We answered the questions on the Datasheets for Datasets (https://arxiv.org/pdf/1803.09010.pdf) to further describe our new benchmark.

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.
Videos with real humans have various ethical, privacy, and copyright related costs and concerns. As an alternative, we conduct pre-training with a mix of real videos with humans removed and synthetic videos. Our goal is to reduce the gap between representations learned with and without data containing real humans by combining these two privacy-preserving data sources.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
The dataset was created by Howard Zhong, Samarth Mishra, Donghyun Kim, SouYoung Jin, Leonid Karlinsky, Hilde Kuehne, Rameswar Panda, Venkatesh Saligrama, Aude Oliva, and Rogerio Feris at MIT-IBM Watson AI Lab.

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.
N/A

Any other comments?
None

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.
The pre-training dataset in our benchmark consists of No-Human Kinetics and Synthetic dataset. The No-Human Kinetics dataset consists of real videos from the

Kinetics dataset where the humans are removed by inpainting, followed by the approach in the HAT paper. We use the Synthetic dataset collected in SynAPT, which is from three different pre-existing datasets/generators: ElderSim, PHAV, and SURREACT.

How many instances are there in total (of each type, if appropriate)?
The No-Human Kinetics dataset consists of 150 classes, each containing up to 1,000 examples. The Synthetic data also contains 150 classes, each containing up to 1,000 examples. We followed SynAPT for the selection of the classes.

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).

No-Human Kinetics: We remove humans from a randomly selected sample of 150 classes of Kinetics-400, which has 400 classes.
Synthetic Data: We used the synthetic data from SynAPT.

What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.
We provide a link to download the original Kinetics dataset and the detailed instructions to remove real humans from the videos with the inpainting approach proposed by HAT. We also provide a link to the SynAPT benchmark to download the synthetic data.

Is there a label or target associated with each instance? If so, please provide a description.
For No-Human Kinetics, the label is one of 150 human action classes. For Synthetic data, the label is in a different set of 150 human action classes.

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.
N/A

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.
N/A

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.
For No-Human Kinetics, we followed the data splits of Kinetics. For synthetic data, we followed the data splits generated by SynApt.

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

In No-Human Kinetics, there are some frames where humans are not fully removed. We found this occurs most often when the human is moving fast or if there is a crowd of humans.

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 dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

No-Human Kinetics: The user must download the original Kinetics dataset. After the dataset is downloaded, the user must run the HAT framework to remove the humans from the videos. Kinetics provides links to YouTube videos, so this dataset is constant unless the videos are taken down. Similarly, unless the HAT github is taken down, users can access the specific commit required to conduct the inpainting. We have an archive of the relevant code and model checkpoints in HAT on our server.

Synthetic Data: The dataset is released in the SynAPT github. Please refer to that for licenses and terms of use.

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' nonpublic communications)?
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.
Although some human actions might come off as offensive (e.g. kicking, punching), we remove humans from the video so that is no longer an issue.

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-Human Kinetics: It is only possible in the case where the background has identifying information of the human (eg. interior of house, text in background), or if the inpainting model did not fully remove the human.

Synthetic Data: No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race 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?
None.

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 the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
No-Human Kinetics: The unprocessed data released through the Kinetics dataset was originally from YouTube videos of real humans. We run the HAT framework to remove the humans from each frame of the video to create the No-Human Kinetics dataset.
Synthetic Data: The data was generated by computer graphics simulators.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?
No-Human Kinetics: Starting from the Kinetics dataset, we use the HAT framework to remove humans. The two models in the HAT framework are the SeMask segmentation model and the E2GFVI inpainting model. These models have been trained and validated on standard segmentation or inpainting benchmarks.

Synthetic Data: ElderSim, SURREACT, and PHAV computer graphics simulators were used to produce the data.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
For both No-Human Kinetics and the Synthetic dataset, please refer to the aforementioned details regarding the sampling strategy.

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)?
As mentioned, the data was sampled from publicly available assets. We use open source models to remove the humans in No-Human Kinetics.

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 underlying real data, Kinetics, was collected from YouTube in 2017.

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.
N/A, the No-Human Kinetics have humans removed. The synthetic videos for SynAPT do not involve any human subject.

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

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.

N/A, we did not collect data as the Kinetics data is publicly available. Instead, we processed this data by removing real humans from the data.

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.
N/A

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).
N/A

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.
N/A

Any other comments?
Our goal is to curate a safer pre-training dataset. While we cannot guarantee every human is removed, for the most part, the human is no longer identifiable.

Processing/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 remaining questions in this section.
No-Human Kinetics: We removed humans through inpainting from the Kinetics dataset using the HAT framework.

Synthetic Data: No preprocessing and cleaning were necessary as we drew the data from SynAPT.

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.
No-Human Kinetics: The raw data was the Kinetics dataset.
https://www.deepmind.com/open-source/kinetics

Synthetic Data: Raw data can still be downloaded from the public assets it was sourced from.

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point. N/A. Any other comments?
No-Human Kinetics: https://github.com/princetonvisualai/HAT
Synthetic Data: None.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.
We’ve used a combination of No-Human Kinetics and Synthetic data to learn privacy-preserving action recognition representations. Specifically, we evaluate the transferability of representations on the downstream tasks of the SynAPT benchmark.

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.
N/A.

What (other) tasks could the dataset be used for?
Different ratios of No-Human Kinetics and Synthetic data during pre-training can make the model depend more on contextual features or temporal action features.

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 dataset consumer 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 risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?
One of the main purposes of the No-Human Kinetics and Synthetic data is to overcome the ethics, privacy, copyright concerns, and other concerns of existing real video datasets. As of now, we do not see these issues impacting future uses.

Are there tasks for which the dataset should not be used? If so, please provide a description.
Please adhere to the original terms of use of the three synthetic video assets as well as the agreements for the Kinetics dataset and HAT framework.

Any other comments?
None.

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.
Yes, documentation on how to recreate the No-Human Kinetics dataset and the synthetic dataset is released.

How will the dataset be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
The documentation and instructions will be available in a GitHub repository.

When will the dataset be distributed?
The dataset is released in 2023.

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.

Please agree to the following ToUs:

- Kinetics
  The Kinetics dataset is licensed by Google Inc. under a Creative Commons Attribution 4.0 International License.
- HAT
  https://github.com/princetonvisualai/HAT/blob/main/LICENSE
- ElderSim:
  https://drive.google.com/drive/folders/1HomRAWYLiW_dREcwIE_tQUP7qRyV5ZMj
- PHAV:
  https://academictorrents.com/details/7a8b49530d40331d4fbdf0511844d52996683196
- SURREACT:
  https://www.di.ens.fr/willow/research/surreal/data/requestaccess.php

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.

Users must follow the licenses for Kinetics and HAT and sign the ToU for ElderSim, PHAV, and SURREACT.

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?

None.

Maintenance

Who will be supporting/hosting/maintaining the dataset?
The original asset curators will host the assets. Howard Zhong et al. will maintain the benchmark's GitHub to preprocess the data.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
Howard Zhong can be contacted via howardzh@mit.edu.

Is there an erratum? If so, please provide a link or other access point.
No.

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 dataset consumers (e.g., mailing list, GitHub)?
This will be posted on the GitHub.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the 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.
N/A.

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 dataset consumers.
Older versions can be viewed on the GitHub.

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 dataset consumers? If so, please provide a description.
Others may do so and should contact the original authors regarding any fixes or extensions.

Any other comments?
None.

References:

SynAPT: Kim, Y. W., Mishra, S., Jin, S., Panda, R., Kuehne, H., Karlinsky, L., ... & Feris, R. (2022). How Transferable are Video Representations Based on Synthetic Data?. *Advances in Neural Information Processing Systems,* 35, 35710-35723.

HAT: Chung, J., Wu, Y., & Russakovsky, O. (2022). Enabling Detailed Action Recognition Evaluation Through Video Dataset Augmentation. *Advances in Neural Information Processing Systems,* 35, 39020-39033.

Kinetics: Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., ... & Zisserman, A. (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950.*

ElderSim: Hwang, H., Jang, C., Park, G., Cho, J., & Kim, I. J. (2021). Eldersim: A synthetic data generation platform for human action recognition in eldercare applications. *IEEE Access.*

PHAV: de Souza, C. R., Gaidon, A., Cabon, Y., Murray, N., & López, A. M. (2020). Generating human action videos by coupling 3D game engines and probabilistic graphical models. *International Journal of Computer Vision*, 128(5), 1505-1536.

SURREACT: Varol, G., Laptev, I., Schmid, C., & Zisserman, A. (2021). Synthetic humans for action recognition from unseen viewpoints. *International Journal of Computer Vision*, 129(7), 2264-2287.