HourVideo Datasheet

Questions from the Datasheets for Datasets paper, v7.

Jump to section:

- Motivation
- Composition
- Collection process
- Preprocessing/cleaning/labeling
- Uses
- Distribution
- Maintenance

Motivation

For what purpose was the dataset created?

This benchmark dataset was created to assess extremely long-form video language understanding, specifically for videos spanning 20 to 120 minutes. While prior works have proposed videos that span several minutes, our dataset contains questions that often require watching over an hour of video content to correctly answer.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The creators of the dataset are from the Vision and Learning Lab at Stanford University.

Who funded the creation of the dataset?

The creation of this benchmark dataset was funded solely by the Stanford Vision and Learning Lab at Stanford University.

Any other comments?

No.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

Each instance contains one long video alongside a set of text and images that comprise multiple-choice questions with five answer options.

Specifically, the dataset consists of videos containing people performing various activities. Accompanying the videos, we have various multiple choice questions for evaluating the understanding of the video seqence. The videos from our dataset are exclusively taken from Ego4D.

How many instances are there in total (of each type, if appropriate)?

We have 12,976 questions spanning across 500 videos.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

The dataset is a subsample of Ego4D in which the provided captions and videos yielded high-quality video-question pairs. We specify our exact selection and collection process in the collection process section.

What data does each instance consist of?

Unprocessed videos at 30 fps in mp4 format, natural language questions as text, and five ordered multiple choice options as answers in text. An integer label from 1-5 denotes the correct answer. Only for the navigation question category, the answers are listed as images instead of text.

Is there a label or target associated with each instance?

Yes, each instance contains a single correct answer from the multiple-choice question. We denote the label with an integer label from 1-5 indicating which of the 5 ordered options are correct.

Is any information missing from individual instances?

No information is missing from any instance.

Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)?

Each long video has a unique string ID. The relationships between videos and questions are linked via the video ID. Thus, each of our videos contain multiple different questions.

Are there recommended data splits (e.g., training, development/validation, testing)?

The entire dataset is only to be used for zero-shot evaluation. Thus, it should NOT be used for training and only for model testing purposes.

Are there any errors, sources of noise, or redundancies in the dataset?

Different questions in our dataset may refer to the same video, however each video clip is distinct. We manually checked for errors at multiple steps of the dataset generation process, as described in Appendix B.3.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The dataset links to the existing Ego4D dataset. The dataset is archived and available to download via AWS. It requires the agreement to the Ego4D License (linked here). Ego4D is open for academic research, commercial product development, among other use-cases specified in the license. Please read the linked license for more details.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?

No, none of the data is considered confidential, and the dataset does not contain private information.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

We do not anticpate that any of the content of our dataset could be viewed as offensive, insulting, threatening, or anxiety-inducing. The videos are of relatively common everyday activities and are recorded in controlled settings.

Does the dataset relate to people?

Yes, the dataset contains videos of people performing various activities.

Does the dataset identify any subpopulations (e.g., by age, gender)?

No, our dataset does not explicitly identify any subpopulations.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

Yes, some individuals can be directly viewed in the videos. The face and clothing of the different actors in the dataset are sometimes visible in the raw videos.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?

Some videos may reveal the religious beliefs of individuals by showing specific religious attire they are wearing. We are committed to ensuring that this data is managed in compliance with ethical standards and relevant data protection regulations.

Any other comments?

No.

Collection process

How was the data associated with each instance acquired?

The video data was directly used from Ego4D, a publicly accessible dataset. The text data for video question answering was created via a combination of human contractors and LLMs. We describe the details of this question generation in Appendix C.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?

The video and narration data were collected in accordance with the official Ego4D guidelines for data access, which can be found at: Ego4D Data Access Guidelines. To generate the text data within our dataset, we utilized API access for GPT-4 via OpenAI. This approach enabled us to create three distinct questions for each video clip sampled from the Ego4D dataset. Once the questions were generated for each sampled video clip, we implemented a series of filtering procedures to ensure the quality and relevance of the data.

These procedures included:

QAW Refinement with LLMs using Human Feedback (Stage 3): The purpose of this phase is to refine $\QAW*{2}$, which may contain invalid questions, incorrect answers, simplistic incorrect options, and other issues arising from the noisy narrations in the Ego4D dataset. For example, different narrators within the same video might refer to a dishwasher as a "plate rack," or describe an individual inconsistently. To address these issues, we implemented a human feedback system where annotators were tasked with assessing the validity of each question to ensure it aligns with the video content. This helps verify the accuracy of the given answer, since annotators can provide the correct answer if the given answer is found incorrect, and they also ensure that all incorrect options are factually wrong and clearly distinguishable from the correct answer. Over 400 hours of human effort were dedicated to gathering feedback for all $\QAW*{2}$. This feedback was then used to automatically refine $\QAW*{2}$, resulting in $\QAW*{3}$. **Blind Filtering (Stage 4)**: The objective of this phase is to eliminate questions that can be answered through prior knowledge or trivially without needing information from the video. This was achieved by utilizing two separate blind LLMs (GPT-4-turbo and GPT-4) to review $\QAW*{3}$ without any video input. Any MCQ correctly answered by at least one LLM was excluded, ensuring that the remaining $\QAW*{4}$ required video content for accurate answers. **Expert Refinement (Stage 5)**: This stage aimed to enhance the quality of $\QAW*{4}$ by utilizing a selected group of expert human annotators. Experts dedicated over 250 hours to carefully examining and refining $\QAW*{4}$, addressing any remaining issues from prior stages. For instance, a broad question like "Where did the camera wearer leave the keys?" was refined into a more precise query: "Where did the camera wearer leave the bike keys after returning home from grocery shopping?" This process resulted in a high-quality $\QAW_{5}$. For a more detailed explanation of these procedures, please refer to Section 2.2 in the main paper.

How were these mechanisms or procedures validated?

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

Our dataset is sampled from the larger Ego4D dataset. We manually reviewed a 1,470 subset of videos ranging from 20 to 120 minutes in length. We assessed each video's potential for generating relevant questions for the various tasks in our predefined task suite. After this review and manual filtering, we were left with approximately 500 videos. More details can be found in our paper and supplementary materials.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The authors and contractors were involved in the data annotation process. The contractors were based in China, and were paid on average $5 per hour, significantly higher than the $1.27 hourly minimum wage in the country.

Over what timeframe was the data collected?

The videos in Ego4D were collected in a timespan ranging from 2019 to 2021. HourVideo was annotated during the first half of 2024.

Were any ethical review processes conducted (e.g., by an institutional review board)?

No.

Does the dataset relate to people?

Yes, the dataset contains videos of people.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

We obtained our videos containing people from Ego4D.

Were the individuals in question notified about the data collection?

Yes, the individuals signed explicit agreements with the creators of the Ego4D dataset.

Did the individuals in question consent to the collection and use of their data?

To the best of our knowledge, all the individuals in Ego4D agreed to the collection and use of their data.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

We are unsure about the details of the exact consent agreements signed by participants of the Ego4D dataset.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?

An extensive analysis of the potential impacts of the Ego4D dataset has been studied in their original paper (arXiv link). We have a more in-depth an analysis on the potential societal impacts of our dataset in the Appendix D, Appendix E.2, Appendix E.3 of our paper.

Any other comments?

No.

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

The generated questions and answers were reviewed and refined by manual curation. Detailed procedures are outlined in Section 2.2. The video clips obtained from Ego4D did not undergo any preprocessing.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

Yes, we save every step of our question generation pipeline. We do not plan to share them to prevent prompt leakage for the question generation process.

Is the software used to preprocess/clean/label the instances available?

We do not plan to share specific GPT-4 API calls used to generate our question set. We used Microsoft Excel to manually label and clean the data instances.

Any other comments?

No.

Uses

Our dataset is useful for evaluating the understanding abilities of AI systems that can take in video and text input.

Has the dataset been used for any tasks already?

The dataset has been used for the video question answering task. Specifically, in a 5-way multiple-choice setting.

Is there a repository that links to any or all papers or systems that use the dataset?

There is currently no such repository.

What (other) tasks could the dataset be used for?

The dataset could be used for open-ended video question-answering by leveraging natural language scorers including model-based evaluations. Individual question-types (e.g., summarization) could be used for evaluating specific model abilities.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

No.

Are there tasks for which the dataset should not be used?

The dataset should not be used for training so that it can continue to be used for zero-shot evaluation.

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

Our annotations will be available to download from our official GitHub repository. The raw videos are only accessible via official Ego4D distribution channels.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?

Our annotations and dataset will be distributed via GitHub.

When will the dataset be distributed?

The dataset will be distributed upon acceptance of our paper.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

Our dataset only requires agreement to the Ego4D license. We publish our work under the CC 4.0 License.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

The only restriction is the Ego4D license.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

There are no additional regulatory restrictions beyond the Ego4D license.

Any other comments?

No.

Maintenance

Who is supporting/hosting/maintaining the dataset?

Our annotations will be available to download from GitHub. The official Ego4D videos are maintained by Meta AI.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The dataset curator can be contacted at keshik [at] stanford [dot] edu

Is there an erratum?

There is no erratum. If a correction is made, the datasheet will be corrected.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

Yes, we plan to update our dataset with new instances and other possible future corrections. We release v1.0 of our dataset with this paper, and plan to have future numbered versions of the dataset.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?

There are no limits on the retention of the data associated with our dataset.

Will older versions of the dataset continue to be supported/hosted/maintained?

Yes, if revision of the dataset are published then the older versions will be available to download from GitHub.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

Additions to the dataset are preferred to use standard open-source tools such as GitHub pull requests or creating forks for significantly large changes. We provide the necessary information, code, and data files to effectively retrace our process to create the dataset thus allowing for external contributions to every aspect of our dataset.

Any other comments?

No.