EgoSchema Datasheet

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

EgoSchema is a diagnostic benchmark for assessing very long-form video-language understanding capabilities of modern multimodal systems. While some prior works have proposed video datasets with long clip lengths, we posit that merely the length of the video clip does not truly capture the temporal difficulty of the video task that is being considered. To remedy this, we introduce temporal certificate sets, a general notion for capturing the intrinsic temporal understanding length associated with a broad range of video understanding tasks & datasets. Please see Section 3.2 in the main paper for more details.

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

The authors created the dataset within the Malik Group at Berkeley AI Research, UC Berkeley. The authors created it for the public at large without reference to any particular organization or institution.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

EgoSchema creation is funded by the ONR MURI award number N00014-21-1-2801

Any other comments?

No

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Each instance in the dataset represents a 3-minute video and text that contains a question and five answer options.

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

EgoSchema has a total of 5063 instances each containing one video, one question, and five answer options. You can see further statistics on the whole data on our website EgoSchema.github.io.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The video component of our dataset derives from the broader Ego4D dataset. For our research, we selectively extracted non-overlapping three-minute segments from the Ego4D video data, each segment consisting of a minimum of 30 human-annotated narrations (where each narration refers to a timestamped sentence). Detailed statistic of the number of viable clips for different possible length and narration density choices is discussed in Supplementary Section 7. The selected subset is very diverse in human behavior as can be seen by the activity statistics presented on EgoSchema.github.io.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance in our dataset comprises raw mp4 video data, captured at a rate of 30 frames per second and with a high resolution. Accompanying this video data, there are six text elements - one question and five corresponding answer options one of which is marked as the correct answer to the question.

Is there a label or target associated with each instance? If so, please provide a description.

Each instance is associated with a label ranging from 1 to 5 that indicates which of the five answer options is correct.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g. because it was unavailable). This does not include intentionally removed information but might include, e.g., redacted text.

All instances are complete.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

Some instances may have the same video but different questions and answers. It will be indicated by a clip unique identifier in the final dataset.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

EgoSchema is designed specifically for zero-shot testing. Its primary purpose is to be able to asses the out of the box long-term video-language understanding capabilities of modern multimodal models.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

The dataset was very carefully manually curated to mitigate any incidence of errors within the questions and answers. Although different questions may be posed for the same clip, it is ensured that there is no overlap between any two distinct clips. Further related details are also discussed in the limitations section in the main paper.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

Entirety of the dataset will be made publicly available at our project website EgoSchema.github.io. We will also provide a download tool for preprocessing all the videos such as cutting clips, associating the question/answer text etc. Text will be released in a JSON format, hosted on our github repository. EgoSchema will be publicly released under the Ego4D license, which allows public use of the video and text data for both research and commercial purposes.

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)? If so, please provide a description.

No

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Some videos do contain people. However, the Ego4D authors employed an array of de-identification procedures primarily centered on ensuring a controlled environment with informed consent from all participants, and, where applicable, in public spaces with faces and other personally identifiable information suitably obscured. We strictly import all RGB information from Ego4D without any addition of our own.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No

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? If so, please describe how.

No, Ego4D has employed an array of deidentification procedures in order to obscure any personally identifiable information such as people's faces.

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)? If so, please provide a description.

No

Any other comments?

No

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The video data, which is directly observable, was procured from the publicly accessible Ego4D dataset. In contrast, the text data was generated through the use of Large Language Models (LLMs) including GPT4, BARD, and Claude. These LLMs employed visual narrations from each video within the Ego4D dataset to generate the corresponding text.

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

The video and narration data were downloaded in accordance with the official Ego4D guidelines for data access: https://ego4d-data.org/docs/start-here/download-data. For the generation of the text data within our dataset, we utilized API access for GPT4 via OpenAI, for BARD via Google, and for Claude via Anthropic. This allowed us to generate three distinct questions for each video clip sampled from the Ego4D dataset. Upon the generation of these questions for each sampled video clip, we implemented a series of filtering procedures including Rule-based filtering, Blind filtering, and Manual curation. See Section 3.1.2 in the main paper for a more detailed explanation.

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

The video component of our dataset derives from the broader Ego4D dataset. For our research, we selectively extracted non-overlapping three-minute segments from the Ego4D video data, each segment consisting of a minimum of 30 human-annotated narrations (where each narration refers to a timestamped sentence). Detailed statistic of the number of viable clips for different possible length and narration density choices is discussed in Supplementary Section 7.

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)?

Our research utilized the services of Quantigo, a specialized data labelling company. The teams of Quantigo employees that were based in Bangladesh were compensated at a rate of 5 dollars per hour, at a wage significantly higher than the market hourly rate in Bangladesh. This was done to ensure fair compensation for the complex tasks performed while also contributing to the highest quality of the work delivered. It's important to note that our collaboration with Quantigo followed ethical guidelines, with the fair treatment of all employees involved and the appropriate respect for their expertise and labor. For exact instructions for human curation, see Supplementary Section 11.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The original videos within the Ego4D dataset were collected across various occasions spanning from 2019 to 2021. As for the EgoSchema, the textual information was collected over several sprints during the first half of 2023 based on the Ego4D narrations.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes

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

The video and narration data were acquired in accordance with the official Ego4D guidelines for data access: https://ego4d-data.org/docs/start-here/download-data. The Ego4D authors had in turn ensured consent of the people involved.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Ego4d paper followed several procedures to ensure the preservation of privacy and the upholding of ethical standards. Notably, these procedures included obtaining informed consent from those wearing the cameras and adhering to de-identification requirements for personally identifiable information (PII). Given that the video collection was conducted by Ego4D, we are not in a position to provide specific instructions that were given to the camera wearers. The Ego4D privacy statement is available at https://ego4d-data.org/pdfs/Ego4D-Privacy-and-ethics-consortium-statement.pdf

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Ego4d paper privacy procedures have included obtaining informed consent from those wearing the cameras. Given that the video collection was conducted by Ego4D, we are not in a position to provide specific instructions that were given to the camera wearers. See Ego4D privacy statement.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Ego4d paper privacy procedures have included allowing camera users to ask questions and withdraw at any time. Additionally, they were free to review and redact their own video. Given that the video collection was conducted by Ego4D, we are not in a position to provide specific instructions that were given to the camera wearers. You can find the Ego4D privacy statement at https://ego4d-data.org/pdfs/Ego4D-Privacy-and-ethics-consortium-statement.pdf.

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? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

While we recognize the importance of this topic, we would, once more, refer to the Ego4D paper for an in-depth discussion. Ego4D acknowledges the potential privacy risks associated with the use of wearable devices in data collection and has taken several steps such as depersonalizing any sensitive information, blurring out faces and bodies, etc. towards maintaining privacy. The same carries over to the video data in EgoSchema as well. Broadly, very long-form video understanding is a core capability for agents that are to perceive the natural visual world. Hence, developing datasets such as EgoSchema will be critical to unlocking this key AI capability. Additionally, according to Ego4D privacy statement, all videos from Ego4D were reviewed by an approved member of one of the participant's universities or institutes to identify and assess potential privacy concerns.

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)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

The set of generated questions and answers from output was filtered by those LLMs and finally curated by humans. A detailed description can be found in Section 3. There was no preprocessing done on the video clips sampled from Ego4D.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

Human curation was employed to rectify errors in the question-answer sets, particularly cases where the identified correct answer was wrong or a wrong answer was actually correct. Given the crucial role of this step in ensuring the accuracy of our dataset, we do not find it necessary to release a version of the dataset prior to human curation. However, all the discarded "raw" data is indeed also saved.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

The APIs for the Large Language Models (LLMs) are publicly accessible. The prompts for filtering and instructions for human curation are provided in Supplementary Section 6 and Supplementary Section 11 respectively. Additionally all necessary code for generation, filtering etc. is provided in the supplementary materials.

Any other comments?

No

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

We have used our dataset as a benchmark for several video question-answering models (please see §4.2 for more details). Additionally, we will be providing the code for zero-shot evaluation on our project github repository.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

It will be made public on our website once more papers will start to use our dataset.

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

Besides multiple-choice video question-answering tasks our dataset also can be used for open-ended video question-answering. There might be several more video-language tasks that could also be perhaps benchmarked, but currently we leave that exploration for future work.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

No

Are there tasks for which the dataset should not be used? If so, please provide a description.

No

Any other comments? No

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? If so, please provide a description.

The dataset will be made publicly available and can be used for both research and commercial purposes under the Ego4D license.

How will the dataset be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

The dataset will be distributed as a JSON file describing the unique identifier for each clip, the associated question, the five answer options, the label, and additional clip information that facilitates the tracing of the clip back to the original Ego4D data, such as the Ego4D video identification of the clip's source video, among other details. In addition, download tools to acquire and pre-process the video RGB data will also be provided on our website.

When will the dataset be distributed?

The full dataset will be made available upon the acceptance of the paper before the camera-ready deadline.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

EgoSchema will be publicly released under the Ego4D license, which allows direct public use of the video and text data for both research and commercial purposes.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No

Any other comments?

No

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The authors of the paper will be maintaining the dataset, pointers to which will be hosted on github repo https://github.com/egoschema/EgoSchema along with the code for download and preprocessing tool, with the actual data hosted either on Amazon AWS as an S3 bucket or as a google drive folder.

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

We will post the contact information on our website. We will be available through github issues as well as through email.

Is there an erratum? If so, please provide a link or other access point.

We will host an erratum on the Github repo in the future, to host any approved errata suggested by the authors or the video research community.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Yes, we plan to host an erratum publicly. There are no specific plans for a v2 version, but there does seem plenty oppurtunities for exciting future dataset work based on EgoSchema.

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)? If so, please describe these limits and explain how they will be enforced.

No.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

N/A There are no older versions at the current moment. All updates regarding the current version will be communicated via our website.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Contributions will be made possible using standard open-source tools, submitted as pull requests to the relevant GitHub repository. Moreover, we will provide information on how to trace sampled clips back to their original source within the Ego4D dataset. This will enable users to access additional Ego4D data, such as narrations, summaries, and object detections, as applicable.

Any other comments?

No