Technical Summary:
HourVideo: 1-Hour Video-Language Understanding

Kai-Yu Lu

2025/10/23

1 Research Problem and Motivation

Long-form video understanding refers to the ability of a model to perceive, track, and reason over visual information that extends for tens of minutes or hours, instead of a few seconds or short clips. The paper studies long-form *video-language understanding*, where models must answer natural language questions about hour-long egocentric videos that capture everyday activities such as cooking, cleaning, construction, shopping, or navigation in indoor and outdoor environments.

Existing multimodal benchmarks predominantly evaluate models on single images or short video clips, typically a few seconds to at most a few minutes. These setups are sufficient to test object recognition, short-term temporal reasoning, or simple question answering, but they do not adequately test whether a system can maintain and integrate information over very long temporal horizons. In practice, many real-world applications such as assistive augmented reality agents, household robots, and monitoring systems require sustained reasoning across long time spans, including tracking objects and people, understanding causal chains of actions, predicting future events, and navigating based on past observations.

A further challenge is that many existing long-video datasets still allow questions that can be solved by looking at one short segment or even a single frame, which effectively reduces the task to short-clip understanding instead of true long-range reasoning. Large language models also possess extensive prior knowledge and textual biases, especially for popular media content. If benchmarks use movies or television shows, questions may be answerable from language priors alone, without any genuine use of the visual signal. The benchmark therefore needs to focus on everyday, mundane egocentric videos and carefully designed questions that truly require combining information across multiple temporal segments.

2 Related Work

2.1 Short-form video question answering

Earlier video question answering benchmarks focus mainly on short videos. Examples include MSRVTT-QA, which provides question answering over short web videos, ActivityNet-QA, which targets complex but still relatively short web videos, TVQA, which focuses on question answering about short clips from television shows, How2QA, which uses instructional videos, and NExT-QA, which emphasizes temporal action reasoning over shorter clips. These datasets typically involve video lengths on the order of seconds to a few minutes and contain large numbers of questions, but they primarily test short-term perception and reasoning.

2.2 Existing long-form video benchmarks

More recent work has targeted longer videos. EgoSchema is a diagnostic benchmark for long-form video-language understanding, with videos averaging around three minutes and questions designed to probe temporal and causal understanding. However, the average video length still remains far below one hour, and many questions in long-form benchmarks can often be answered from a limited part of the video, which weakens the pressure to maintain long-range dependencies.

2.3 Limitations of prior benchmarks

These prior datasets reveal three key limitations that motivate HourVideo:

- **Limited temporal horizon.** Most benchmarks provide short clips, so they cannot fully evaluate memory and reasoning over tens of minutes or hours.
- **Narrow task coverage.** Many benchmarks focus on one or a few task types, such as factual recall or simple temporal relations, and do not jointly assess summarization, perception, visual reasoning, and navigation capabilities in a unified benchmark.
- **Potential reliance on priors.** For benchmarks based on popular media content, large language models may exploit textual priors or cultural knowledge to answer questions without actually processing the video.

HourVideo is designed to address these gaps by providing hour-scale egocentric videos, a broad suite of task types and sub-tasks, and questions that are deliberately constructed to require visual evidence scattered across time, thereby reducing the influence of language-only priors.

3 Dataset Construction

3.1 Source videos and scenarios

The HourVideo benchmark uses long egocentric videos sourced from the Ego4D dataset, which is a large-scale collection of first-person recordings of daily life activities. From this source, 1,470 candidate videos with durations between 20 and 120 minutes were manually reviewed by five human experts to assess suitability for long-form question construction. The final benchmark selects 500 egocentric videos covering 77 everyday scenarios, including cooking, cleaning, laundry, construction and renovation, crafting, gardening, driving, office work, shopping, and navigation around homes and streets.

Overall, the dataset contains approximately 381 hours of video footage. Individual video durations range from 20 to 120 minutes, with an average length of 45.7 minutes, which is reported as fifteen times longer than prior work in long-form video-language understanding. Among the 500 videos, 113 exceed one hour in length. Each video is associated with on average 26 high-quality five-way multiple-choice questions, for a total of 12,976 questions in the benchmark.

3.2 Task suite and question prototypes

HourVideo introduces a task suite designed to cover both perceptual and cognitive abilities in long-form video-language understanding. The tasks are organised into four main categories with 18 sub-tasks:

- **Summarization.** Key events and object-centric summarization, temporal sequencing of high-level activities, and compare or contrast questions between different phases or locations.

- **Perception.** Information retrieval tasks including factual recall, sequence recall, and temporal distance estimation, as well as tracking questions that require counting and identifying distinct individuals.
- **Visual reasoning.** Spatial reasoning about relationships, proximity, and layout, temporal reasoning about duration, frequency, and prerequisites, predictive reasoning about likely next actions, and causal and counterfactual reasoning.
- **Navigation.** Room-to-room navigation questions and object retrieval questions that require understanding the spatial layout and paths needed to reach a target location or object.

For each sub-task, the benchmark authors manually design question prototypes that specify the required information types and temporal dependencies. Examples include asking which tool was used more frequently for a given activity, how long after a certain event another event occurred, how to navigate from the kitchen to the backyard, or what would happen to cooking time if a different appliance were used.

3.3 Multi-stage MCQ generation pipeline

The dataset uses a five-stage pipeline to construct high-quality five-way multiple-choice questions (MCQs):

1. **Video curation.** Human experts select 500 suitable Ego4D videos that contain rich, narratable activities and diverse scenes, while avoiding content that is too sparse or repetitive.
2. **Candidate MCQ generation.** For each task and video, question templates are instantiated using structured representations derived from Ego4D narrations. Videos are segmented into 20-minute intervals, and for each segment an automatically compiled summary includes lists of tools, food items, technology, people, pets, and locations. Large language models are prompted with detailed instructions, task-specific templates, and in-context examples to generate candidate MCQs, referred to as MCQ2.
3. **LLM refinement with human feedback.** Trained annotators inspect each MCQ2, verify that the question is valid and answerable from the video, correct any wrong ground-truth answers, and ensure that distractor options are clearly incorrect but plausible. Their feedback is encoded in prompts to automatically refine MCQ2 into MCQ3. This stage involves more than 400 hours of human effort.
4. **Blind filtering.** To remove questions that could be answered purely from prior knowledge or textual biases, two blind large language models (GPT-4-turbo and GPT-4) answer the MCQ3 questions without any video input. Any question that at least one blind model answers correctly is discarded, yielding a filtered set MCQ4.
5. **Expert refinement and manual creation.** A smaller group of expert annotators further refine MCQ4 into final MCQ5 by tightening semantics, clarifying references, and eliminating residual issues. For some tasks, especially causal, counterfactual, spatial layout, and navigation, questions are manually written by experts rather than generated through the automated pipeline, because such questions require nuanced understanding that is difficult to automate. This stage uses more than 300 hours of expert effort.

In total, more than 800 hours of human effort are combined with large language model based generation and refinement to construct the final MCQ set. The dataset is released as a JSON file, accompanied by a PyTorch dataloader and supporting images for navigation and spatial layout tasks, as well as sample annotated videos for inspection.

Property	Value
Number of videos	500 egocentric videos from Ego4D
Total video duration	Approximately 381 hours of video
Video length range	Between 20 and 120 minutes per video
Average video length	Approximately 45.7 minutes per video
Videos longer than 60 minutes	113 videos exceed one hour
Number of scenarios	77 everyday activity scenarios (such as cooking, cleaning, shopping, driving, gardening)
Total MCQs	12,976 five-way multiple-choice questions
Average MCQs per video	Approximately 26 MCQs per video
Task and sub-task coverage	4 main tasks and 18 sub-tasks across summarization, perception, visual reasoning, and navigation

Table 1: Overview of HourVideo dataset statistics.

3.4 Dataset statistics

Table 1 summarises key statistics for HourVideo.

4 Query Protocol and Task Definitions

4.1 Multiple-choice question format and evaluation

Each HourVideo question is a five-way multiple-choice question. For each video, there are several questions belonging to different tasks and sub-tasks. Each question consists of a textual prompt and five candidate answers, exactly one of which is correct.

The primary evaluation metric is accuracy, defined as the fraction of questions for which the model selects the correct option. Accuracy is reported per sub-task, per task (for example summarization, perception, visual reasoning, navigation), and as an overall average across all questions.

Evaluating long-form MCQs raises a specific challenge. Ideally, each question would be answered independently from scratch, in order to prevent information leakage across questions. However, this independent evaluation requires repeated processing of the same long video, which is computationally very expensive for models that already operate near the limit of their context length. The benchmark addresses this by grouping questions into batches per task or sub-task for a given video and evaluating the model on all questions in the batch together. For predictive reasoning tasks, timestamps are provided to trim the video segments that must be processed.

4.2 Task and sub-task definitions

The four main tasks and their sub-tasks are defined as follows.

Summarization. Questions require high-level description of the content of the entire video or large portions of it.

- Key events and objects: summarise the main activities and object interactions, such as what the camera wearer did in a supermarket or during a construction project.
- Temporal sequencing: describe the sequence of actions in tasks such as preparing and cooking a meal.

- Compare and contrast: compare activities in different locations or time intervals, such as how behaviour in an apartment differs from behaviour in a restaurant.

Perception. Questions probe detailed factual and temporal knowledge about specific events.

- Factual recall: identify which items were picked up, which tools were used, or which locations were visited.
- Sequence recall: determine what action occurred immediately before or after a reference event, such as checking out at a cashier.
- Temporal distance: estimate how much time elapsed between two events, such as between starting to eat and discarding a food container.
- Tracking: count and identify distinct individuals that the camera wearer interacted with in a store or other setting.

Visual reasoning. Questions require reasoning beyond direct recall, incorporating spatial, temporal, predictive, causal, and counterfactual reasoning.

- Spatial reasoning: assess relationships, proximity, and layout, such as which object is closer to a reference point, or which schematic layout matches the observed room.
- Temporal reasoning: compare durations and frequencies of activities (for example how long was spent cooking versus playing the piano, or which saw was used more frequently), and identify prerequisite steps before a complex task such as baking cookies.
- Predictive reasoning: predict the most likely next action after a given sequence of events, such as what the camera wearer will do after leaving a cashier.
- Causal reasoning: explain why an event occurred, such as why a child moved a step stool near a kitchen counter.
- Counterfactual reasoning: reason about hypothetical changes, such as how total cooking time would change if a different appliance were used.

Navigation. Questions test understanding of spatial layout across rooms and the ability to plan paths.

- Room-to-room navigation: describe how to move from one location to another, for example from a building entrance to an apartment or from the kitchen to the backyard.
- Object retrieval: describe how to reach a target object starting from a given room, such as how to retrieve a motorcycle or a television remote from the kitchen.

5 Modeling Approach

The HourVideo benchmark does not propose a new model architecture. Instead, it compares several families of baseline systems under a unified evaluation formulation. The goal is to quantify how well existing models handle long-form video-language understanding and to highlight the performance gap relative to human experts.

5.1 Unified functional formulation

All evaluated models are abstracted by the following functional relationship:

$$A = M(V, \tau, Q) \tag{1}$$

In Equation (1), V denotes the long-form video input, for example an hour-long egocentric recording. The symbol τ denotes a textual prompt or instruction that tells the model how to answer multiple-choice questions, for instance by asking it to choose exactly one option. The symbol Q denotes the text of the multiple-choice question, including the five candidate answers. The symbol M denotes the multimodal model being evaluated, which may take both video and text as input. The symbol A denotes the model's textual answer, such as the letter of the chosen option. Intuitively, this equation describes the process of feeding a video and a question into a model and obtaining an answer.

5.2 Blind large language models

To quantify how many questions can be answered from language priors alone, the benchmark defines a blind large language model baseline that ignores visual input:

$$A = M(\tau, Q) \tag{2}$$

In Equation (2), the symbols M, τ, Q, and A have the same meanings as in Equation (1), but the video input V is omitted. The model receives only the task-agnostic prompt and the question text. In practice, GPT-4 is used as M in this setting. This formulation measures how far a powerful language model can go using only textual heuristics and world knowledge, without looking at the video at all.

5.3 Socratic models

Most current vision-language models cannot process entire hour-long videos directly. The benchmark therefore uses a Socratic model approach, which decomposes long video understanding into two stages: captioning short segments, and then reasoning over the resulting text.

First, the video V with total length t minutes is divided into one-minute segments $V[i]$ for $i = 1, \ldots, t$. Each segment is captioned by a video captioning model:

$$z_i = \text{VideoCaptioner}(V[i]) \tag{3}$$

In Equation (3), $V[i]$ is the visual content of minute i of the video, VideoCaptioner$(\cdot)$ denotes an automatic video captioning model such as GPT-4 or LLaVA-NeXT-34B-DPO, and z_i denotes the generated textual caption that summarises what happens during that minute. This step converts each short video segment into a short textual description.

Next, all captions are concatenated into a long textual sequence called the world state history, which encodes the temporal evolution of the scene. The question answering model then takes the prompt, the sequence of captions, and the question as input:

$$A = M([\tau, z_1, z_2, \ldots, z_t, Q]) \tag{4}$$

In Equation (4), the bracket $[\tau, z_1, z_2, \ldots, z_t, Q]$ denotes the concatenation of the prompt, the ordered captions, and the question into a single long text sequence. The model M is a large language model such as GPT-4 that operates purely on text. The output A is the selected answer option. Intuitively, this formulation instructs the model to reason about the long video indirectly by reading a time-stamped narrative of what occurred.

Model	Summarization	Perception	Visual Reasoning	Navigation	Average Accuracy
GPT-4 (blind LLM)	24.4%	20.0%	19.1%	17.6%	19.6%
LLaVA-NeXT-34B-DPO (Socratic)	34.6%	26.7%	19.1%	21.8%	22.3%
GPT-4 (Socratic)	41.0%	29.4%	22.8%	24.0%	25.7%
Tarsier-7B (short-form video model)	32.2%	24.7%	27.4%	17.9%	26.7%
Gemini 1.5 Pro (native multimodal)	55.8%	38.2%	35.7%	28.1%	37.3%

Table 2: Accuracy of various models on HourVideo tasks.

In the experiments, the video segments are sampled at 0.5 frames per second and resized to a resolution of 512×384 pixels before captioning. The final question answering is performed by GPT-4, because LLaVA-NeXT-34B-DPO does not support the very long context lengths required to ingest the entire caption sequence.

5.4 Native multimodal models

Native multimodal models are trained end-to-end on video, image, audio, and text data and are capable of handling very long contexts directly. Gemini 1.5 Pro is an example of such a model, with support for multimodal context lengths beyond two million tokens. In the benchmark, Gemini 1.5 Pro is evaluated by directly supplying the sampled frames of the long video together with the textual prompt and question, as in Equation (1). A sampling rate of 0.5 frames per second and a resolution of 512×384 pixels are used, with a temperature setting of 0.1 to obtain stable outputs.

6 Empirical Results

6.1 Overall performance comparison

Table 2 summarises the performance of different model families on HourVideo across the four main tasks and on average. Random guessing in a five-way multiple-choice setting corresponds to 20% accuracy.

Several observations emerge from these results.

First, the blind GPT-4 baseline attains an average accuracy of 19.6%, which is only marginally above random chance. This indicates that HourVideo successfully reduces the influence of language-only priors and requires genuine visual understanding.

Second, Socratic models that use captioned segments provide clear gains. The GPT-4 Socratic configuration reaches 25.7% average accuracy, with particularly strong improvement in summarization (41.0% versus 24.4% for blind GPT-4), reflecting the advantage of explicit textual world state histories.

Third, Tarsier-7B, a strong short-form video model pre-trained in part on Ego4D for video captioning, achieves an average of 26.7%, which is comparable to the GPT-4 Socratic approach but still far from human-level performance.

Fourth, Gemini 1.5 Pro, a native multimodal model capable of direct long-context video processing, significantly outperforms Socratic and blind baselines, reaching 37.3% average accuracy and leading on 14

out of 18 sub-tasks. This suggests that end-to-end long-context multimodal models are promising directions for long-form understanding.

Despite these improvements, a large gap remains compared to human experts. On a subset of 213 questions over 11.2 hours of video, human experts achieve 85.0% accuracy overall, with task-level accuracies of approximately 83.3% on summarization, 82.3% on perception, 83.3% on visual reasoning, and 86.7% on navigation.

6.2 Task-level and sub-task trends

The detailed task and sub-task analysis reveals additional trends:

- The GPT-4 Socratic model performs particularly well in summarization sub-tasks, where the explicit caption sequence aligns well with the need to aggregate global information.
- Gemini 1.5 Pro provides substantial gains in spatial reasoning, temporal comparison, and predictive and causal reasoning, reflecting its ability to integrate information over long visual contexts.
- Even the best models struggle with counterfactual reasoning and navigation sub-tasks, which require more abstract reasoning and path planning, indicating that these capabilities remain underdeveloped.

6.3 Evaluation protocol ablation

The benchmark includes an ablation study comparing the task-level evaluation protocol with an ideal but more expensive per-question evaluation protocol. Using Gemini 1.5 Pro on a subset of 25 videos (15.9 hours, 570 MCQs), task-level evaluation yields 38.9% accuracy with approximately 120.8 million tokens processed and an estimated evaluation cost of 846 USD. Per-question independent evaluation yields 36.8% accuracy, processes around 374.4 million tokens, and costs approximately 2,621 USD.

The results show a modest performance drop of 2.1 percentage points when moving to independent per-question evaluation, while the computational cost more than triples. This supports the benchmark's choice to use task-level evaluation as a practical yet reasonably faithful approximation of independent evaluation for large-scale benchmarking.

6.4 Model refusal analysis

The supplementary experiments analyse refusal rates, that is, the fraction of questions for which a model abstains from answering. Blind GPT-4 and Socratic GPT-4 have refusal rates around 0.35% and 0.13% respectively, while LLaVA-34B-DPO has a refusal rate of 0.18%. In contrast, Gemini 1.5 Pro refuses to answer around 16.45% of questions across 445 evaluated videos and 10,842 MCQs.

These higher refusal rates likely reflect additional safety filters, privacy constraints, or other internal content policies in proprietary multimodal models. From a benchmarking perspective, this emphasises the need to report not only accuracy but also the proportion of questions for which models decline to respond.

7 Summary

7.1 Research contributions

The HourVideo benchmark makes several contributions to the study of long-form video-language understanding:

- It introduces a large-scale benchmark of 500 egocentric videos and 12,976 five-way multiple-choice questions, with an average video length of 45.7 minutes and many videos exceeding one hour.
- It presents a carefully designed task suite with 18 sub-tasks spanning summarization, perception, visual reasoning, and navigation, supported by handcrafted question prototypes that enforce long-range temporal reasoning.
- It develops a multi-stage question generation pipeline that combines large language models with extensive human annotation and blind filtering, resulting in high-quality questions that are difficult to answer without genuine video understanding.
- It provides a comprehensive empirical evaluation of blind language models, Socratic models, short-form video models, and native long-context multimodal models, revealing a substantial gap between current models and human-level performance.

7.2 Limitations and future directions

The benchmark also has several limitations that suggest directions for future work.

First, despite the extensive multi-stage refinement process, some inconsistencies and residual noise in multiple-choice questions are still possible. Complex egocentric scenes and imperfect narrations can lead to ambiguous references, rare edge cases, or subtle timing ambiguities that challenge both annotators and models.

Second, HourVideo focuses on egocentric videos from Ego4D and therefore emphasises everyday activities in domestic and work environments. While this focus is appropriate for embodied agents and household assistants, it does not yet cover a broader range of domains such as sports broadcasts, educational videos, or diverse online media platforms. Expanding to more heterogeneous long-form sources would improve coverage of real-world scenarios.

Third, the current benchmark primarily targets visual information and text. Although video and language already form a rich multimodal space, many real-world applications also involve audio cues, tactile feedback, or other sensory modalities. Incorporating audio and potentially additional modalities would support more comprehensive evaluation of multimodal agents.

Finally, the benchmark is designed as an evaluation dataset and does not release full ground-truth annotations for all videos, which is appropriate for testbed usage but limits its role as a training resource. Future work may explore complementary long-form video corpora for pretraining while reserving HourVideo for held-out evaluation.

7.3 Key takeaways

Three key takeaways emerge from this study:

1. Long-form video-language understanding is significantly more challenging than short-form video question answering. Even strong models such as Gemini 1.5 Pro achieve around 37.3% accuracy on HourVideo, far below human performance of 85.0%.
2. Carefully constructed tasks and multi-stage multiple-choice question generation can substantially reduce the influence of language-only priors and force models to rely on true visual understanding over long time horizons.
3. End-to-end long-context multimodal models outperform Socratic and blind baselines but still struggle with navigation, causal reasoning, and counterfactual questions, highlighting open research directions in memory, abstraction, and planning for video-language models.