HourVideo: 1-Hour Video-Language Understanding

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He,
Cristobal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, Li Fei-Fei
Stanford University
hourvideo.stanford.edu

Dataset/Code/Demo

Motivation

- Humans excel at comprehending visual stimuli over **long time horizons**, enabling them to **perceive, plan, and act.**
- Similarly, general-purpose AI systems require sustained visual processing capabilities.
- **Research Gap:** Lack of systematic methods to **holistically evaluate** long-form video understanding.

Our Proposed Task Suite

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

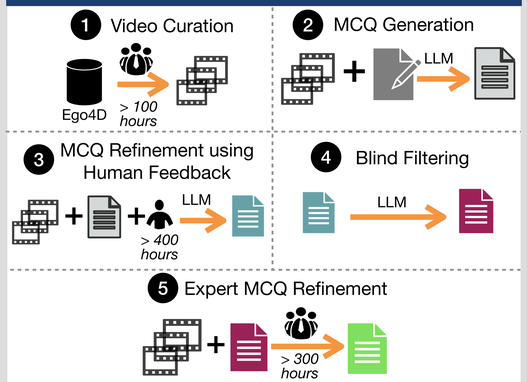
Perception
- Information Retrieval: Factual Recall, Sequence Recall, Temporal Distance
- Tracking

Navigation
- Room-to-Room
- Object Retrieval

Visual Reasoning
- Spatial: Relationship, Proximity, Layout
- Temporal: Duration, Frequency, Pre-requisites
- Predictive
- Causal
- Counterfactual

We design question prototypes for each task (see Table 1 in our paper).

HourVideo Generation Pipeline



A Benchmark Dataset for Hour-Long Video-Language Understanding

500 egocentric videos (381 hours)

77 everyday scenarios

Summarization
Perception
Visual Reasoning
Navigation
18 total tasks

12,976 high quality MCQs

Summarization (714)	
Key Events/ Objects Identification	467
Temporal	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

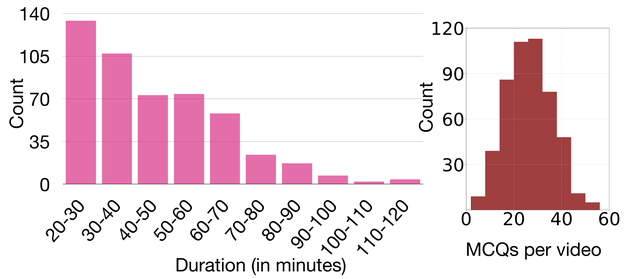
Spatial (3173)	
Relationship	1889
Proximity	1239
Layout	45

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

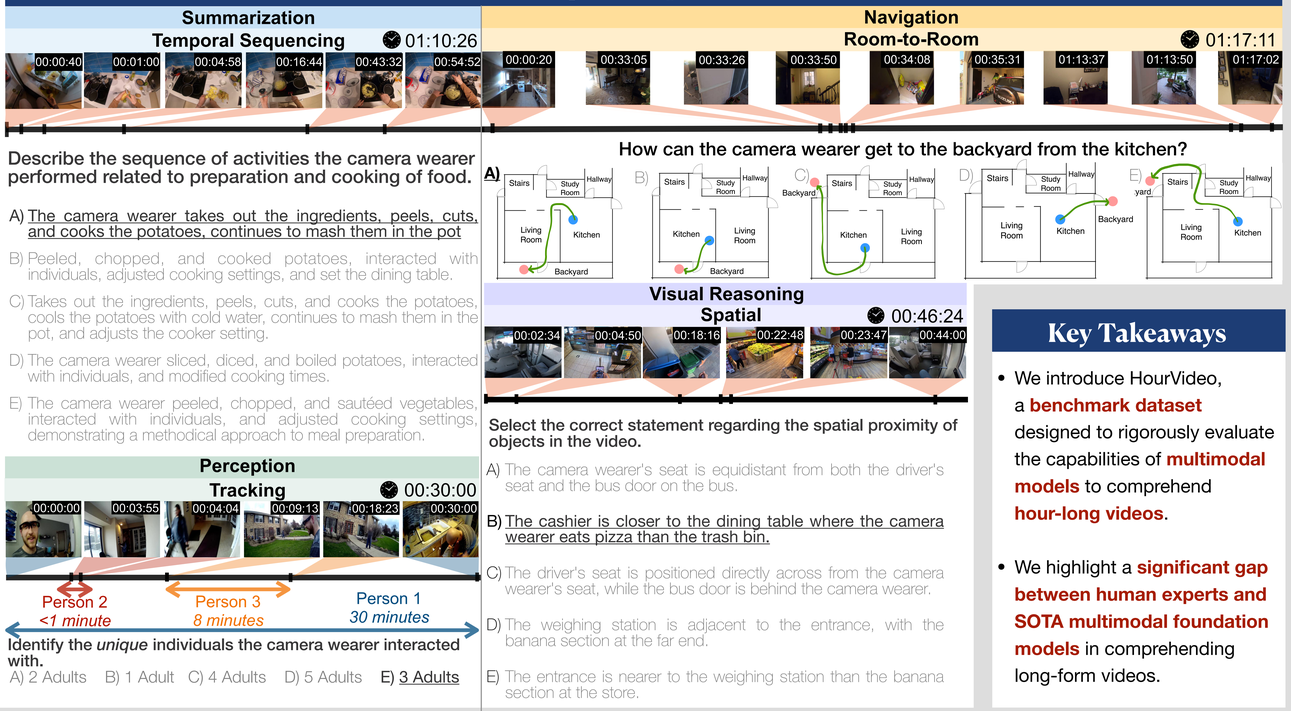
Predictive (407)

Causal (150)

Counterfactual (151)



Example MCQs from HourVideo

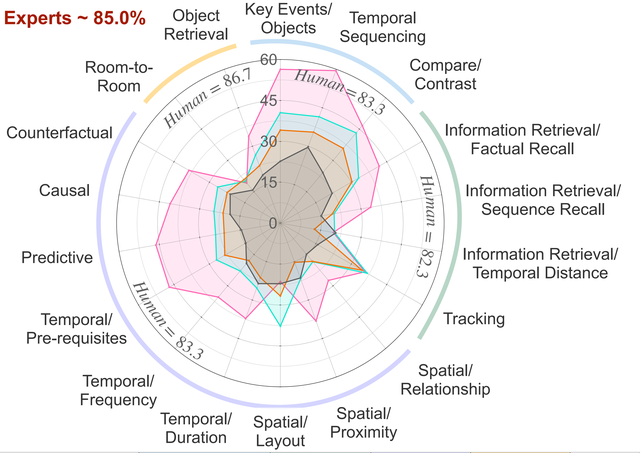


Key Takeaways

- We introduce HourVideo, a **benchmark dataset** designed to rigorously evaluate the capabilities of **multimodal models** to comprehend **hour-long videos**.
- We highlight a **significant gap between human experts and SOTA multimodal foundation models** in comprehending long-form videos.

How do today's Multimodal Models perform on HourVideo?

Human Experts ~ 85.0%



	Summarization	Perception	Visual Reasoning	Navigation	Avg.
Blind LLMs					
GPT-4	24.4	20.0	19.1	17.6	19.6
Socratic LLMs					
LLaVA-NeXT-34B	34.6	26.7	19.1	21.8	22.3
GPT-4	41.0	29.4	22.8	24.0	25.7
Native Multimodal LLMs					
Gemini 1.5 Pro	**55.8**	**38.2**	**35.7**	**28.1**	**37.3**