
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

Abstract

We present **HourVideo**, a benchmark dataset for one-hour video-language understanding. Our dataset consists of a novel task suite comprising summarization, perception (*recall, tracking*), visual reasoning (*spatial, temporal, predictive, causal, counterfactual*), and navigation (*room-to-room, object retrieval*) tasks. HourVideo includes 500 manually curated egocentric videos from the Ego4D dataset, spanning durations of 20 to 120 minutes, and features **12,976 high-quality, five-way multiple-choice questions**. Benchmarking results reveal that multimodal models, including GPT-4V and LLaVA-NeXT, achieve marginal improvements over random chance. In stark contrast, human experts significantly outperform the state-of-the-art long-context multimodal model, Gemini Pro 1.5 (85.0% vs. 37.3%), highlighting a substantial gap in multimodal capabilities. Our benchmark, evaluation toolkit, prompts, and documentation are available at hourvideo.stanford.edu.

1 Introduction

Our world presents an endless stream of visual stimuli. Humans demonstrate a remarkable ability to process visual stimuli over long time horizons, enabling them to perceive, plan and act in the real world. Consider the routine task of cooking a meal. This activity involves a continuous and adaptive visual process: identifying and using ingredients and tools, monitoring state changes of various dishes, and adjusting cooking duration/techniques based on visual cues such as color and texture. Such sustained visual processing is crucial to achieving the desired culinary outcomes. Naturally, endowing this capability to autonomous agents has been a long-standing goal in Artificial Intelligence.

In recent years, large multimodal models [1–3] have emerged as a promising approach toward achieving this goal. Typically, these models are evaluated using multiple datasets that test capabilities such as object recognition [4, 5], image comprehension [6–8], and action recognition [9]. However, these benchmarks are often restricted to single images or short video clips, usually lasting from a few seconds to no more than three minutes [9–12]. While these benchmarks have spurred significant advancements, a deeper exploration into long-form video-language understanding is essential to develop multimodal systems that can form the basis for future autonomous agents and assistants.

A significant challenge in evaluating long-form video-language understanding capabilities is designing tasks that genuinely necessitate *long-term* comprehension, i.e., tasks that require long-range dependencies. Merely posing questions that can be answered by watching a brief segment of a lengthy video effectively reduces the task to a combination of temporal localization and short-clip understanding. Furthermore, while intriguing narrative inquiries can certainly be formulated for long-form videos such as television shows and films, it is imperative to ensure that the questions are not trivially answerable due to the vast prior knowledge encoded in modern large language models.

In this work, we introduce **HourVideo**—a benchmark dataset designed for long-form video-language understanding. To design tasks that require *long-term* comprehension, we first propose a novel task

Correspondence to {keshik, agrim}@stanford.edu

suite (Tab. 1), comprising **summarization**, **perception** (*recall, tracking*), **visual reasoning** (*spatial, temporal, predictive, causal, counterfactual*), and **navigation** (*room-to-room, object retrieval*) tasks. For each task, we manually create question prototypes designed to ensure that correctly answering them requires identification and synthesis of information across multiple temporal segments within the long-form videos. Guided by our task suite, we curated 500 egocentric videos from the Ego4D dataset [13]—covering 77 unique everyday activities and ranging from 20 to 120 minutes—to generate questions based on our prototypes. The combination of our comprehensive task suite and everyday mundane egocentric videos provides a robust framework to rigorously evaluate multimodal models’ capabilities in understanding long-form videos. Finally, we developed a question-answer generation pipeline utilizing the expertise of trained human annotators (800+ hours of effort) and large language models (LLMs), resulting in a collection of 12,976 high-quality, five-way multiple-choice questions.

We comprehensively evaluate state-of-the-art multimodal models on HourVideo (Tab. 3, Fig. 4), including GPT-4V [2], Gemini 1.5 Pro [3], and LLaVA-NeXT [14] in a zero-shot setting. Our findings reveal that GPT-4V and LLaVA-NeXT achieve only marginal improvements over a random predictor (20%), obtaining accuracies of 25.7% and 22.3%, respectively. Gemini 1.5 Pro, designed specifically for long-context multimodal understanding, obtains an accuracy of 37.3%, which, while better, is still substantially lower than the average performance of human experts at 85.0%. These results suggest that while the multimodal community has made meaningful progress, a significant gap remains to be bridged before these systems can match human-level long-form video understanding capabilities. We hope that HourVideo will serve as a benchmark to facilitate research in this direction and enable the development of multimodal models that can understand endless streams of visual data.

2 Benchmark Design and Construction

While open-ended question answering closely emulates human interaction, automating the evaluation of free-form natural language responses remains challenging. Given that our primary goal is to assess long-form video-language understanding capabilities, we opt for a five-way multiple-choice question-answering (MCQ) task. This approach simplifies the evaluation process by enabling the calculation of an aggregate question-answering accuracy metric. In the following section, we describe our task suite and question-answer generation pipeline in detail, both of which are designed to curate high-quality five-way multiple-choice questions (MCQs).

2.1 Task Suite

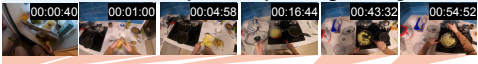
Creating a comprehensive benchmark for long-form video-language understanding is challenging, primarily because formulating meaningful questions that require processing and synthesizing information across various temporal segments is highly nontrivial, even for expert human annotators. Moreover, we note that even benchmarks for image or short video clip understanding are difficult to construct. As a result, we typically observe two common strategies for benchmark creation: (1) pre-defined label spaces testing for a specific skill or within narrow domains (e.g., Kinetics [9] and Something-Something [15]); or (2) gluing together different datasets, each designed to test a specific model capability [16–19]. In contrast, a single benchmark that can comprehensively test a suite of model capabilities can significantly benefit the research community.

We draw inspiration from both lines of research methodologies and introduce a novel suite of tasks designed to benchmark long-form video-language understanding capabilities for one-hour-long videos. Our task suite encompasses a comprehensive set of perceptual and cognitive tasks, including summarization, perception (recall, tracking), visual reasoning (spatial, temporal, predictive, causal, counterfactual), and navigation (room-to-room, object retrieval) tasks. Our strategy draws inspiration from the two common approaches previously discussed: (1) designing narrowly focused question prototypes to significantly streamline the question-answer creation process, and (2) creating a diverse suite of tasks that holistically evaluate a broad spectrum of multimodal capabilities. Our task suite with manually designed question prototypes are shown in Table 1. In particular, there are 18 sub-tasks in our proposed task suite and example MCQs from HourVideo are shown in Fig. 1.

Summarization

Temporal Sequencing

01:10:26



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

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

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.

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) The cashier is closer to the dining table where the camera wearer eats pizza than the trash bin.


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.

Predictive

00:46:24



After shopping and interacting with the Cashier at the checkout, what will the camera wearer do next?

A) Looks at their phone while pushing a trolley, exits the store, hands over the trolley to a woman, then cycle back home.

B) Looks at their phone while pushing a trolley, exits the store, hands over the trolley to Woman, buys a drink at the exit, then run towards the bus stand.

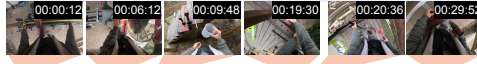
C) Looks at their phone while pushing a trolley, exits the store, hands over the trolley to a man, then walk towards the bus stop.

D) Looks at their phone while pushing a trolley, exits the store, hands over the trolley to Woman, then run towards the bus stand.

E) Looks at their phone while pushing a trolley, exits the store, buys a drink at the exit, then run towards the bus stand.

Temporal

00:32:30



Select the correct statement regarding frequencies of different tool usage in the video

A) During the construction activities, the miter saw was used more frequently compared to the cordless drill.

B) During the woodworking tasks, the tape measure was used more often compared to the ruler.

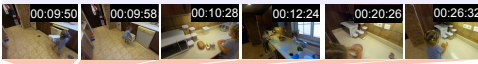
C) During the deck construction, the hammer was used more frequently compared to the impact driver.

D) During woodworking activity, the tape measure was used more frequently compared to the circular saw.

E) During the woodworking activity, the miter saw was used more frequently compared to the tape measure.

Causal

01:17:11



Why did the child move the step stool near the kitchen countertop?

A) The child moved the stepstool near the kitchen countertop to reach it and help with preparing dough and beating eggs.

B) The child moved the stepstool near the kitchen countertop to access it and retrieve the cookie jar.

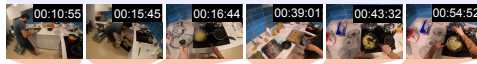
C) The child moved the stepstool near the kitchen countertop to reach the sink over it and wash her hands.

D) The child moved the step stool near the kitchen countertop to reach it and help with preparing dough.

E) The child moved the stepstool near the kitchen countertop to access the top drawer above it and find the measuring spoons.

Counterfactual

01:10:26



What if the camera wearer used the oven to make mashed potatoes?

A) Overall cooking time would have increased as the oven was also used by the camera wearer to bake cookies.

B) Overall cooking time would have increased as the oven would consume more time compared to using induction stove.

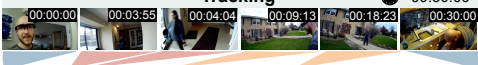
C) Overall cooking time would have increased as the oven was also used by the man to bake cookies.

D) Overall cooking time would have increased as the oven would consume more time compared to using gas cooker.

E) Overall cooking time would have increased as the oven would consume more time compared to using microwave.

Perception Tracking

00:30:00



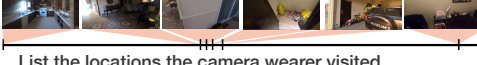
Person 2 <1 minute Person 3 8 minutes Person 1 30 minutes

Identify the *unique* individuals the camera wearer interacted with.

A) 2 Adults B) 1 Adult C) 4 Adults D) 5 Adults E) 3 Adults

Information Retrieval / Factual Recall

01:17:11



List the locations the camera wearer visited.

A) Kitchen, BBQ Area, Storage Room, Garage, Room, Pavements

B) Kitchen, Garden, Storage Room, Garage, Room, Pavements

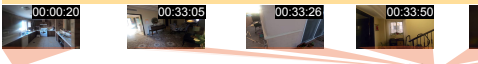
C) Kitchen, Bathroom, Storage Room, Garage, Room, Pavements,

D) Kitchen, Room, Balcony, Storage Room, Garage, Living Room

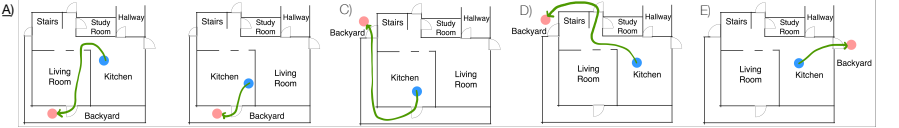
E) Kitchen, Balcony, Storage Room, Garage, Room, Pavements

Navigation

01:17:11



Room-to-Room Navigation: How can the camera wearer get to the backyard from the kitchen?



Object Retrieval: How can the camera wearer retrieve the motorcycle from the kitchen?

A) Exit the kitchen towards the stairs and exit through the door. The motorbike is outside.

B) Exit the kitchen through the door into the backyard, and the motorbike is on the right.

C) Exit the kitchen and turn left. Walk through the living room and go through the door into the backyard. The motorbike is on the right.

D) Exit the kitchen to the living room and turn left. Go through the door to the backyard; the motorbike is on the right.

E) Exit the kitchen and turn left. Walk down the hallway and turn right before the stairs. Exit the door and the motorbike is outside.

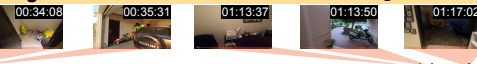


Figure 1: Example MCQs from HourVideo for different tasks. The correct answers are underlined.

Summarization	
Key Events/ Objects	<i>Summarize the key interactions of the camera wearer in the [supermarket].</i>
Temporal Sequencing	<i>Describe the sequence of activities performed by the camera wearer to [prepare the desert].</i>
Compare/ Contrast	<i>How did the camera wearer’s activities in the [apartment] differ from those in the [restaurant]?</i>
Perception	
Information Retrieval	
• Factual Recall	<i>What [dairy products] did the camera wearer [pick up] in the [supermarket]?</i>
• Sequence Recall	<i>What did the camera wearer do immediately after [weighing tomatoes] at the [supermarket]?</i>
• Temporal Distance	<i>How long after starting to [eat pizza] did the camera wearer [dispose of the pizza box]?</i>
Tracking	<i>List the unique [individuals] the camera wearer interacted with at the [drugstore].</i>
Visual Reasoning	
Spatial	
• Relationship	<i>Where was the [microwave] placed in relation to the [stove] in the [kitchen]?</i>
• Proximity	<i>Is the [microwave] closer to the [fridge] compared to the [sink]?</i>
• Layout	<i>Which is the correct [IMAGE] depicting the layout of the camera wearer’s [apartment]?</i>
Temporal	
• Duration	<i>Which activity did the camera wearer spend more time on: [cooking] or [playing the piano]?</i>
• Frequency	<i>Did the camera wearer use the [circular saw] or [crosscut saw] more frequently to [cut wood]?</i>
• Pre-requisites	<i>What preparation steps did the camera wearer take before [baking cookies]?</i>
Predictive	<i>What is the most likely activity the camera wearer will do next after [doing laundry]?</i>
Causal	<i>Why did the camera wearer [leave the garage for the second time]?</i>
Counterfactual	<i>What if the camera wearer used the [oven] to [cook mashed potatoes]?</i>
Navigation	
Room-to-Room	<i>How did the camera wearer get from the [building entrance] to the [apartment]?</i>
Object Retrieval	<i>How can the camera wearer retrieve the [TV remote] if they are in the [kitchen]?</i>

Table 1: **Our Proposed Task Suite with Question Prototypes.** This table shows all 4 tasks and 18 sub-tasks proposed in **HourVideo**, along with the corresponding handcrafted question prototypes designed to evaluate long-form video-language understanding capabilities.

2.2 Dataset Generation Pipeline

In this section, we provide an overview of the question-answer creation pipeline that we developed to create HourVideo. The pipeline is summarized in Fig. 2.

Video curation, Stage 1. A crucial design consideration for this benchmark is the selection of video sources and types. We chose the Ego4D [13] dataset for our videos for multiple reasons: (1) its egocentric perspective aligns well with the typical visual input for autonomous agents and assistants; (2) it features extensive visual narrations, which aid in creating diverse multiple-choice questions; and (3) it is readily accessible under the Ego4D license. We manually reviewed 1,470 videos, ranging from 20 to 120 minutes, from the Ego4D dataset, assessing their potential to generate relevant questions for various tasks in our task suite. Following this process, we curated 500 videos.

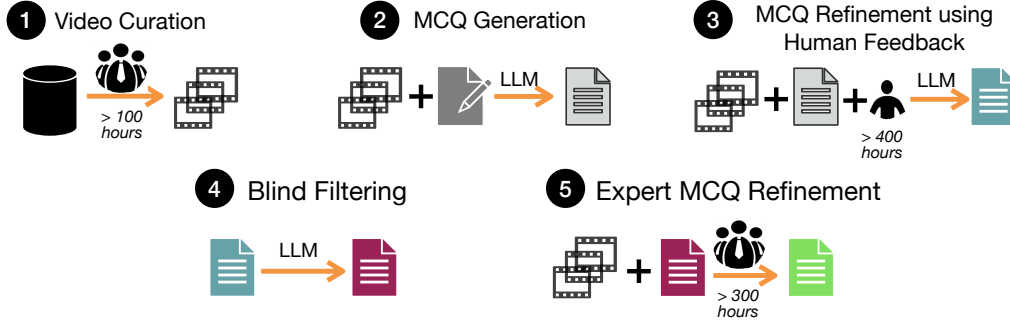


Figure 2: **Our dataset generation pipeline.** We develop a dataset generation pipeline consisting of five stages to create HourVideo. We leverage over *800 hours of human effort* in total corresponding to Video curation (Stage 1), MCQ Refinement using Human Feedback (Stage 3) and Expert MCQ Refinement (Stage 5) stages. We use LLMs in MCQ Generation (Stage 2) and MCQ Refinement using Human Feedback (Stage 3). We note that causal, counterfactual and navigation questions are manually generated by human experts (See Sec. 2.2 for details).

Candidate MCQ Generation, Stage 2. The objective of this stage is to produce high-quality MCQs for each task, requiring analysis and synthesis of information across multiple temporal segments in a long-form video. Initially, we manually develop question template(s) for each task in the suite. As shown in Table 1, transforming a question template into an actual question involves incorporating video-specific information tailored to the task and template. To facilitate this, we utilize the detailed narrations from the Ego4D dataset, transforming them into a structured format that can be processed by an LLM. Specifically, we segment the video at 20-minute intervals, with each segment’s representation including a summary and a list of tools, food items, technology, humans, pets, and physical locations encountered by the camera wearer in the video. We note that synthesizing a structured representation and a question template into a valid question with correct and incorrect answers presents a significant challenge, even for advanced LLMs. Consequently, for each task, we formulate detailed prompts that offer question prototypes, comprehensive instructions, in-context examples, and step-by-step guidance on how to transform a question template into a valid candidate MCQ_2 . In total, we developed 25 task-specific prompts.

MCQ Refinement with LLMs using Human Feedback, Stage 3. The purpose of this phase is to refine MCQ_2 , created in the previous stage. MCQ_2 may contain invalid questions, incorrect answers, trivial incorrect options, and various other issues. We identified that a significant source of these issues stemmed from relying on the noisy narrations in Ego4D. For example, different narrators within the same video could refer to a dishwasher as a "plate rack" or use other terms, and an individual might be described as an "adult," "person with a red and white shirt," "man Y," or "teenager" at various times in the narration. These inconsistencies, combined with our automatic question generation in the first stage, could lead to generation of invalid MCQs. To address noisy MCQs, we implement a human feedback system where trained annotators are tasked with: 1) assessing the validity of each question to ensure it aligns with the video content, 2) verifying the accuracy of the given answer—if found incorrect, they provide the correct answer in free-form text, 3) ensuring that all incorrect options are factually wrong and clearly distinguishable from the correct answer. We gather human feedback for all MCQ_2 , involving over 400 hours of human effort. We then design prompts, to automatically refine MCQ_2 using this human feedback to produce MCQ_3 . We engaged seven trained annotators in this stage.

Blind filtering, Stage 4. Modern LLMs possess extensive prior knowledge and can thus easily answer certain questions without needing to analyze the videos. The objective of this phase is to eliminate questions that can be answered through prior knowledge or can be trivially answered without requiring any information from the video. To address this, we do blind filtering of MCQ_3 , utilizing two separate blind LLMs (GPT-4-turbo and GPT-4). Specifically, we exclude any MCQ that is correctly answered by at least one LLM without video input. Although this method may aggressively remove MCQs, it ensures that the remaining MCQ_4 are of high quality and specifically tailored to test long-form video-language understanding.

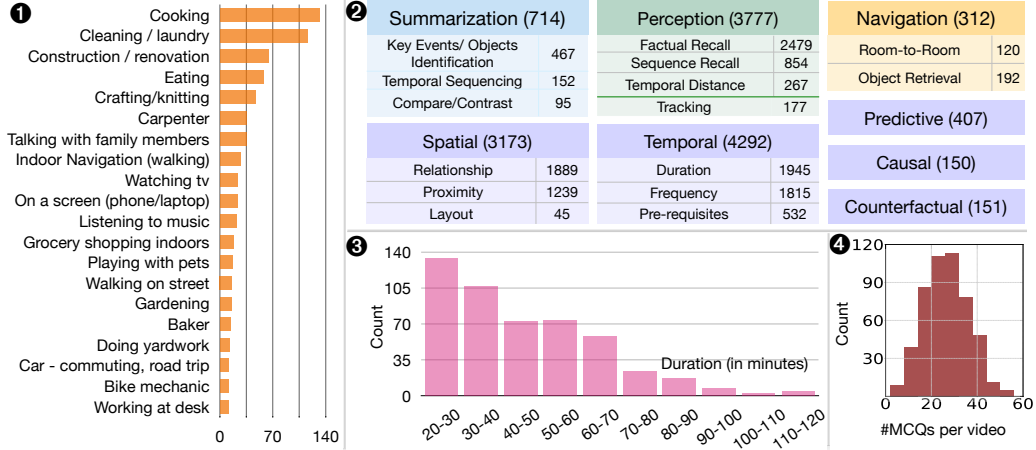


Figure 3: **Dataset Statistics.** ①: HourVideo includes 500 videos sourced from the Ego4D dataset, spanning 77 everyday scenarios. The bar chart shows the top 20 scenarios. ②: We report the number of MCQs per task/sub-task. In total, there are 12,976 questions in HourVideo. ③: We show the distribution of video duration in HourVideo. The average duration of videos in HourVideo is 45.7 minutes, with 113 videos extending beyond one hour. ④: We show the distribution of number of MCQs per video. On average, each video contains 26 MCQs.

Expert Refinement, Stage 5. The aim of this stage is to enhance the quality of MCQ₄ by utilizing a selected group of expert human annotators. This stage serves as a comprehensive step to address various remaining issues that might have persisted through prior stages. Examples of expert refinement include transforming a broad question like "Where did the camera wearer leave the keys?" into a more precise query: "Where did the camera wearer leave the bike keys after returning home from shopping?" Over 300 hours of expert human effort are employed in this stage to carefully examine and refine MCQ₄, culminating in a high-quality MCQ₅. We engaged four human experts in this stage.

Manual Generation. Despite our extensive efforts to automate fully or partially, we discovered that certain tasks did not align well with the pipeline we described earlier. Specifically, for causal, counterfactual, spatial layout and navigation tasks, we found it more effective to manually generate questions with human experts rather than processing through our multi-stage pipeline. Consequently, for these tasks in our benchmark, we generated high-quality questions, albeit in a smaller quantity. In total, 5.1% of the MCQs in HourVideo were manually generated.

Implementation details. We used GPT-4 in our pipeline as it offers impressive capabilities to follow complex multi-step instructions. We used the Chain-of-Thought [20] prompting strategy and a temperature of 0.1 for all stages involving LLMs in our pipeline.

2.3 HourVideo Statistics

HourVideo consists of 500 videos from the Ego4D dataset, covering 77 *daily life scenarios* such as cooking, cleaning, eating, watching TV, baking, etc. (Fig. 3). The dataset includes 381 hours of video footage, with video durations ranging from 20 to 120 minutes (Figure 3). On average, each video is approximately 45.7 minutes long, which 15× larger than prior work in long-form video-language understanding [12]. Additionally, 113 videos in our dataset exceed one hour in length. Each video is accompanied by an average of 26 high-quality, five-way multiple-choice questions, totaling 12,976 questions in the dataset. Finally, we strive to ensure an even distribution of MCQs across all tasks in our suite, with the exception of causal, counterfactual, and navigation tasks, where questions were manually generated for a selected group of videos.

3 Experiments

3.1 Evaluation Protocol

HourVideo includes five-way multiple-choice questions, for which we report accuracies per task and in aggregate across the entire dataset. A significant challenge in evaluating MCQs over long videos is preventing information leakage across questions. Ideally, each MCQ should be evaluated independently to avoid this issue, but unfortunately, this approach is computationally expensive and time-consuming. Therefore, for our evaluation, we assess the questions in batches, with each batch containing all questions related to a specific task or sub-task. For predictive tasks (reasoning), we provide precise timestamps to trim the videos for targeted evaluation. Details on tasks and sub-tasks requiring independent evaluation are provided in the Supplementary material.

3.2 Baselines

In this section, we compare the performance of different multimodal models on understanding long videos in a zero-shot setting. Specifically, we evaluate three classes of models: (1) Blind LLMs, (2) Socratic Models [21], and (3) Native multimodal models. All these models operate under a common function $A = M(V, \tau, Q)$ where V, τ, Q, M, A refer to the long-form video input, prompt (instruction), multiple-choice question, multimodal model, and text output respectively.

Blind LLMs. Modern LLMs possess extensive prior knowledge, enabling them to easily answer certain questions without the need to analyze videos. Furthermore, it is likely that some questions can be trivially answered by exploiting biases in the question-answer pairs. The ‘blind’ LLM baseline is designed to evaluate this by asking the LLM to answer the multiple-choice question without considering any visual information from the video, i.e., $A = M(\tau, Q)$, where τ is a generic task-agnostic prompt prepended to the question Q . We use GPT-4 [22] as our LLM for this baseline.

Socratic Models. Most current state-of-the-art multimodal models are unable to process very long videos. Therefore, to benchmark these models, we use the Socratic models approach [21]. In this approach, the video V , with a total duration of t minutes, is segmented into one-minute intervals, each denoted as $V[i]$ for minute i . Each segment $V[i]$ is independently captioned, yielding a sequence of captions $z_1, z_2, z_3, \dots, z_t$, where $z_i = \text{Video-Captioner}(V[i])$. These captions are aggregated to form a comprehensive language-based representation of the video, referred to as the world state history, which includes timestamps. This textual representation, along with a generic task-agnostic prompt τ , serves as the input for long-form video-question answering: $A = M([\tau, z_1, z_2, \dots, z_t, Q])$. We sample one-minute video clips at a rate of 0.5 fps and a resolution of 512×384 . We test using both GPT-4 [22] and LLaVA-NeXT-34B-DPO [14] as the Video-Captioner. Finally, we use GPT-4 for actual question answering, as LLaVA-NeXT-34B-DPO does not support the extended context length required to process our world state history.

Native Multimodal Models. Multimodal video models, such as Gemini 1.5 Pro [3], are trained *jointly* on multimodal data, including audio, video, images, and text. These models are particularly adept at handling very long context lengths (2M+), making them ideal for end-to-end evaluation using our benchmark. Evaluating these models is straightforward, as they can directly process hour-long videos as $A = M(V, \tau, Q)$. For all experiments, we use a sampling rate of 0.5 frames per second, a resolution of 512×384 , and a temperature setting of 0.1.

Human performance. Due to the high costs associated with human evaluations, we sampled 14 videos from our benchmark, which included more than 18 scenarios in total including crafting/painting, cooking, construction/renovation, gardening, cleaning/laundry and yard work. We ask three human experts to conduct evaluations on 11.2 hours of video content, encompassing a total of 213 MCQs. The human experts achieve an accuracy of **85.0%**. The results are shown in Fig. 4.

3.3 Results

We report all task and sub-task level quantitative results in Tab. 3. Qualitative evaluations, including human evaluation numbers, are presented in Fig. 4. We remark that random guessing corresponds to 20% accuracy. Below, we discuss our key observations.

	Summarization			Perception				Visual Reasoning							Navigation		Avg.		
	Key Events/ Objects	Temporal Sequencing	Compare/ Contrast	Factual Recall	Sequence Recall	Temporal Distance	Tracking	Relationship	Proximity	Layout	Duration	Frequency	Pre-requisites	Predictive	Causal	Counterfactual	Room-to-Room	Object Retrieval	
Blind LLMs																			
GPT-4	22.7	29.6	24.2	21.9	15.2	20.6	15.8	14.9	21.4	22.2	23.6	19.3	14.7	14.5	18.7	21.2	15.8	18.8	19.6
Socratic Models																			
LLaVA-34B-DPO	34.0	35.5	35.8	30.3	19.3	12.7	34.5	18.3	15.3	26.7	21.3	17.9	23.5	20.9	21.3	22.4	20.8	22.4	22.3
GPT-4	40.5	41.5	43.2	33.1	20.0	20.2	36.7	18.5	21.7	37.8	25.3	22.9	27.1	24.1	24.7	26.5	20.0	26.6	25.7
Multimodal Models																			
Gemini 1.5 Pro*	56.4	59.5	46.7	41.8	33.6	19.7	35.7	27.4	38.2	21.4	37.2	35.4	46.8	46.3	41.0	38.7	19.2	33.9	37.3

Table 3: **Baseline results on HourVideo.** We report results for Blind LLMs (GPT-4), Socratic models with GPT-4 and LLaVA-NeXT-34B-DPO video captions, and Gemini 1.5 Pro. Gemini 1.5 Pro outperforms Blind LLMs and Socratic LLMs by a significant margin across all tasks (14 out of 18 sub-tasks).

Blind LLMs vs. Socratic LLMs. On aggregate, blind LLMs achieve an accuracy of 19.6%, indicating that our benchmark requires access to video content for effective performance. Comparing Blind LLMs and Socratic models, both variants of Socratic models perform marginally better than blind LLMs. It is worth noting that the GPT-4-based Socratic model approach performs considerably better on the summarization task (41.1%) than blind LLMs (24.4%) and LLaVA-NeXT-34B-DPO (34.6%). Qualitative comparisons are shown in Fig. 4.

Socratic models vs. Native Multimodal Models. Gemini 1.5 Pro outperforms Socratic models by a considerable margin across all 4 tasks—summarization, perception, visual reasoning, and navigation—indicating that similar models may be promising avenues toward long-form video-language understanding. On aggregate, Gemini 1.5 Pro outperforms the GPT-4-based Socratic model by 11.6%. Despite these significant improvements, it is important to note that Gemini’s performance, at 37.3%, still lags significantly behind that of human experts, who achieve 85.0%.

Independent vs. Task-level MCQ evaluation. To investigate the validity of our proposed task/sub-task level evaluation method, we conducted an ablation study where each multiple-choice question (MCQ) was evaluated independently. For this, we used 15.9 hours of video and 570 MCQs across 25 randomly selected videos. We used Gemini 1.5 Pro, which demonstrated the highest performance on HourVideo (37.3%). The results and evaluation costs are shown in Tab. 2. There is a minor drop (2.1%) in performance when evaluating each MCQ independently; however, the associated costs increase by more than threefold. These results highlight the efficiency and validity of our proposed task-level/subtask level evaluation method. We will require future benchmark submissions to indicate whether they used task-level or individual MCQ evaluation when submitting their results, allowing for greater transparency and comparability between methods.

	Performance	Total Tokens	Evaluation Cost
Task-level	38.9%	120,818,343	\$846
Individual	36.8%	374,396,885	\$2621

Table 2: Performance and evaluation cost comparison for our proposed task/sub-task level vs. individual MCQ evaluation.

4 Related Work

Dataset Comparison. Existing video benchmarks [23, 24, 11, 25–30], primarily focus on specific domains or short videos, which limit their ability to assess long-form video understanding comprehensively. Efforts like WebVid10M [31], InternVid [32], and Panda-70M [33] include detailed captions

*Our Gemini 1.5 Pro model evaluation includes 445 videos, covering 10,842 MCQ. For details, see Supplementary material.

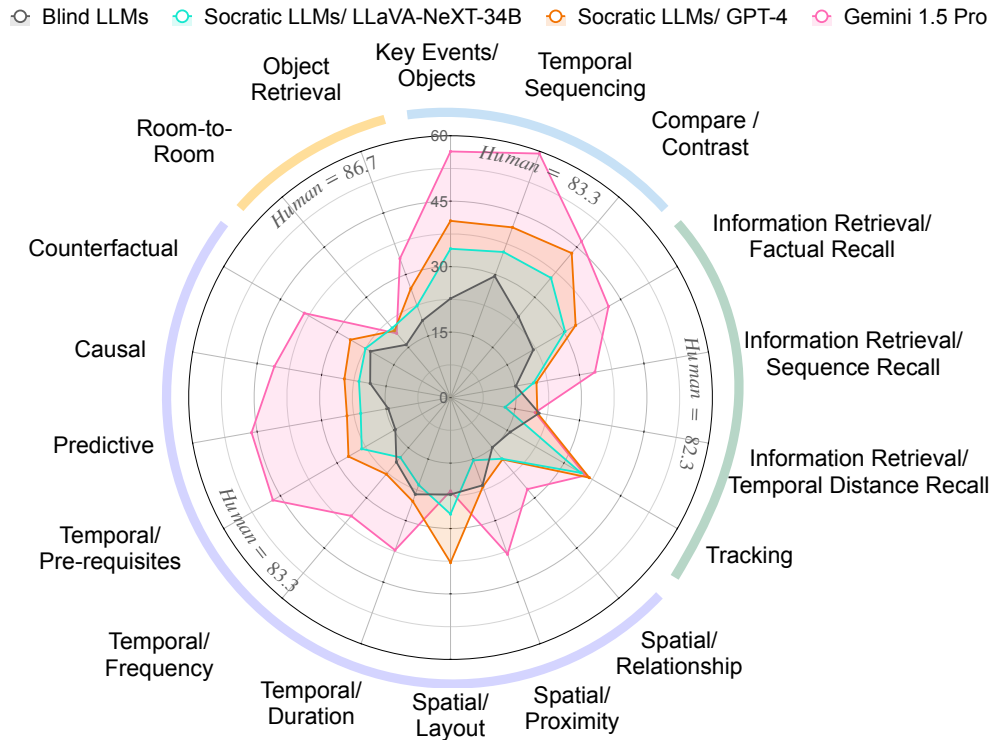


Figure 4: Comparison between different multimodal foundation models on HourVideo across different tasks/sub-tasks. We include average human expert performance for summarization, perception, visual reasoning and navigation tasks. As one can observe, current multimodal models significantly lack long-form video-language understanding capabilities.

to provide video pretraining data but consist primarily of short video clips less than one minute in length and do not provide QA pairs. Recent works have introduced several benchmarks specifically designed for long video understanding, such as Next-QA [34], Next-GQA [35], VideoChatGPT [36], EgoSchema [12], MovieChat-1K [37] and MovieNet-QA [38]. However, the average video length in these datasets is still relatively short, with Ego-Schema having an average duration of 3 minutes. In contrast, we focus on long-form video-language understanding, with videos averaging 45 minutes in duration (Table 4).

Video Understanding Tasks.

Significant efforts have been made to design tasks appropriate for evaluating multimodal large language models (MLLMs) [39–46]. The evaluation of vision-language models (VLMs) focuses mainly on visual perception tasks such as image-text matching, retrieval, captioning, object detection, and visual grounding tasks) [41–43]. Methods revolving around contrastive learning

Benchmark	# Videos	Avg. len. (mins)	# Questions
MSRVTT-QA [23]	2,990	0.25	72,821
ActivityNet-QA [11]	800	1.85	8,000
TVQA [25]	2,179	0.19	15,253
How2QA [26]	1,166	0.25	2,852
NExT-QA [34]	1,000	0.66	8,564
EgoSchema [12]	5,063	3.0	5,063
HourVideo	500	45.7	12,976

Table 4: **Dataset statistics** comparison between video understanding benchmarks.

on image-text pairs have proven to be effective methods for learning transferable representations for these visual tasks [47–49], and have been shown to be effective in more specific domains such as multi-disciplinary scientific understanding [46, 50] and multi-modal mathematical reasoning [44, 45]. Later work has improved upon the visual reasoning capabilities of VLMs [1, 51–58] and their ability

to reason across complex spatio-temporal video data [59–66]. To better evaluate spatio-temporal abilities, specific benchmarks [12, 28, 30, 67, 68] have been developed. However, the questions in many of these datasets are often not challenging enough to fully evaluate the capabilities of models in understanding long-form video content and can often be answered from only a single frame [69]. In contrast, our benchmark focuses on evaluating the capabilities needed to reason over a significantly longer duration and with more sophisticated reasoning. The questions in our dataset are designed to be highly challenging, with novel video question categories such as navigation highlighting our benchmark’s ability to effectively assess the limitations of current state-of-the-art models in long video understanding.

Long-Form Video Understanding. To extend video-language models [70–79] to long videos, the main challenge lies in efficiently encoding the temporal and spatial dynamics over a long horizon. One widely used strategy is to maintain a memory bank to store history information in long videos [80–85]. Alternatively, other methods have been proposed to compact spatio-temporal tokens into a smaller set of compressed or merged tokens to reduce redundancy and alleviate computational burden [37, 76, 86–92]. Another line of work leverages language as a bridge by first generating textual descriptions for shorter video clips sub-sampled from the longer video and then employing an LLM to aggregate the short captions for longer video understanding [93, 94]. In contrast, approaches like TimeChat [95] and VTimeLLM [96] aim to enhance temporal localization capabilities by encoding timestamp knowledge into visual tokens or using multi-stage training methods. Despite these extensive efforts, long-video understanding remains a significant challenge for the current generation of video understanding models.

5 Conclusion

We introduce **HourVideo**, a novel benchmark dataset designed to rigorously evaluate the capabilities of multimodal models to comprehend one-hour-long videos. Our dataset consists of a novel task suite comprising summarization, perception (*recall, tracking*), visual reasoning (*spatial, temporal, predictive, causal, counterfactual*), and navigation (*room-to-room, object retrieval*) tasks. This benchmark includes 500 egocentric videos from the Ego4D dataset, spanning durations of 20 to 120 minutes, and features 12,976 high-quality five-way multiple-choice questions. Our zero-shot evaluation on HourVideo reveal that multimodal models such as GPT-4V and LLaVA-NeXT exhibit performance levels only slightly better than random guessing. In stark contrast, human expert performance substantially surpasses state-of-the-art long-context multimodal model Gemini 1.5 Pro (85.0% accuracy versus 37.3%), highlighting significant research gap. We aim to establish HourVideo as a benchmark challenge to spur the development of advanced multimodal models capable of truly understanding endless streams of visual data.

Limitations and future work. Despite our substantial efforts to create a high-quality benchmark dataset, we remark that there may still be some inconsistencies within the multiple-choice questions. Additionally, while this is currently the largest long-form video-language understanding benchmark of its kind to the best of our knowledge, we acknowledge the need for more holistic benchmarks that include diverse video sources such as sports and YouTube videos. Lastly, we note that incorporating support for the audio modality is essential for more comprehensive evaluation of multimodal models. We discuss broader impact of HourVideo in Supplementary D.

Acknowledgments and Disclosure of Funding

This work was in part supported by the Stanford Institute for Human-Centered Artificial Intelligence (HAI), ONR N00014-23-1-2355, and Microsoft. This work was supported by API credit grants from Google DeepMind and OpenAI. We thank Vishal Dharmadhikari for assistance with setting up Gemini 1.5 evaluations, Hashem Elezabi and Canon Grace Pham for help with data curation. We also thank our reviewers for their valuable feedback and insights.

References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. [1](#), [9](#)
- [2] OpenAI. GPT-4V(ision) system card, 2023. [2](#)
- [3] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. [1](#), [2](#), [7](#)
- [4] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, pages 211–252, 2014. [1](#)
- [5] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. [1](#)
- [6] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. [1](#)
- [7] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 2019.
- [8] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709, 2019. [1](#)
- [9] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. [1](#), [2](#)
- [10] Luwei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.
- [11] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9127–9134, 2019. [8](#), [9](#)
- [12] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *NeurIPS*, volume 36, 2024. [1](#), [6](#), [9](#), [10](#)
- [13] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. [2](#), [4](#), [18](#), [19](#), [20](#), [23](#), [24](#)
- [14] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024. [2](#), [7](#)
- [15] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thureau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017. [2](#)
- [16] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. [2](#)
- [17] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. Multibench: Multiscale benchmarks for multimodal representation learning. *arXiv preprint arXiv:2107.07502*, 2021.

- [18] Madeline C Schiappa, Shruti Vyas, Hamid Palangi, Yogesh S Rawat, and Vibhav Vineet. Robustness analysis of video-language models against visual and language perturbations. *arXiv preprint arXiv:2207.02159*, 2022.
- [19] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruysen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. [2](#)
- [20] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 2022. [6](#)
- [21] Andy Zeng, Maria Attarian, brian ichter, Krzysztof Marcin Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aavek Purohit, Michael S Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic models: Composing zero-shot multimodal reasoning with language. In *The Eleventh International Conference on Learning Representations*, 2023. [7](#)
- [22] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. [7](#)
- [23] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017. [8](#), [9](#)
- [24] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: toward spatio-temporal reasoning in visual question answering. In *CVPR*, 2017. [8](#)
- [25] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. TVQA: Localized, compositional video question answering. In *EMNLP*, 2018. [8](#), [9](#)
- [26] Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luwei Zhou, Xin Eric Wang, William Yang Wang, et al. Value: A multi-task benchmark for video-and-language understanding evaluation. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*, 2021. [9](#)
- [27] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024.
- [28] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. *ArXiv preprint*, 2023. [10](#)
- [29] Xiuyuan Chen, Yuan Lin, Yuchen Zhang, and Weiran Huang. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. *ArXiv preprint*, 2023.
- [30] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *ArXiv preprint*, 2024. [8](#), [10](#)
- [31] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. [8](#)
- [32] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. [8](#)
- [33] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. *arXiv preprint arXiv:2402.19479*, 2024. [8](#)
- [34] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, pages 9777–9786, 2021. [9](#)
- [35] Junbin Xiao, Yao Angela, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *arXiv*, page preprint, 2023. [9](#)

- [36] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024. 9
- [37] Enxin Song, Wenhao Chai, Guan hong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*, 2023. 9, 10
- [38] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 709–727. Springer, 2020. 9
- [39] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *ArXiv preprint*, 2023. 9
- [40] Chaoyou Fu, Renrui Zhang, Haojia Lin, Zihan Wang, Timin Gao, Yongdong Luo, Yubo Huang, Zhengye Zhang, Longtian Qiu, Gaoxiang Ye, et al. A challenger to gpt-4v? early explorations of gemini in visual expertise. *ArXiv preprint*, 2023.
- [41] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv preprint*, 2023. 9
- [42] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *ArXiv preprint*, 2023.
- [43] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *ArXiv preprint*, 2023. 9
- [44] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *ArXiv preprint*, 2023. 9
- [45] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *ArXiv preprint*, 2024. 9
- [46] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhua Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *ArXiv preprint*, 2023. 9
- [47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 9
- [48] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.
- [49] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, Proceedings of Machine Learning Research, pages 12888–12900. PMLR, 2022. 9
- [50] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *ArXiv preprint*, 2024. 9
- [51] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 9

- [52] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.
- [53] OpenAI. Gpt-4 technical report, 2023.
- [54] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigt-4: Enhancing vision-language understanding with advanced large language models. *ArXiv preprint*, 2023.
- [55] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *ArXiv preprint*, 2023.
- [56] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. mplug-docowl: Modularized multimodal large language model for document understanding. *ArXiv preprint*, 2023.
- [57] Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023.
- [58] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *ArXiv preprint*, 2024. 9
- [59] Shuangrui Ding, Rui Qian, and Hongkai Xiong. Dual contrastive learning for spatio-temporal representation. In *Proceedings of the 30th ACM international conference on multimedia*, pages 5649–5658, 2022. 10
- [60] Shuangrui Ding, Weidi Xie, Yabo Chen, Rui Qian, Xiaopeng Zhang, Hongkai Xiong, and Qi Tian. Motion-inductive self-supervised object discovery in videos. *arXiv preprint arXiv:2210.00221*, 2022.
- [61] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3299–3309, 2021.
- [62] Shuangrui Ding, Rui Qian, Haohang Xu, Dahua Lin, and Hongkai Xiong. Betrayed by attention: A simple yet effective approach for self-supervised video object segmentation. *arXiv preprint arXiv:2311.17893*, 2023.
- [63] Rui Qian, Yuxi Li, Huabin Liu, John See, Shuangrui Ding, Xian Liu, Dian Li, and Weiyao Lin. Enhancing self-supervised video representation learning via multi-level feature optimization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7990–8001, 2021.
- [64] Rui Qian, Shuangrui Ding, Xian Liu, and Dahua Lin. Static and dynamic concepts for self-supervised video representation learning. In *European Conference on Computer Vision*, pages 145–164. Springer, 2022.
- [65] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.
- [66] Rui Qian, Shuangrui Ding, Xian Liu, and Dahua Lin. Semantics meets temporal correspondence: Self-supervised object-centric learning in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16675–16687, 2023. 10
- [67] Shicheng Li, Lei Li, Shuhuai Ren, Yuanxin Liu, Yi Liu, Rundong Gao, Xu Sun, and Lu Hou. Vitatecs: A diagnostic dataset for temporal concept understanding of video-language models. *ArXiv preprint*, 2023. 10
- [68] Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiayi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *ArXiv preprint*, 2023. 10
- [69] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "video" in video-language understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2917–2927, 2022. 10
- [70] KunChang Li, Yanan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *ArXiv preprint*, 2023. 10

- [71] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *ArXiv preprint*, 2023.
- [72] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *ArXiv preprint*, 2023.
- [73] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Ming-Hui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *ArXiv preprint*, 2023.
- [74] Enxin Song, Wenhao Chai, Guan hong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tianbo Ye, Yang Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. *ArXiv preprint*, 2023.
- [75] Haogeng Liu, Qihang Fan, Tingkai Liu, Linjie Yang, Yunzhe Tao, Huaibo Huang, Ran He, and Hongxia Yang. Video-teller: Enhancing cross-modal generation with fusion and decoupling. *ArXiv preprint*, 2023.
- [76] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *ArXiv preprint*, 2023. 10
- [77] Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *ArXiv preprint*, 2023.
- [78] Raehyuk Jung, Hyojun Go, Jaehyuk Yi, Jiho Jang, Daniel Kim, Jay Suh, Aiden SJ Lee, Cooper Han, Jae Lee, Jeff Kim, Jin-Young Kim, Junwan Kim, Kyle Park, Lucas Lee, Mars Ha, Minjoon Seo, Abraham Jo, Ed Park, Hassan Kianinejad, SJ Kim, Tony Moon, Wade Jeong, Andrei Popescu, Esther Kim, EK Yoon, Genie Heo, Henry Choi, Jenna Kang, Kevin Han, Noah Seo, Sunny Nguyen, Ryan Won, Ye Eun Park, Anthony Giuliani, Dave Chung, Hans Yoon, James Le, Jenny Ahn, June Lee, Maninder Saini, Meredith Sanders, Soyoung Lee, Sue Kim, and Travis Couture. Pegasus-v1 technical report. *ArXiv preprint*, 2024.
- [79] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava : Parameter-free llava extension from images to videos for video dense captioning. *ArXiv preprint*, 2024. 10
- [80] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 284–293, 2019. 10
- [81] Ivana Balažević, Yuge Shi, Pinelopi Papalampidi, Rahma Chaabouni, Skanda Koppula, and Olivier J Hénaff. Memory consolidation enables long-context video understanding. *arXiv preprint arXiv:2402.05861*, 2024.
- [82] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597, 2022.
- [83] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9226–9235, 2019.
- [84] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022.
- [85] Jiahao Wang, Guo Chen, Yifei Huang, Limin Wang, and Tong Lu. Memory-and-anticipation transformer for online action understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13824–13835, 2023. 10
- [86] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023. 10
- [87] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *ArXiv preprint*, 2023.
- [88] Suyuan Huang, Haoxin Zhang, Yan Gao, Yao Hu, and Zengchang Qin. From image to video, what do we need in multimodal llms? *arXiv preprint arXiv:2404.11865*, 2024.
- [89] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36, 2024.

- [90] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. *arXiv preprint arXiv:2404.05726*, 2024.
- [91] Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *arXiv preprint arXiv:2311.08046*, 2023.
- [92] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. *arXiv preprint arXiv:2404.03384*, 2024. 10
- [93] Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. Video recap: Recursive captioning of hour-long videos. *arXiv preprint arXiv:2402.13250*, 2024. 10
- [94] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. *arXiv preprint arXiv:2312.17235*, 2023. 10
- [95] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. *ArXiv preprint*, 2023. 10
- [96] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. *ArXiv preprint*, 2023. 10
- [97] Jiawei Wang, Liping Yuan, and Yuchen Zhang. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*, 2024. 20, 21

HourVideo Supplementary Material

- Section **A** : HourVideo Release v1.0
- Section **B** : Data Generation Pipeline: Additional details
 - Section **B.1** : Prompt Design
 - Section **B.2** : Narration Compilation Details
 - Section **B.3** : Human Feedback and Expert Refinement Details
- Section **C** : Additional Experiments and Evaluation Details
 - Section **C.1** : Evaluation details
 - Section **C.2** : Additional Baselines
 - Section **C.3** : Model Refusal Rates
- Section **D** : Broader Impact
- Section **E** : Additional Information for Checklist
 - Section **E.1** : Amount of Compute
 - Section **E.2** : Limitations
 - Section **E.3** : Potential Negative Societal Impact

A HourVideo Release v1.0

We are releasing HourVideo v1.0, our proposed benchmark dataset for one-hour video-language understanding. The benchmark dataset is provided as a single JSON file for ease of use and for straightforward integration with existing benchmarking pipelines. For each video, the dataset includes metadata and contains multiple-choice questions covering multiple tasks from our proposed task suite. Each task is accompanied by a set of multiple-choice questions, each with five possible answers. For predictive visual reasoning tasks, relevant timestamps are provided to allow precise video trimming. Additionally, a PyTorch dataloader is provided to efficiently load the video and the benchmark dataset. We provide all the 500 video_ids used in our benchmark, and users can simply download the corresponding videos from the Ego4D website after reviewing and accepting the Ego4D license agreement. We provide 2 sample videos with annotations from HourVideo. All materials are available at hourvideo.stanford.edu.

- **Structure:** HourVideo v1.0 release is organized as follows :
 - **data/**
 - * `HourVideo_v1_0.json`: Contains all 12976 questions in the benchmark dataset.
 - * `navigation_images/`: Contains all images which are part of the navigation task.
 - * `spatial_layout_images/`: Contains all images which are part of the spatial layout (reasoning/spatial) task.
 - * `sample_annotations/`: Given that **HourVideo** is an evaluation benchmark, ground truth annotations will not be released to public. For review purposes, we provide ground truth annotations for 2 sample videos as csv files.
 - * `csv/`: We provide the benchmark in individual csv files for each video to enhance accessibility, allowing users to conveniently view the contents for each video separately.
 - **src/**
 - * `video_utils.py`: A script for video processing functionalities.
 - * `hourvideo_data_loader.py`: A PyTorch DataLoader script designed to efficiently load and preprocess the dataset.
 - * `baselines/`: Contains all prompts and code for captioning/ question answering for Blind LLMs, Socratic models and Multimodal Video Models. **Remark:** Except for LLaVA-NeXT-34B-DPO captioning experiments, all other experiments require access to proprietary models including GPT-4 and Gemini 1.5 Pro.
- **Documentation:** We provide a comprehensive datasheet explaining the benchmark dataset’s purpose and intended usage.
- **License:** HourVideo will be made publicly available under MIT License. Do note that Ego4D videos are publicly available under the Ego4D License [13].
- **Versioning and Updates:** We will maintain HourVideo, with all updates and new versions announced publicly.
- **Contact Information:** For additional inquiries, please contact keshik@stanford.edu.

B Data Generation Pipeline: Additional details

B.1 Prompt Design

We meticulously design 25 prompts in total for tasks/ sub-tasks in our proposed task suite. For 9 out of 15 tasks, we generate questions first, followed by jointly generating answers and wrong answers. For the predictive visual reasoning and temporal pre-requisites tasks, we jointly generate questions and answers first, followed by generating wrong answers. For causal, counterfactual, spatial layout and navigation tasks, we generate questions, answers and wrong answers manually. We also designed prompts for narration compilation (See Fig. B.2) and paraphrasing answers for the summarization, temporal pre-requisites, and predictive visual reasoning tasks.

B.2 Narration Compilation Details

We segment all our videos at 20 minute intervals and extract a semi-structured representation which includes `title`, `description`, `start_identifier`, `end_identifier`, `list of tools`,

list of food items, list of technology objects, list of humans interacted, list of pets interacted and list of unique locations in the video segment. Finally, these segments are compiled by a LLM to form a single structured representation for each video. The prompt is shown in Fig. B.2. Considering that Ego4D offers two independently collected sets of narrations for each video, we select the narration set with the higher token count. This design choice is based on our empirical observation that a larger number of tokens typically ensures more comprehensive coverage of visual elements. These results are shown in Fig. B.1.

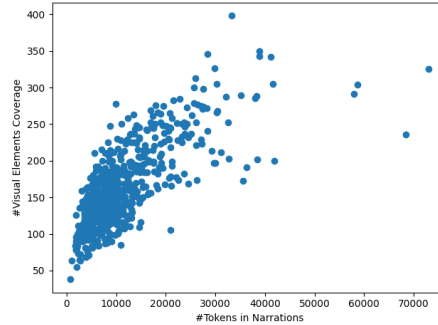


Figure B.1: This plot shows visual elements coverage vs. total number of narration tokens. We use collection of objects in ImageNet-21K, VisualGenome, Tencent1M and Places365 to quantify visual coverage. We use Tiktoken library to calculate the total number of tokens. We used Ego4D dataset [13] to perform this experiment.

B.3 Human Feedback and Expert Refinement Details

For *MCQ Refinement with Large Language Models using Human Feedback (Stage 3)*, we engaged seven annotators who had been trained to provide human feedback based on examples created by our team. Continuous quality assessments were conducted throughout this stage to ensure the integrity and high quality of the feedback obtained for MCQ Refinement. More than 400 hours of human effort were spent in this stage. For *Expert Refinement (Stage 5)*, we engaged four human experts dedicating over 250 hours of human effort for *QAW Refinement*.

C Additional Experiments and Evaluation Details

C.1 Evaluation details

Evaluation Protocol. As outlined in the main paper, we have developed an evaluation protocol that assesses multimodal models at the level of individual tasks and sub-tasks. The specific tasks and sub-tasks requiring independent evaluation are listed as follows:

- Summarization
- Perception/Information Retrieval/Factual Recall
- Perception/Information Retrieval/Sequence Recall
- Perception/Information Retrieval/Temporal Distance
- Perception/Tracking
- Visual Reasoning/Spatial/Relationship
- Visual Reasoning/Spatial/Proximity
- Visual Reasoning/Spatial/Layout
- Visual Reasoning/Temporal/Duration
- Visual Reasoning/Temporal/Frequency
- Visual Reasoning/Temporal/Pre-requisites
- Visual Reasoning/Predictive
- Visual Reasoning/Causal
- Visual Reasoning/Counterfactual
- Navigation/Room-to-Room
- Navigation/Object Retrieval
- Navigation/Room-to-Room (Image-based)
- Navigation/Object Retrieval (Image-based)

Narration Compilation

MAIN INSTRUCTIONS:

In the "Long-Form Egocentric Video Narrative Compilation" task, you are working with detailed narrations from long-form, real-world, egocentric videos. Your goal is to compile these narrations into a chronological narrative that accurately reflects the linear progression of events in the video, from the perspective of 'C', the camera wearer

For each segment, structure your summary as follows:

```
""json
{
  "segment_title": "<Generated Title>",
  "segment_description": "<Generated summary>",
  "segment_start_identifier": "<Starting Unique Identifier>",
  "segment_end_identifier": "<End Unique Identifier>",
  "segment_tool_list": "<List of tools used in the segment>",
  "segment_food_list": "<List of food related items used in the segment>",
  "segment_technology_list": "<List of technology related objects used in the segment>",
  "segment_humans_list": "<List of humans/pets that C interacted with in the segment>",
  "segment_pets_list": "<List of pets that C interacted with in the segment>",
  "segment_locations_list": "<List of specific, named locations that C visited or mentioned in the segment>"
}
```

CHRONOLOGICAL NARRATIVE COMPILATION GUIDELINES:

1. Concise Segment-Based Narrative Construction:
.....
2. Clarity, Brevity, and Object Emphasis in Language:
.....
3. Narrative Integrity and Object Relevance:
.....
4. Objective and Efficient Representation:
.....

STRICTLY AVOID:

.....

ADDITIONAL CONSIDERATIONS:

.....

FORMATTING INSTRUCTIONS:

.....

EXAMPLE OUTPUT:

```
[
  {
    "segment_title": "Initial Activities in Living Room and Kitchen",
    "segment_description": "'C' starts in the living room and then moves to the kitchen. She stands up, walks around, interacts with a man named K, and uses her phone. In the kitchen, 'C' opens the fridge, takes out potatoes, and then moves to the kitchen counter where she begins to peel potatoes., highlighting the use of technology and tools like a phone and knife, and the involvement of Man K and a dog.",
    "segment_start_identifier": "T0000_0",
    "segment_end_identifier": "T0010_17",
    "segment_tool_list": ["Phone", "Knife", "Fridge"],
    "segment_food_list": ["Potatoes"],
    "segment_technology_list": ["Phone"],
    "segment_humans_list": ["Man K"],
    "segment_pets_list": ["Dog"],
    "segment_locations_list": ["Kitchen", "Living Room"]
  },
  .....
```

DENSE NARRATIONS:

<Insert Video Narrations below>

Figure B.2: **Our Narration Compilation Prompt** In this Figure, we show our prompt for Narration Compilation task. This prompt is designed to compile dense narrations to a structured format, providing step-by-step instructions, formatting guidelines and output examples for narration compilation. The dense narrations are obtained from Ego4D [13].

This structured approach minimizes information leakage across questions and mitigates the substantial costs associated with individual MCQ evaluation. It is important to note that the costs of individual MCQ evaluation are proportional to the number of questions, emphasizing the need for our proposed assessment strategy.

C.2 Additional Baselines

We conduct an additional experiment using the recently released Tarsier model [97] which reports state-of-the-art results in multiple short-form video understanding benchmarks. Following the exact

setup in Tarsier for long-video understanding, we use the publicly available Tarsier-7B model with 16 frames uniformly sampled from the entire video. The results are reported in Tab. C.1.

Prompts. For all our baseline experiments, we use a generic task-agnostic prompt together with the video and MCQ tests for evaluation. All our prompts for baseline evaluations are included in the evaluation toolkit. We leave advanced prompting strategies for future work.

	Summarization	Perception	Visual Reasoning	Navigation	Avg.
Blind LLMs					
GPT-4	24.4	20.0	19.1	17.6	19.6
Socratic Models					
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
Multimodal Models					
Gemini 1.5 Pro	55.8	38.2	35.7	28.1	37.3
SOTA short-form video model					
Tarsier-7B (16 frames)	32.2	24.7	27.4	17.9	26.7

Table C.1: **Additional results on HourVideo using Tarsier-7B [97].** Tarsier-7B (16 frames) performance is comparable to Socratic LLMs.

C.3 Model Refusal Rates

Proprietary models, such as GPT-4 and Gemini 1.5 Pro can abstain from responding to MCQs for various reasons, including video content filtering, privacy concerns, and other undisclosed factors. In particular, we observed that the model refusal rates were significantly higher for Gemini 1.5 Pro compared to GPT-4. For Socratic models, both GPT-4 and LLaVA-34B-DPO models successfully caption more than 96% of the 1-min segments. We report refusal rates for question-answering in Tab. C.2.

Model	Videos/MCQs answered	Refusal rate
GPT-4 (Blind)	500 / 12,930	0.35%
GPT-4 (Socratic)	500 / 12,959	0.13%
LLaVA-34B-DPO (Socratic)	500 / 12,953	0.18%
Gemini 1.5 Pro	445 / 10,842	16.45%

Table C.2: **Model refusal rates:** We report refusal rates for various models for 500 videos / 12,976 MCQs. For Socratic LLMs, we report the refusal rates for question answering. The refusal rate for Gemini 1.5 Pro is significantly higher compared to GPT-4.

D Broader Impact

The Long-form Video-Language Understanding Benchmark (HourVideo) introduced in this work has the potential to significantly advance the field of AI video understanding and enable a wide range of useful applications. By focusing on long-form video, HourVideo challenges models to demonstrate high-level reasoning and comprehension skills that more closely mirror human intelligence. Success on this benchmark could lead to AI systems that can effectively perceive and interact with the real world over extended periods of time, unlocking transformative capabilities in areas like embodied AI and robotics, autonomous vehicles, smart environments, and augmented/virtual reality.

Embodied AI and robotics, which aim to develop artificial agents that can perceive, navigate, and physically interact with their environment, could benefit greatly from advances in long-form video understanding. A robot or embodied agent that can maintain a coherent, long-term understanding of its surroundings and goals would be far more capable and adaptable than one operating with only short-term perception. It could handle more complex, multi-stage tasks, learn from extended observations, and build rich mental models to support planning and decision making. For example, a home robot with long-form video understanding could tidy up a room by keeping track of object locations, understanding the steps involved in cleaning tasks, and adapting to unexpected obstacles or messes. Similarly, an industrial robot with long-term video comprehension could perform intricate assembly tasks, monitor and maintain complex machinery, or collaborate seamlessly with human

workers. Long-form video understanding is thus a key missing piece in realizing the full potential of embodied AI and robotics.

Progress on HourVideo could also contribute to the development of large world models – AI systems that learn comprehensive, multi-modal representations of the world from vast amounts of data. By processing and consolidating information from extended video sequences, these models could construct more complete and coherent world knowledge that spans time and integrates multiple levels of abstraction. Long-form video understanding would allow these models to not just recognize isolated snapshots, but grasp the flow of events, the persistence and transformation of objects, the rules of physics and causality, and the complex interactions between agents and their environments. This deep, temporally-informed world knowledge could in turn support more advanced reasoning, prediction, planning, and generalization.

In autonomous vehicles, long-form video understanding could enable more robust navigation and decision-making by considering the long-term context and anticipating future events in a driving scene. Intelligent monitoring systems could summarize and flag salient events in surveillance video with greater nuance and fewer false positives.

Long-form video understanding is also crucial for creating compelling augmented reality (AR) and virtual reality (VR) experiences. An AR system that can parse and adapt to a user’s visual context over time would be a far more capable assistant than one that merely labels objects frame-by-frame. In VR, AI characters and environments that evolve responsively to a user’s choices and actions throughout an extended interactive session would provide a deeper sense of immersion and realism.

While these exciting applications underscore the importance of advancing long-form video understanding, it is equally critical to consider the potential risks and ethical implications involved. Video data, particularly long-running egocentric video as used in HourVideo, can be highly sensitive and revealing of personal details. As AI video understanding capabilities grow, robust safeguards must be put in place to protect individual privacy, ensure secure data handling, maintain transparency around data collection and use, and prevent unauthorized surveillance or abuse. The intimate window that AR/VR systems and embodied AI agents have into users’ private spaces and behaviors further heightens these concerns. As world models become more comprehensive and powerful, it will be crucial to ensure they are developed and used in ways that respect privacy, promote fairness and transparency, and align with human values.

In designing HourVideo, we have taken care to use only videos that are licensed for research and to focus the benchmark on high-level semantic understanding rather than invasive personal information extraction. Nonetheless, the overarching trajectory toward machines that can deeply interpret the visual world will require ongoing vigilance and proactive efforts to align their development and deployment with societal values.

In summary, HourVideo offers a valuable step forward for AI video understanding, with promising implications for embodied AI, robotics, large world models, autonomous vehicles, AR/VR, and beyond. However, for long-form video understanding technology to realize its full positive potential, the AI research community must prioritize the responsible development of these powerful capabilities with strong commitments to ethics, privacy, security, and beneficial impact for humanity. We believe our benchmark will shape the progress of video understanding systems to be not only more capable, but also more trustworthy and socially beneficial.

E Additional Information for Checklist

E.1 Amount of Compute

We report the total amount of compute used for captioning 381 hours of video content using LLaVA-NeXT-34B-DPO in Table E.1. For GPT-4, we spent a total of \approx \$10,000 in credits which includes the entire dataset generation pipeline and baseline experiments (Blind LLMs and Socratic LLMs). Gemini 1.5 Pro baseline experiments cost approximately \$105 per one-hour video across all tasks/sub-tasks.

E.2 Limitations

While **HourVideo** significantly advances video-language understanding by incorporating a diverse range of tasks for extended duration with long-range dependencies, it currently uses egocentric

Table E.1: Amount of compute/ API usage used in this project. The GPU hours include computations for initial explorations/ prompt engineering / experiments to produce the reported values. CO2 emission values are computed using <https://mlco2.github.io/impact/>

Experiment	Hardware	GPU hours	Carbon emitted in kg
Main paper : Table 2 (LLaVA-NeXT-34B)	A6000	120	9.00
Main paper : Table 2 (LLaVA-NeXT-34B)	RTX A5000	24	1.66
Additional Compute for Hyper-parameter tuning	RTX A5000	12	1.80
Total		156	12.46

videos from the Ego4D dataset [13]. In the future, we plan to extend it to non-egocentric videos as our task suite and data generation pipeline do not rely on ego-centric properties. Additionally, the benchmark does not currently support audio modalities, which we acknowledge are crucial for holistic understanding of long-form videos, and we leave this for future work. In terms of annotations, every effort has been made to ensure high quality and consistency; however, the subjective nature of interpreting complex video content means that some degree of interpretative variance is inevitable. Lastly, we acknowledge that the compute required for processing extensive video content may limit accessibility for some researchers.

E.3 Potential Negative Societal Impact

Advancements in **HourVideo** benchmark could significantly enhance AI capabilities towards building autonomous agents. However, these technologies could also, for example, fuel the development of more sophisticated surveillance systems, raising significant privacy concerns. While such advancements have potential security benefits, they pose risks if used inappropriately, threatening individual privacy in public and private spaces. It is crucial that developments in video understanding are accompanied by stringent ethical standards and robust privacy safeguards to prevent misuse. We encourage ongoing dialogue and the development of comprehensive policies to ensure these technologies are used responsibly.

Checklist

1. For all authors...
 - (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? **[Yes]** Our contributions are clearly described in the abstract and introduction.
 - (b) Did you describe the limitations of your work? **[Yes]** We describe our limitations in Sec. 5 and more thoroughly in Supplementary E.2
 - (c) Did you discuss any potential negative societal impacts of your work? **[Yes]** We discuss potential negative societal impacts in Supplementary D
 - (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? **[Yes]** Our paper conforms to the NeurIPS ethics review guidelines.
2. If you are including theoretical results...
 - (a) Did you state the full set of assumptions of all theoretical results? **[N/A]**
 - (b) Did you include complete proofs of all theoretical results? **[N/A]**
3. If you ran experiments (e.g. for benchmarks)...
 - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **[Yes]** All code, annotations, and instructions for reproducing our results are provided in our project website hourvideo.stanford.edu.
 - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **[N/A]** There was no training in our benchmark (eval only).
 - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **[No]** Given that our benchmark involves heavy use of costly proprietary models (GPT-4, Gemini 1.5 Pro) for experiments, we did not repeat the experiments. We provide all our prompts/ evaluation code at hourvideo.stanford.edu.
 - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **[Yes]** Given most of our experiments involved the use of proprietary models (GPT-4, Gemini 1.5 Pro), we report the API usage in the Supplementary. For LLaVA-NeXT-34B-DPO captioning experiments, we report the amount of compute in Supplementary E.1.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
 - (a) If your work uses existing assets, did you cite the creators? **[Yes]** We cite EGO4D [13], the source of the videos for our dataset.
 - (b) Did you mention the license of the assets? **[Yes]** Our dataset uses the EGO4D License, as explained in Sec. 2.2.
 - (c) Did you include any new assets either in the supplemental material or as a URL? **[Yes]** We release our proposed video question answering benchmark dataset. Please refer to hourvideo.stanford.edu
 - (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? **[Yes]** We obtained consent via agreeing to the EGO4D License.
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **[No]** Our videos are identical in content to those of EGO4D, thus we do not explicitly discuss this in our paper. EGO4D does contain videos of humans that are identifiable and does not contain offensive content.
5. If you used crowdsourcing or conducted research with human subjects...
 - (a) Did you include the full text of instructions given to participants and screenshots, if applicable? **[Yes]** We supply the text instructions given to our annotators in the Supplementary materials.
 - (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? **[N/A]** No human participant risks.
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **[Yes]** We supply estimates of cost for our human participants in the Supplementary Materials (datasheet).