Multimodal Datasets and Benchmarks for Reasoning about Dynamic Spatio-Temporality in Everyday Environments

Takanori Ugai
Fujitsu Limited.
4-1-1 Kamikotanaka Nakaharaku Kawasaki Kanagawa, 211-8588, Japan
ugai@fujitsu.com
National Institute of Advanced Industrial Science and Technology
2-3-26, Aomi, Koto-ku, Tokyo 135-0064, Japan

Kensho Hara, Shusaku Egami, Ken Fukuda
National Institute of Advanced Industrial Science and Technology
{kensho.hara, s-egami, ken.fukuda}@aist.go.jp

Abstract

We used a 3D simulator to create artificial video data with standardized annotations, aiming to aid in the development of Embodied AI. Our question answering (QA) dataset measures the extent to which a robot can understand human behavior and the environment in a home setting. Preliminary experiments suggest our dataset is useful in measuring AI's comprehension of daily life.

1. Introduction

As Embodied AI continues to develop, understanding the time and place of actions in daily life becomes increasingly important [1–3, 14, 15]. Datasets and benchmarks have been created to support their development, and challenges have been presented [4, 6, 12].

Most of this data consists of recorded images of everyday life, annotations, and descriptions. The annotations were performed manually and were imprecise; not everything in the room was annotated. The behavior of what the person tries to do needs to be fully described in these descriptions.

Nishimura et al. [13] proposed PrimitiveActionOntology[1] to abstract activity labels in recognition datasets based on HomeOntology [19] and International Classification of Functioning, Disability and Health (ICF)[2]. They also proposed a HomeObjectOntology[3] based on VirtualHome assets, objects defined in Charades [17], and objects that occurred in the videos in the video archive called Elderly Behavior Library[4].

We created artificial video data (**MMDL**: Multimodal Dataset of Daily Life) using a 3D VirtualHome-AIST [18] simulator, which is based on VirtualHome [16] and, using VirtualHome2KG [5], created data describing what it is and where it is located for more objects. These data also clarify what the data are from the scripts that are placed in the simulator to make the avatar work. The annotations are mechanically generated with a standard vocabulary based on PrimitiveActionOntology and HomeOntology, which contributes significantly to the development of Embodied AI as they are consistent and free of contradictions.

We also created a question answering (QA) dataset (**MMQADL**: Multimodal Question Answering Dataset of Daily Life) to measure the extent to which the robot could understand a person's daily life from a video. We offer various types of descriptive and quantitative questions for question answering (QA) to gather information on location, action, object, time, and more. We also provide location-selective and descriptive QA examples for training and evaluation data.

This paper presents the findings of initial experiments conducted using two generative AIs, namely Video-LLaVa [10] and Google's Gemini 1.5 Pro Vision. These AIs were fed with a combination of images, natural language sentences, and QA that we created. The purpose of this experiment was to investigate AI's understanding of human behavior in a home environment. The results of the experiment indicate that our dataset is useful in measuring the AI's comprehension of human behavior and the surround-

[1] https://github.com/aistairc/PrimitiveActionOntology

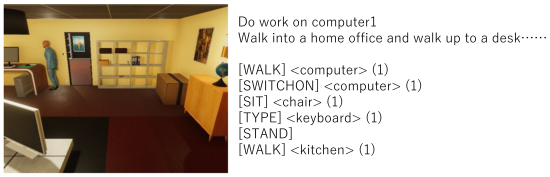
[2] https://www.who.int/standards/classifications/international-classification-of-functioning-disability-and-health

[3] https://github.com/aistairc/HomeObjectOntology

[4] https://www.behavior-library-meti.com/behaviorLib/homes/about

ing environment in a home.

2. MMDL: Simulation movie and detailed annotation



```
Do work on computer1
Walk into a home office and walk up to a desk……

[WALK] <computer> (1)
[SWITCHON] <computer> (1)
[SIT] <chair> (1)
[TYPE] <keyboard> (1)
[STAND]
[WALK] <kitchen> (1)
```

Figure 1. Example of video snapshot and action script

Figure 1 shows an action script titled “Do work on computer” and a snapshot of the video generated from it in VirtualHome-AIST. The first line of the action script is the title, the second line is the description, and the fourth and subsequent lines are the rows of the avatar’s behavior. There are 3,530 different videos, each of which shows a short chunk of behavior, called an activity, of approximately 30 seconds to a minute in length. They were generated from 706 scenarios (action scripts); for one scenario, five videos were generated with different camera positions. The characters’ behavior (avatars) in the videos and the 3D coordinates of approximately 400 objects in the house were annotated as data using VirtualHome2KG. The states of lights and other electrical appliances, such as on/off, and the opening/closing of the fridge and room doors, were also recorded [18], and the 2D positions of the camera image were also annotated for the objects with which the avatar was involved. 2D annotation is provided in the same scene graph format as Action Genome [8].

3. MMQADL: QA dataset for measuring daily life understanding

Listing 1. Example of Question and Answer

```
Q: Where is the man 10 seconds later from the
beginning of the video?
A1: Livingroom
A2: Bedroom
A3: Kitchen
A4: Bathroom
```

QA can pose different types of questions. They can be a choice (Listing 1) or a simple ”yes” or ”no” answer. The questions were designed to gather information about the location, action, object, time, and combination of topics being discussed based on TempCompass [11] and MVBench [9]. In addition, there are questions that focus on the appropriate caption for a video, which can be either short or long.

These questions were classified into two types: descriptive and quantitative. Descriptive questions were used to obtain factual information or details about the topic, event, object, or situation. They typically start with words like “what,” “does,” “where,” and “when.” The quantitative questions were designed to obtain numerical or quantitative data. They typically start with words like “how many,” “how much,” “how long,” and “how often.”

The data were designed to provide answers to 70 types of questions for longer time frames, defined in columns of three to seven activities, and provided in JSON format. The QA data were divided into two parts: learning (80%) and evaluation (20%) data, both containing answers.

The training data not only lacks answers but also provides annotated data with missing locations, actions, and objects that correspond to the answers. Additionally, the questions were categorized as Easy or Hard. Easy questions have only two options, while hard questions have around 30 options for actions, and all objects (about 200) that exist in the house are considered candidates for objects.

Table 1. Score of Precision

	action	location	object	time	caption
Gemini	0.7	0.9	0.4	0.5	0.8
Video-LLaVa	0.5	0.4	0.25	0.1	0.6

4. Preliminary Experiment

In the Knowledge Graph Reasoning Challenge 2024, one of the strategies [7] is to complete the annotations provided in the Knowledge Graph from the video. If we can accurately fill in all the missing parts of the annotation, we can answer all the questions correctly. We have used video clips and questions about the missing actions, locations, objects, and time as input. We tested the Large Language Models to answer the questions, and Table 1 is the result of our experiment. Overall, Gemini performs well. In particular, the distinction between the four types of rooms is mostly accurate. Video-LLaVa, on the other hand, does not understand the time elapsed in the video.

5. Summary

This article discusses the creation of a dataset that supports the development of Embodied AI. The dataset includes artificial movie data and a QA dataset to measure the AI’s comprehension of human behavior in a home environment. The results of the initial experiments show that the dataset is useful for measuring the AI’s understanding of human behavior and the surrounding environment in a home. We are planning to organise a technology contest (Challenge) in the future. All data is publicly available from https://github.com/KGRC4SI/DataSet

Acknowledgement

This paper is based on results obtained from projects, JPNP20006 and JPNP180013, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*, 2022. 1

[2] Valts Blukis, Chris Paxton, Dieter Fox, Animesh Garg, and Yoav Artzi. A persistent spatial semantic representation for high-level natural language instruction execution. In *Proceedings of the 5th Conference on Robot Learning*, pages 706–717. PMLR, 2022.

[3] Hilary Davis, Michael Arnold, Martin R. Gibbs, and Bjorn Nansen. Time, technology, and the rhythms of daily life. In *2010 IEEE International Symposium on Technology and Society*, pages 475–479, 2010. 1

[4] Matt Deitke, Dhruv Batra, Yonatan Bisk, Tommaso Campari, Angel X. Chang, Devendra Singh Chaplot, Changan Chen, Claudia Pérez D'Arpino, Kiana Ehsani, Ali Farhadi, Li Fei-Fei, Anthony Francis, Chuang Gan, Kristen Grauman, David Hall, Winson Han, Unnat Jain, Aniruddha Kembhavi, Jacob Krantz, Stefan Lee, Chengshu Li, Sagnik Majumder, Oleksandr Maksymets, Roberto Martín-Martín, Roozbeh Mottaghi, Sonia Raychaudhuri, Mike Roberts, Silvio Savarese, Manolis Savva, Mohit Shridhar, Niko Sünderhauf, Andrew Szot, Ben Talbot, Joshua B. Tenenbaum, Jesse Thomason, Alexander Toshev, Joanne Truong, Luca Weihs, and Jiajun Wu. Retrospectives on the embodied ai workshop, 2022. 1

[5] Shusaku Egam, Takanori Ugai, Mikiko Oono, Koji Kitamura, and Ken Fukuda. Synthesizing event-centric knowledge graphs of daily activities using virtual space. *IEEE Access*, pages 1–1, 2023. 1

[6] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Koláĭ, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18973–18990, 2022. 1

[7] Tsukasa Hirano, Kengo Ozaki, and Takeshi Morita. Prediction of actions and objects through video analysis using stepwise prompt. In *2024 IEEE 18th International Conference on Semantic Computing (ICSC)*, pages 289–293, 2024. 2

[8] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, Washington, USA, 2020. 2

[9] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. *ArXiv*, abs/2311.17005, 2023. 2

[10] 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 arXiv:2311.10122*, 2023. 1

[11] 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?, 2024. 2

[12] Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models, 2023. 1

[13] Satoshi Nishimura, Shusaku Egami, Takanori Ugai, Mikiko Oono, Koji Kitamura, and Ken Fukuda. Ontologies of action and object in home environment towards injury prevention. *The 10th International Joint Conference on Knowledge Graphs*, 2021. 1

[14] Giuseppe Paolo, Jonas Gonzalez-Billandon, and Balázs Kégl. A call for embodied ai, 2024. 1

[15] Tess Posner and Li Fei-Fei. AI will change the world, so it's time to change AI. *Nature*, 588(7837):118–118, 2020. 1

[16] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8494–8502, 2018. 1

[17] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision – ECCV 2016*, pages 510–526, Cham, 2016. Springer International Publishing. 1

[18] Takanori Ugai, Shusaku Egami, Swe Nwe Nwe Htun, Kouji Kozaki, Takahiro Kawamura, and Ken Fukuda. Synthetic multimodal dataset for empowering safety and well-being in home environments, 2024. 1, 2

[19] Alexandros Vassiliades, Nick Bassiliades, Filippos Gouidis, and Theodore Patkos. A knowledge retrieval framework for household objects and actions with external knowledge. In *Semantic Systems. In the Era of Knowledge Graphs*, pages 36–52, Cham, 2020. Springer International Publishing. 1