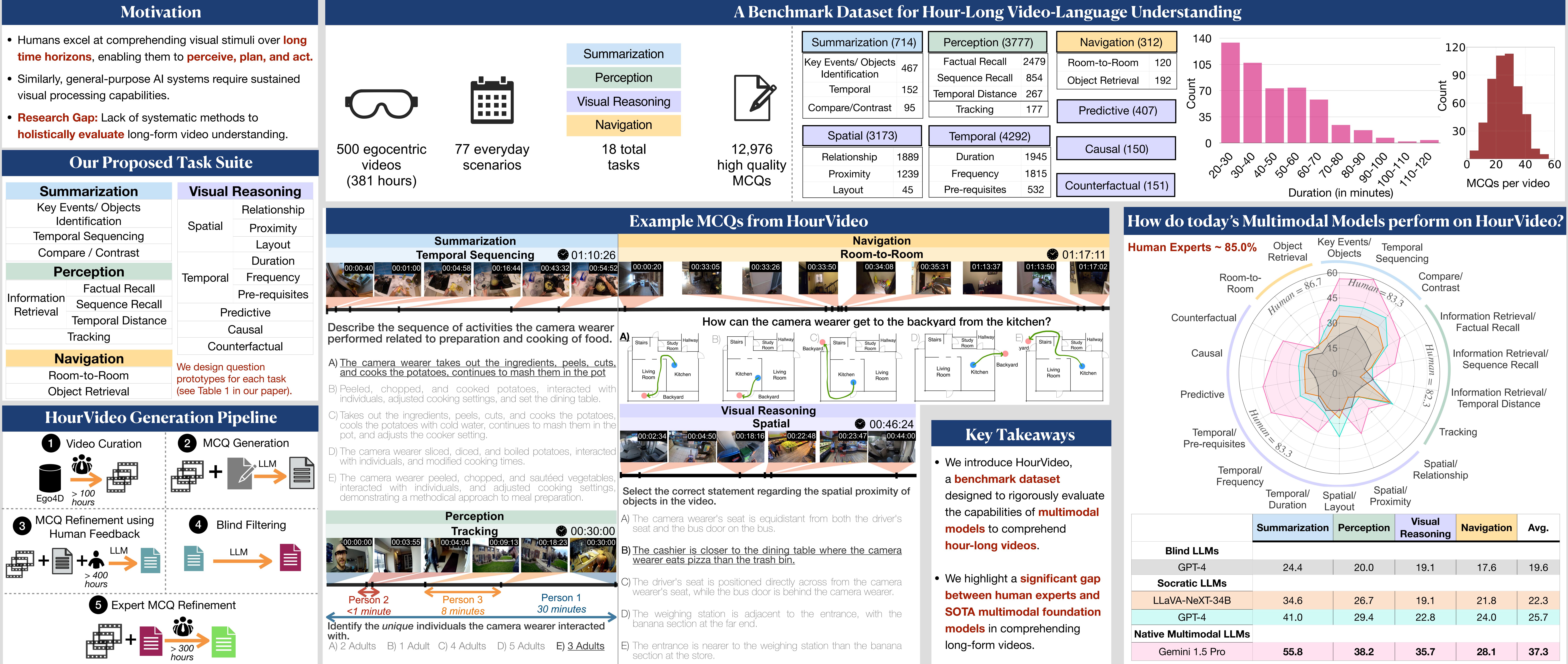# HourVideo: 1-Hour Video-Language Understanding

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

hourvideo.stanford.edu

Stanford University

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

① Video Curation — Ego4D, >100 hours
② MCQ Generation — +LLM
③ MCQ Refinement using Human Feedback — + + +LLM, >400 hours
④ Blind Filtering — LLM
⑤ Expert MCQ Refinement — + + >300 hours

A Benchmark Dataset for Hour-Long Video-Language Understanding

500 egocentric videos (381 hours) · 77 everyday scenarios · 18 total tasks · 12,976 high quality MCQs

Summarization · Perception · Visual Reasoning · Navigation

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

Spatial (3173)	
Relationship	1889
Proximity	1239
Layout	45

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

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

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

Predictive (407)

Causal (150)

Counterfactual (151)

Example MCQs from HourVideo

Summarization — Temporal Sequencing 🕐 01:10:26

Describe the sequence of activities the camera wearer performed related to preparation and cooking of food.

A) <u>The camera wearer takes out the ingredients, peels, cuts, and cooks the potatoes, continues to mash them in the pot</u>

B) Peeled, chopped, and cooked potatoes, interacted with individuals, adjusted cooking settings, and set the dining table.

C) Takes out the ingredients, peels, cuts, and cooks the potatoes, cools the potatoes with cold water, continues to mash them in the pot, and adjusts the cooker setting.

D) The camera wearer sliced, diced, and boiled potatoes, interacted with individuals, and modified cooking times.

E) The camera wearer peeled, chopped, and sautéed vegetables, interacted with individuals, and adjusted cooking settings, demonstrating a methodical approach to meal preparation.

Perception — Tracking 🕐 00:30:00

Person 2 <1 minute · Person 3 8 minutes · Person 1 30 minutes

Identify the *unique* individuals the camera wearer interacted with.
A) 2 Adults   B) 1 Adult   C) 4 Adults   D) 5 Adults   E) <u>3 Adults</u>

Navigation — Room-to-Room 🕐 01:17:11

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

Visual Reasoning — Spatial 🕐 00:46:24

Select the correct statement regarding the spatial proximity of objects in the video.

A) The camera wearer's seat is equidistant from both the driver's seat and the bus door on the bus.

B) <u>The cashier is closer to the dining table where the camera wearer eats pizza than the trash bin.</u>

C) The driver's seat is positioned directly across from the camera wearer's seat, while the bus door is behind the camera wearer.

D) The weighing station is adjacent to the entrance, with the banana section at the far end.

E) The entrance is nearer to the weighing station than the banana section at the store.

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**