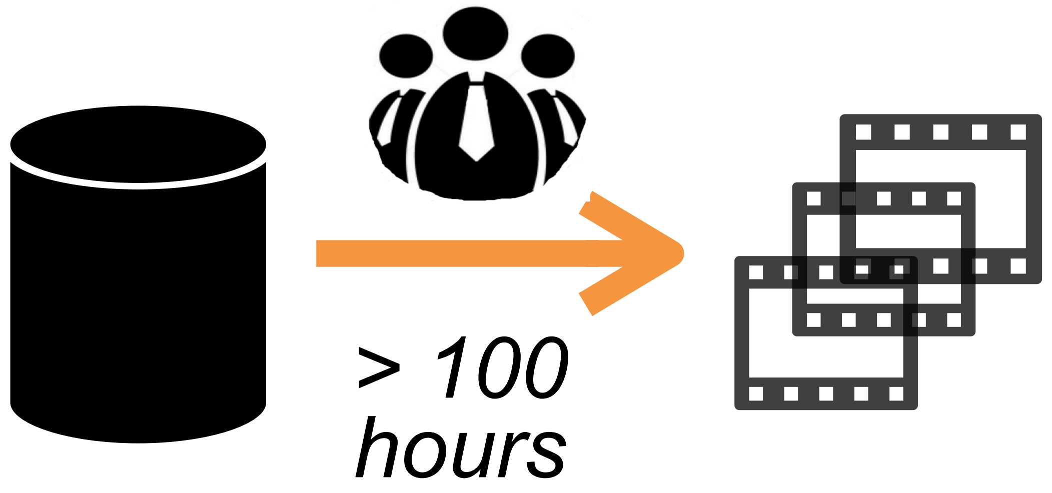
# [Data Generation Pipeline] Video Curation

## 1 Video Curation

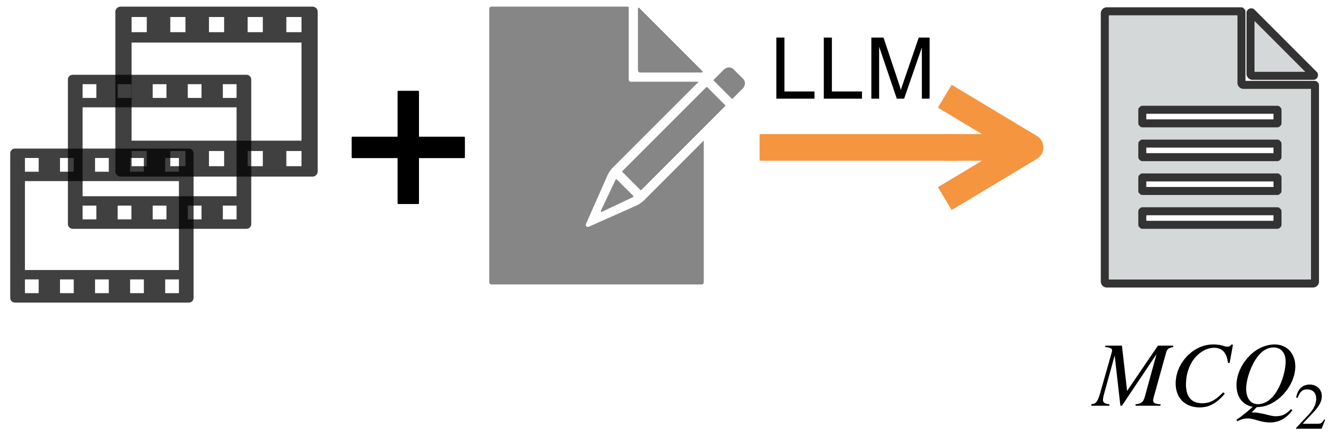


# [Data Generation Pipeline] MCQ Generation

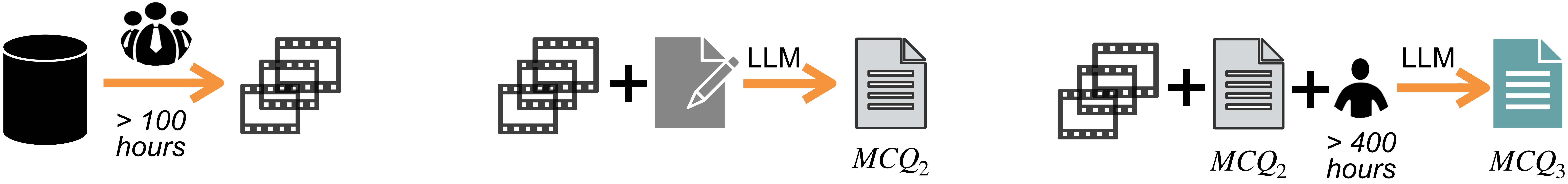
## 1 Video Curation



## 2 MCQ Generation



# [Data Generation Pipeline] MCQ Refinement using Human Feedback



1 Video Curation

> 100 hours

2 MCQ Generation

LLM

$MCQ_2$

3 MCQ Refinement using Human Feedback

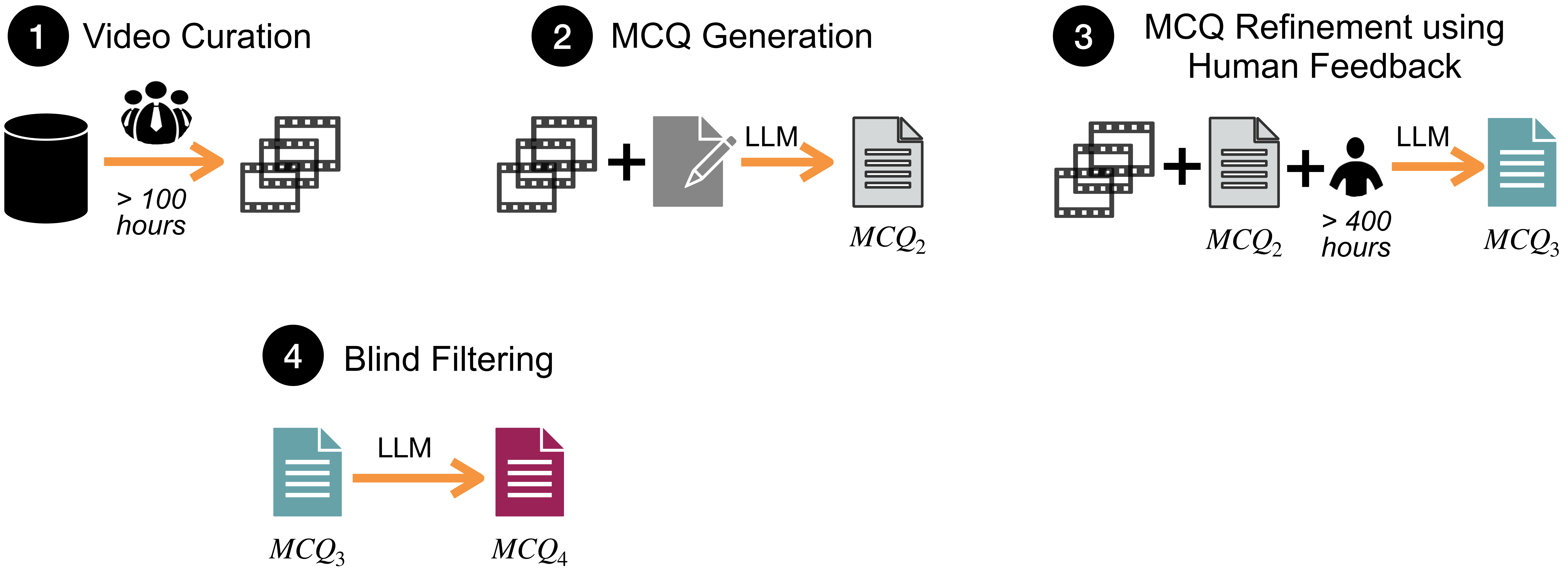
$MCQ_2$

> 400 hours

LLM

$MCQ_3$

# [Data Generation Pipeline] Blind Filtering



1 Video Curation

> 100 hours

2 MCQ Generation

LLM

$MCQ_2$

3 MCQ Refinement using Human Feedback

LLM

$MCQ_2$

> 400 hours

$MCQ_3$

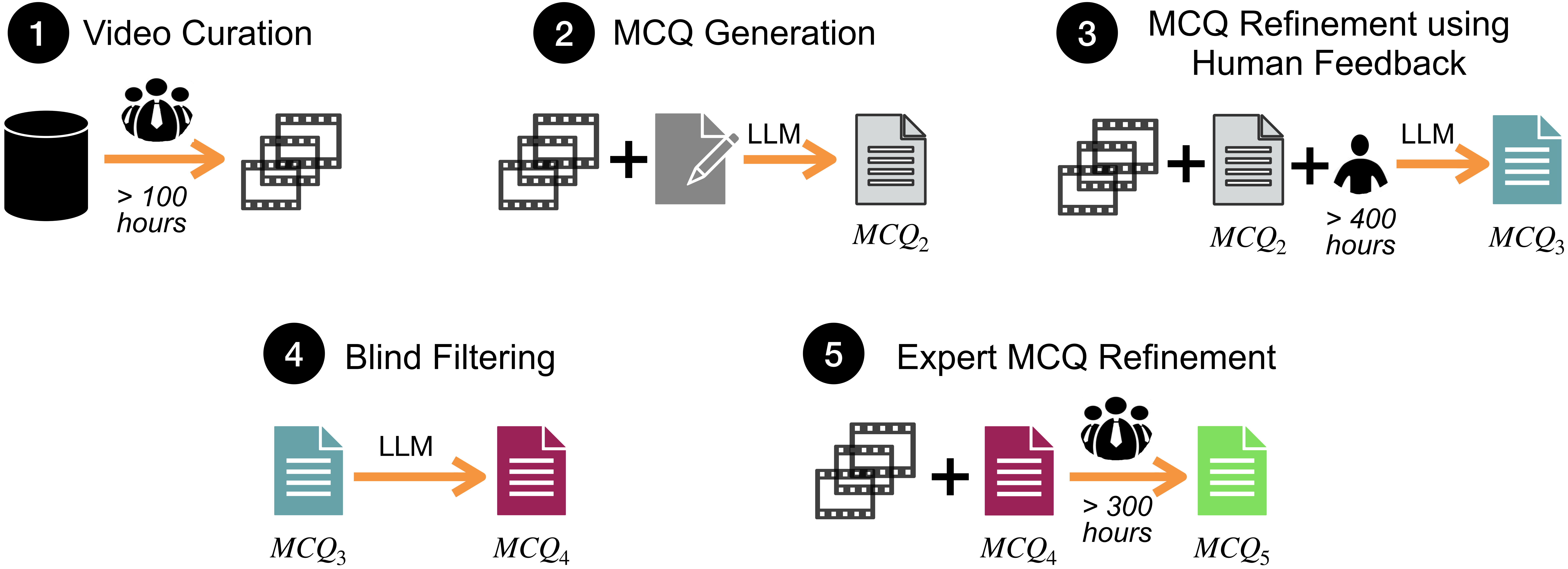
4 Blind Filtering

LLM

$MCQ_3$

$MCQ_4$

# [Data Generation Pipeline] Expert MCQ Refinement



1 Video Curation

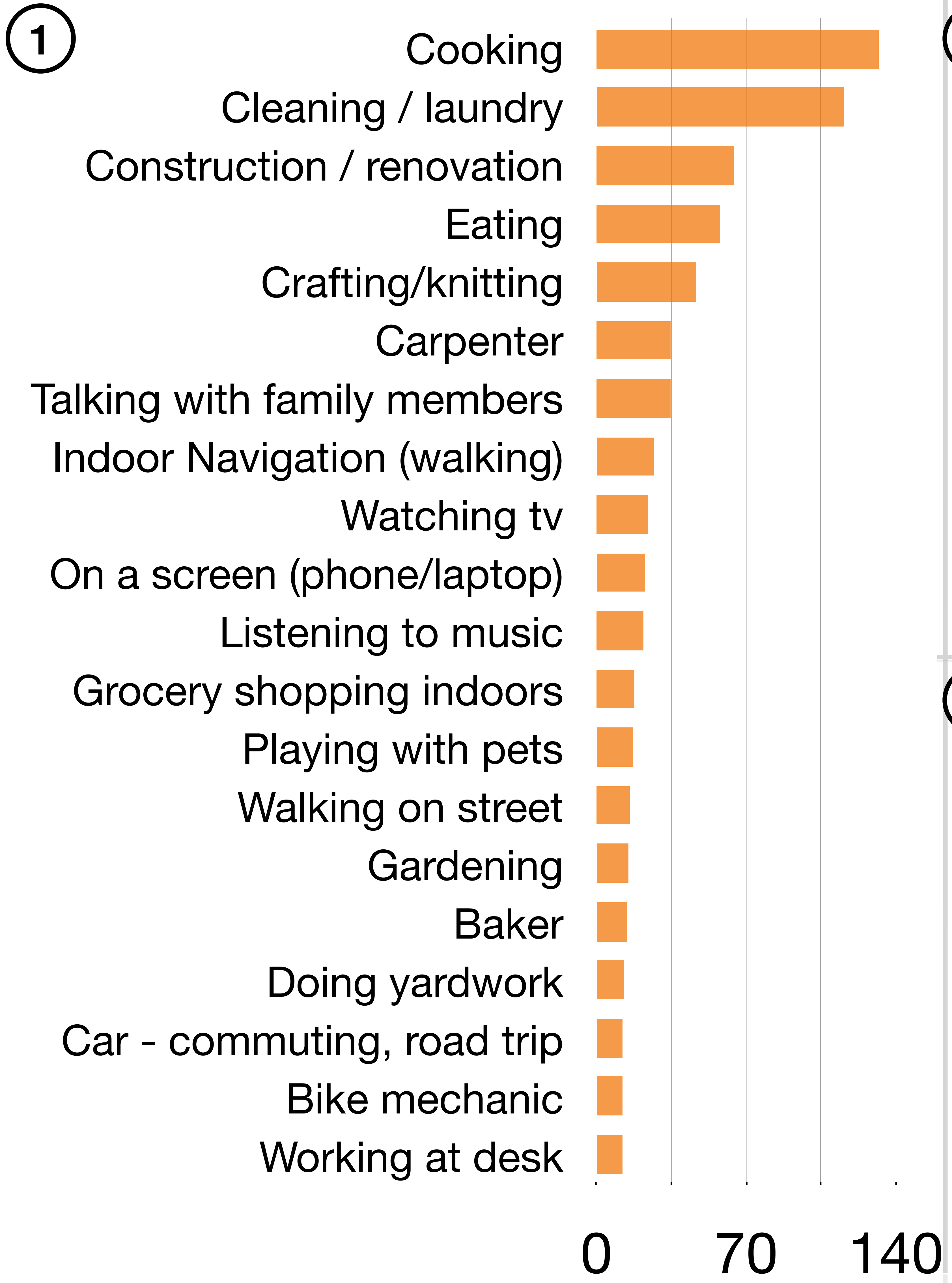
2 MCQ Generation

3 MCQ Refinement using Human Feedback

4 Blind Filtering

5 Expert MCQ Refinement

# HourVideo Statistics



2

| Summarization (714) | |
|---|---|
| Key Events/ Objects Identification | 467 |
| Temporal Sequencing | 152 |
| Compare/Contrast | 95 |

| Perception (3777) | |
|---|---|
| Factual Recall | 2479 |
| Sequence Recall | 854 |
| Temporal Distance | 267 |
| Tracking | 177 |

| Navigation (312) | |
|---|---|
| Room-to-Room | 120 |
| Object Retrieval | 192 |

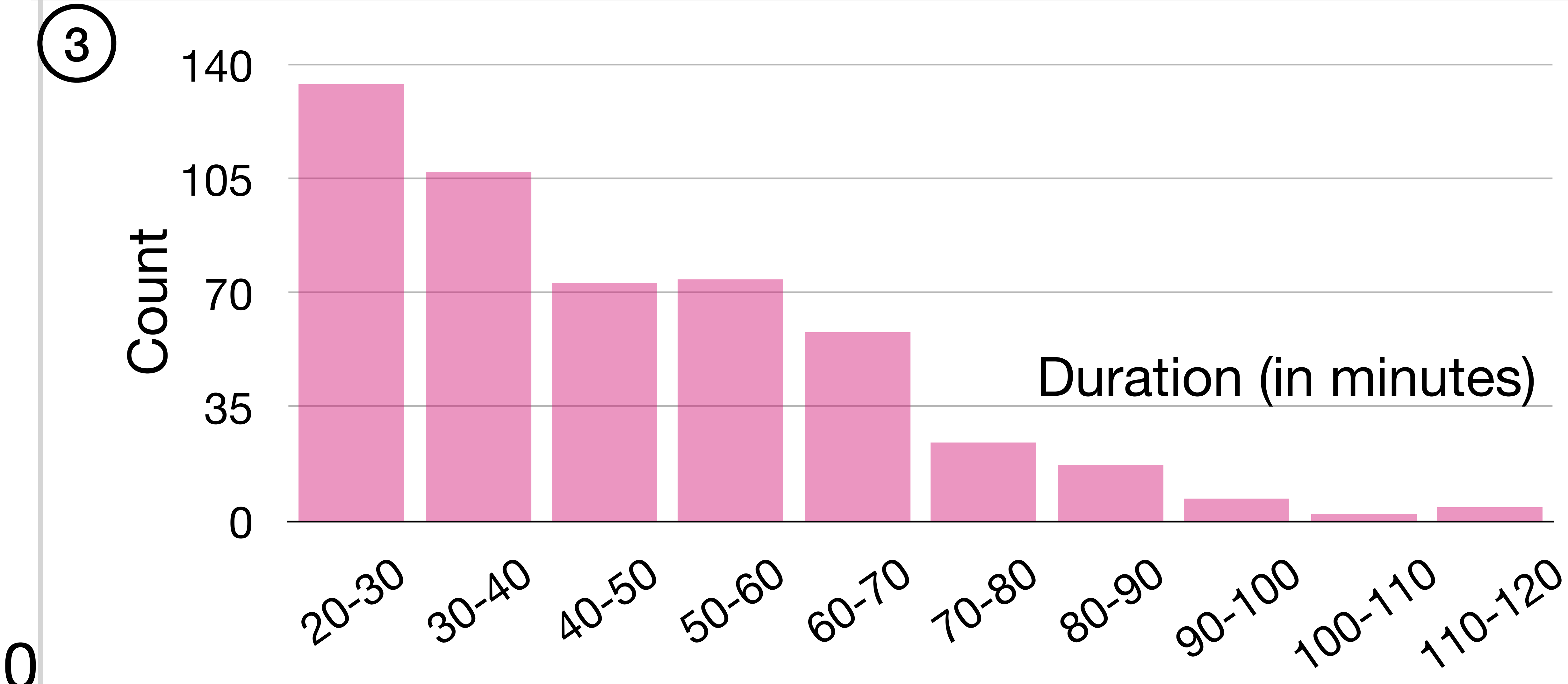
| Spatial (3173) | |
|---|---|
| Relationship | 1889 |
| Proximity | 1239 |
| Layout | 45 |

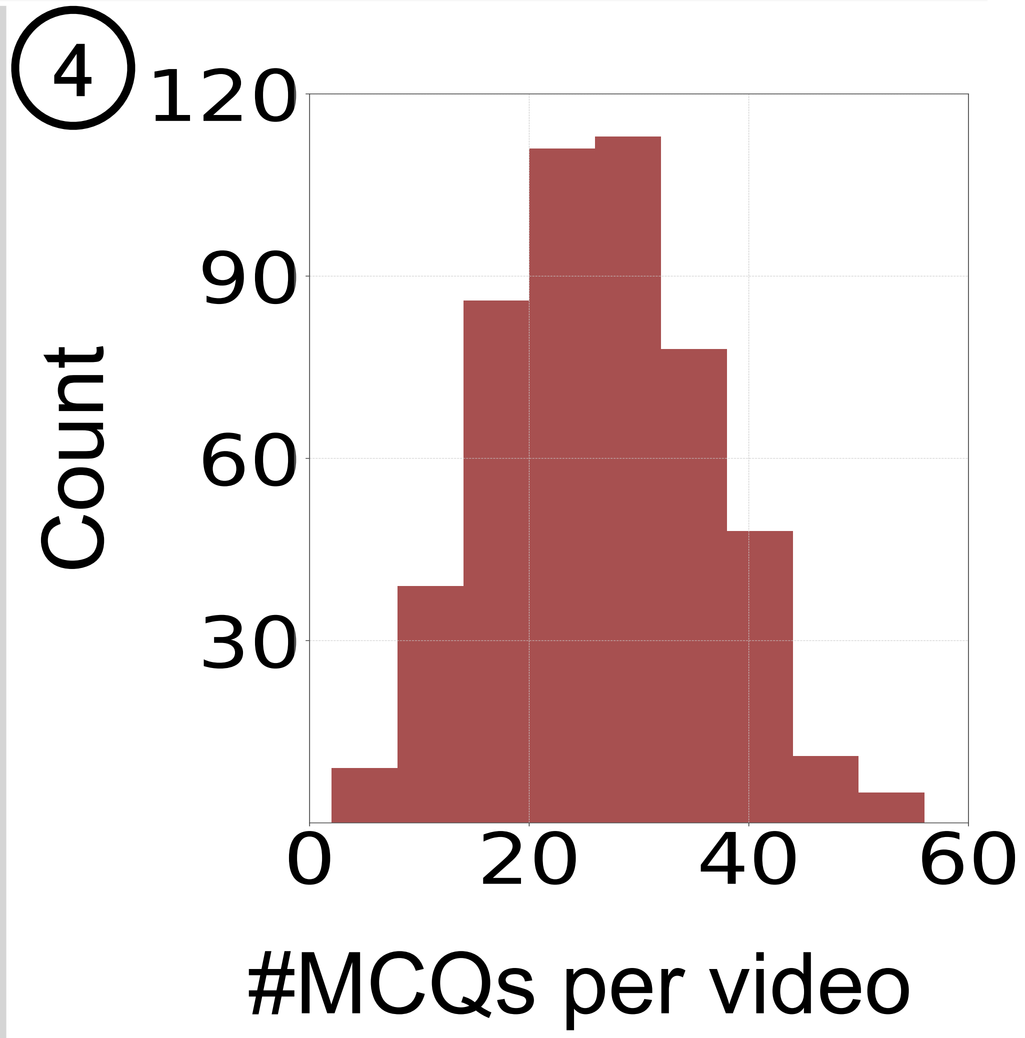
| Temporal (4292) | |
|---|---|
| Duration | 1945 |
| Frequency | 1815 |
| Pre-requisites | 532 |

Predictive (407)

Causal (150)

Counterfactual (151)





# Let’s look at two examples

## Summarization

- Key Events/ Objects Identification
- Temporal Sequencing
- Compare / Contrast

## Perception

- Information Retrieval
  - Factual Recall
  - Sequence Recall
  - Temporal Distance
- Tracking

## Navigation

- Room-to-Room
- Object Retrieval

## Visual Reasoning

### Spatial

- Relationship
- Proximity
- Layout

### Temporal

- Duration
- Frequency
- Pre-requisites

### Predictive

### Causal

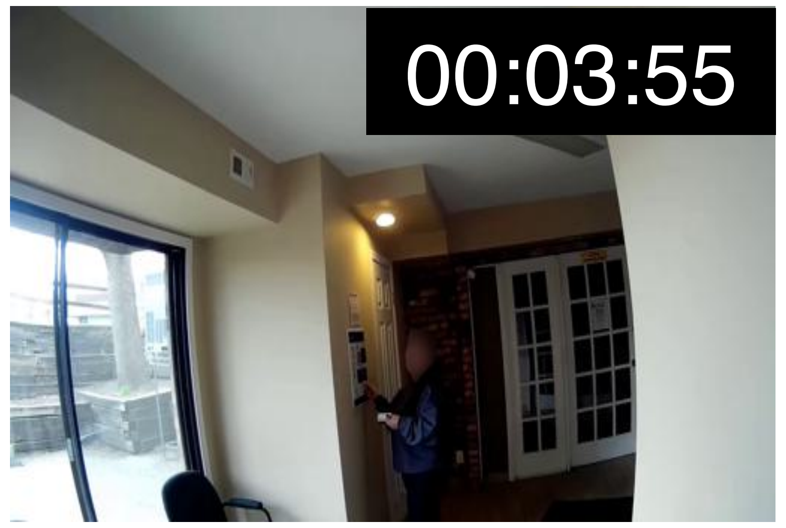
### Counterfactual

# Perception / Tracking

00:30:00

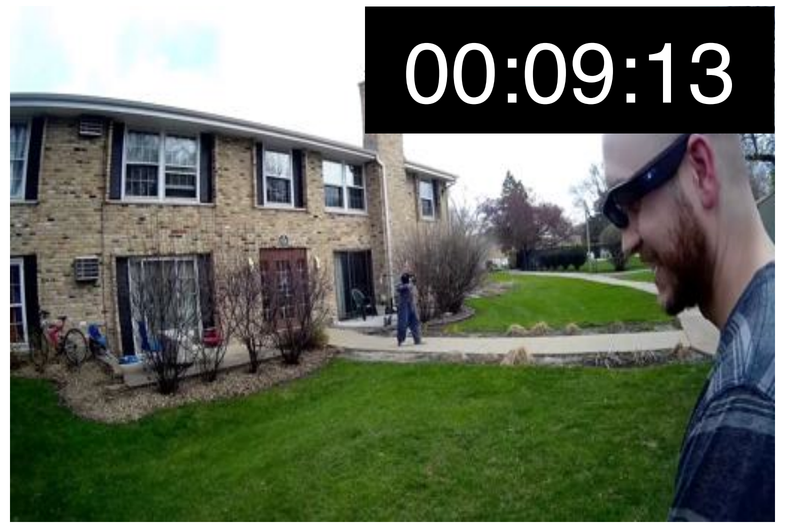
Identify the unique individuals the camera wearer interacted with.

00:00:00



00:03:55

00:04:04



00:09:13

00:18:23

00:30:00

MCQ Test

□ 2 Adults

□ 1 Adult

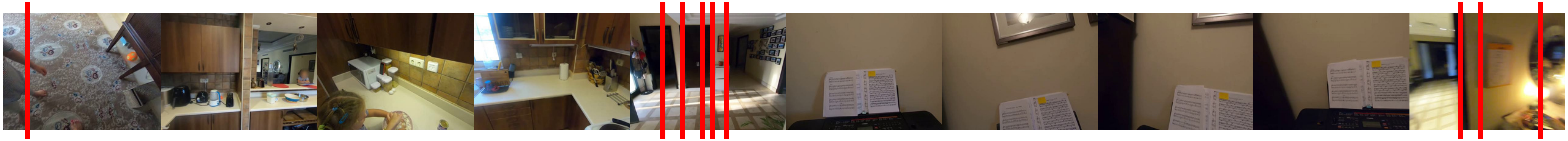
□ 4 Adults

□ 5 Adults

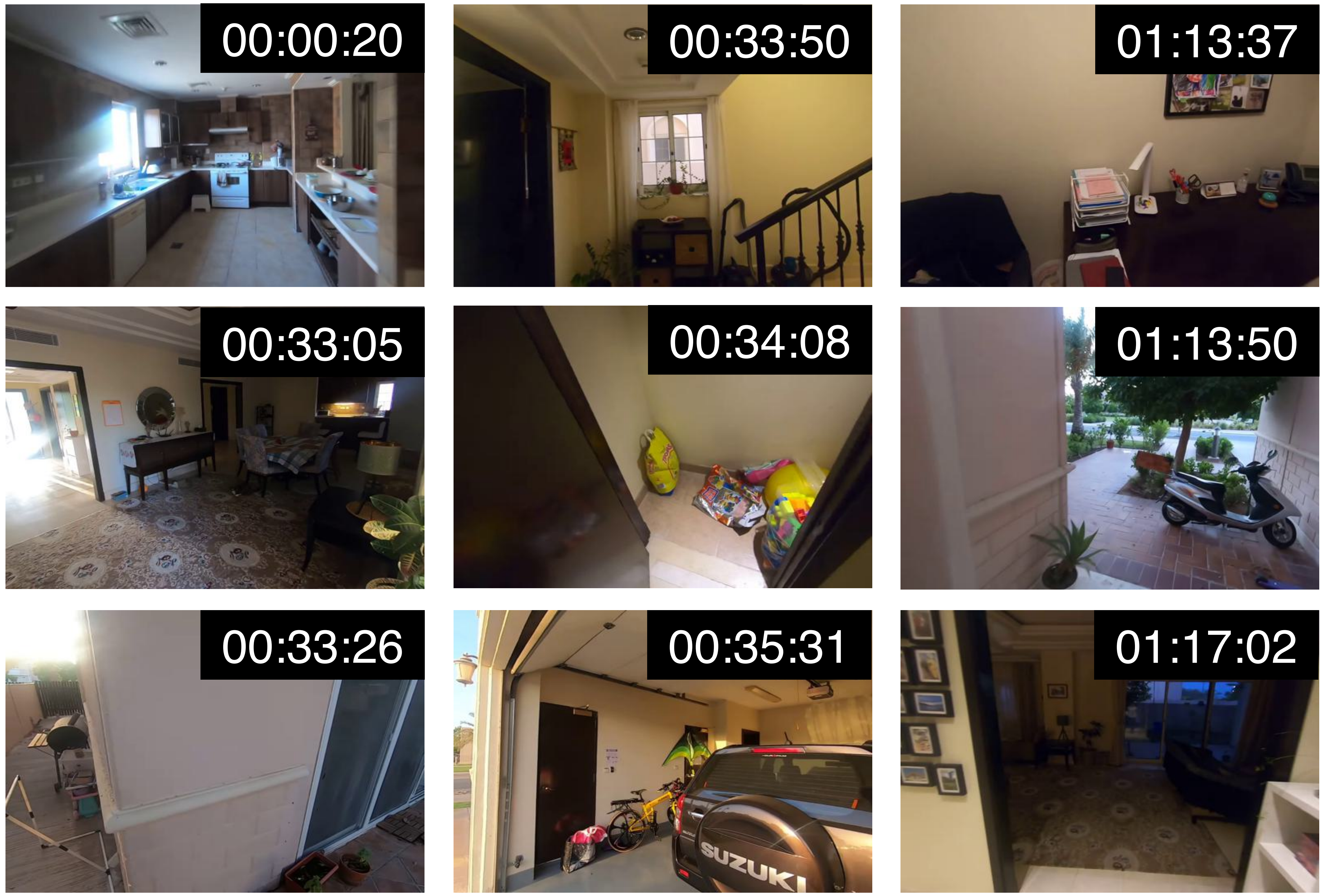
☑ 3 Adults

# Navigation/ Room-to-Room

01:17:11

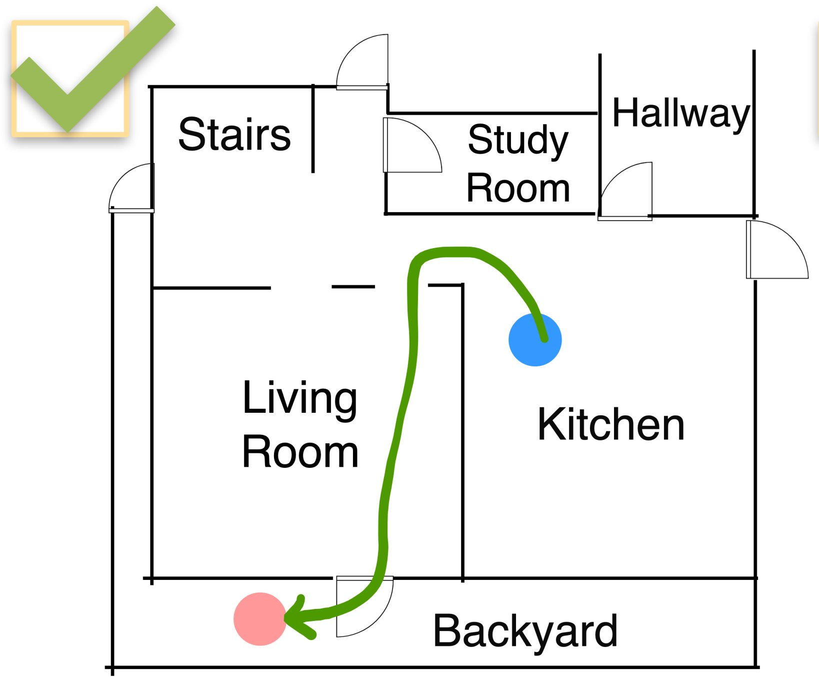


How can the camera wearer get to the backyard from the kitchen?

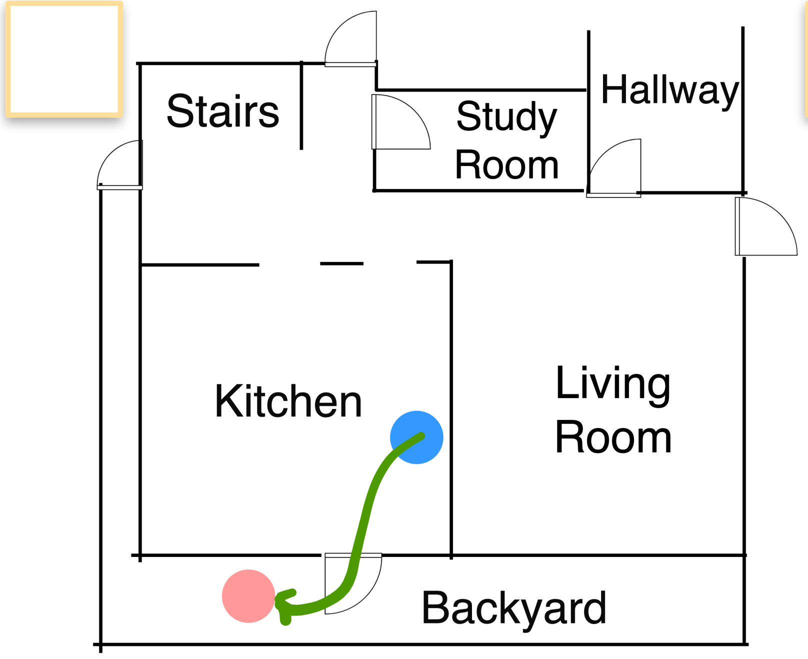


MCQ Test

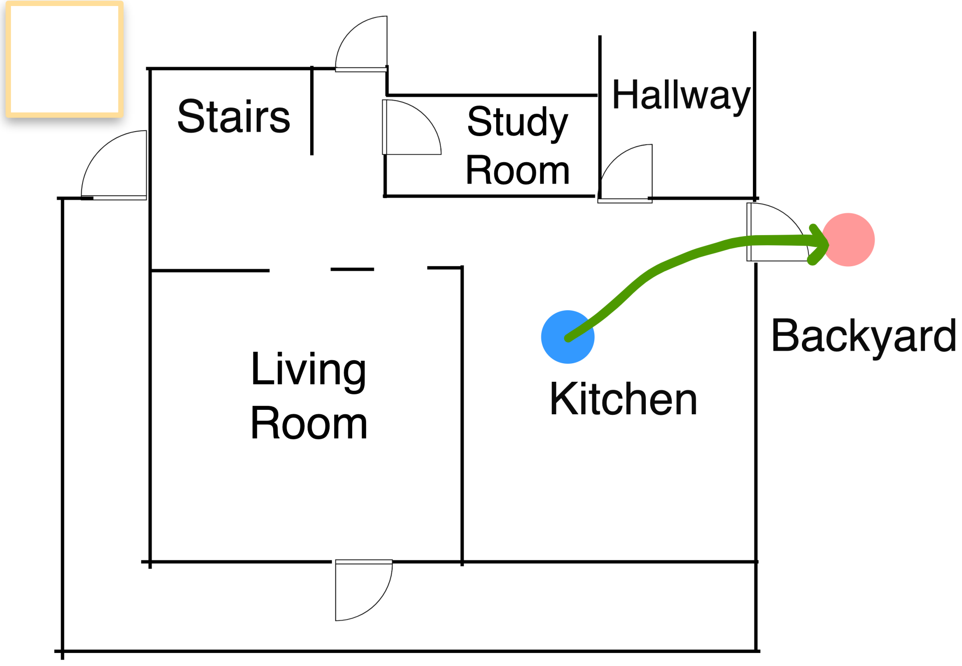
☑



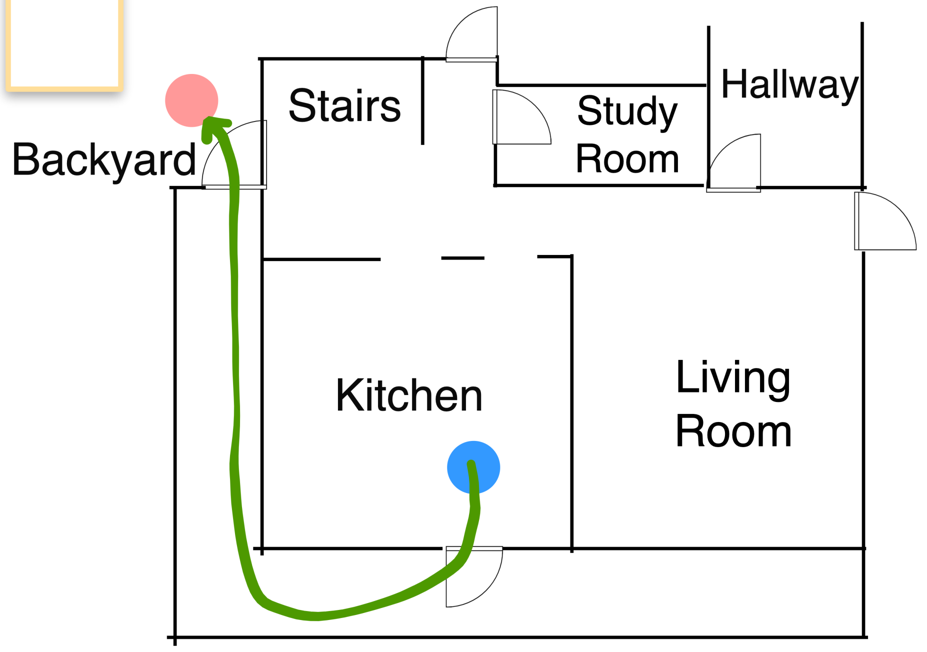
□



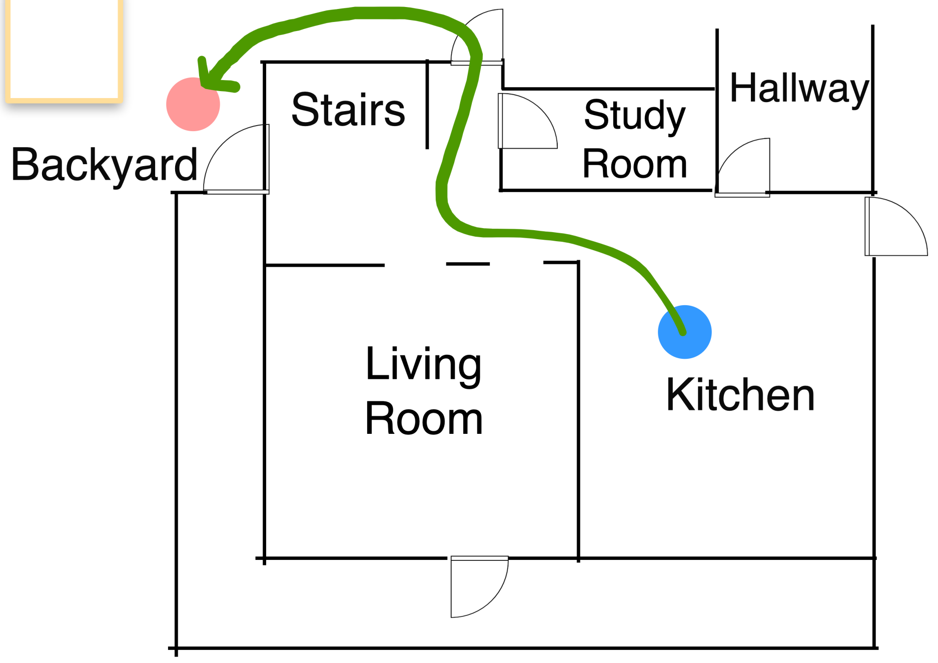
□



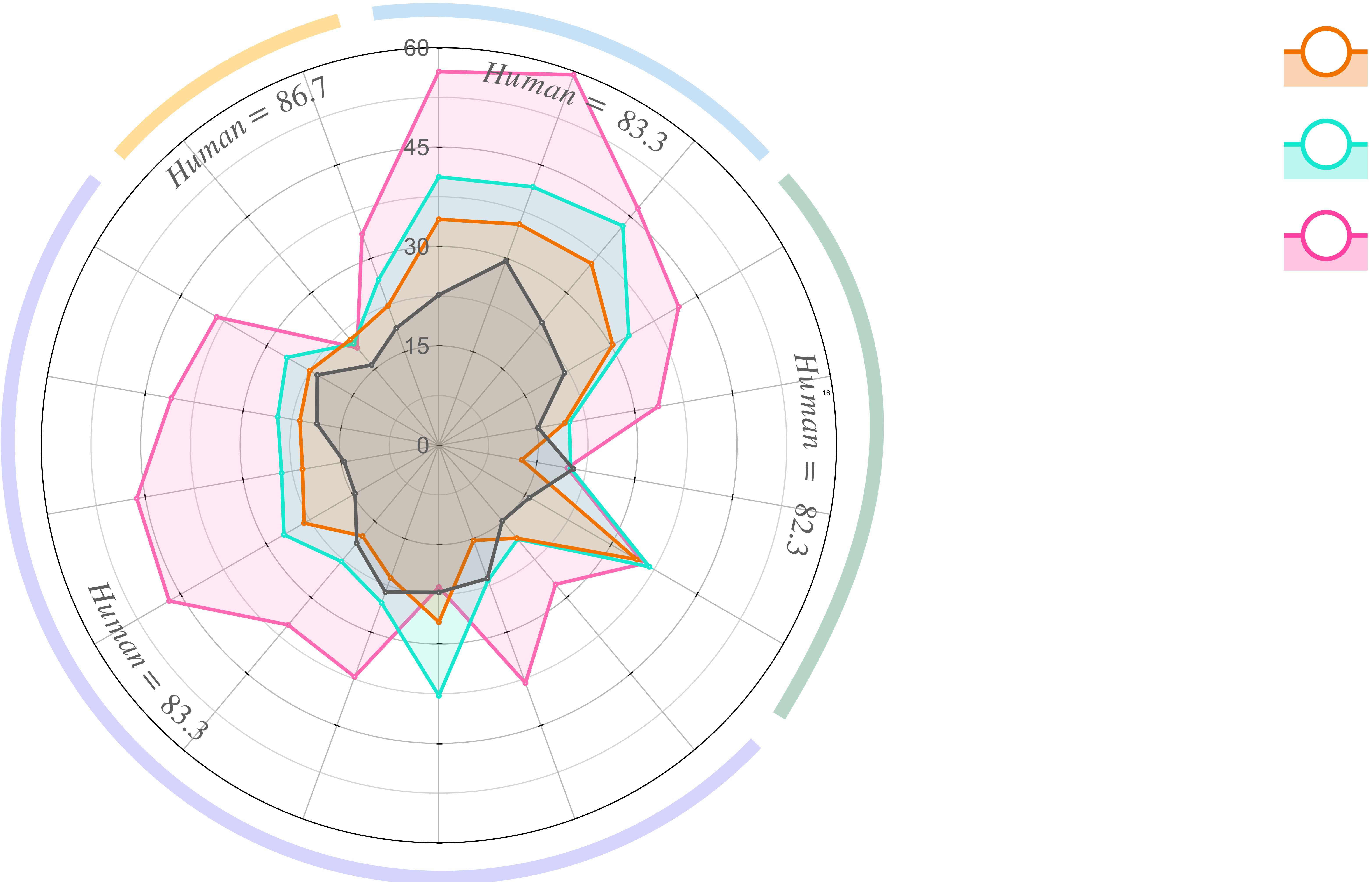
□



□



# How do today’s multimodal models perform on the HourVideo Benchmark?



# Key Takeaways

500 egocentric videos (381 hours)

77 everyday scenarios

Summarization

Perception

Visual Reasoning

Navigation

18 total tasks

12,976 high quality MCQs

# Key Takeaways

Summarization

Perception

Visual Reasoning

Navigation

500 egocentric videos (381 hours)

77 everyday scenarios

18 total tasks

12,976 high quality MCQs

We introduce **HourVideo**, a benchmark dataset designed to rigorously evaluate the capabilities of multimodal models to comprehend hour-long videos.

# Key Takeaways

500 egocentric videos (381 hours)

77 everyday scenarios

Summarization

Perception

Visual Reasoning

Navigation

18 total tasks

12,976 high quality MCQs

We introduce **HourVideo**, a benchmark dataset designed to rigorously evaluate the capabilities of multimodal models to comprehend hour-long videos.

We show that **a significant gap** exists between human experts and SOTA multimodal foundation models in comprehending long-form videos.

# HourVideo: 1-Hour Video-Language Understanding

Paper/ Benchmark/ Code/ Demos

hourvideo.stanford.edu