



## UvA-DARE (Digital Academic Repository)

### The CASTLE 2024 Dataset: Advancing the Art of Multimodal Understanding

Rossetto, Luca; Bailer, Werner; Dang-Nguyen, Duc-Tien; Healy, Graham; Jónsson, Björn Þór; Kongmeesub, Onanong; Le, Hoang-Bao; Rudinac, Stevan; Schöffmann, Klaus; Spiess, Florian; Tran, Allie; Tran, Minh-Triet; Tran, Quang-Linh; Gurrin, Cathal

**DOI**

[10.1145/3746027.3758199](https://doi.org/10.1145/3746027.3758199)

**Publication date**

2025

**Document Version**

Final published version

**Published in**

MM '25

**License**

CC BY

[Link to publication](#)

**Citation for published version (APA):**

Rossetto, L., Bailer, W., Dang-Nguyen, D.-T., Healy, G., Jónsson, B. Þ., Kongmeesub, O., Le, H.-B., Rudinac, S., Schöffmann, K., Spiess, F., Tran, A., Tran, M.-T., Tran, Q.-L., & Gurrin, C. (2025). The CASTLE 2024 Dataset: Advancing the Art of Multimodal Understanding. In *MM '25: Proceedings of the 33rd ACM International Conference on Multimedia : October 27-31, 2025, Dublin Ireland* (pp. 12629–12635). Association for Computing Machinery. <https://doi.org/10.1145/3746027.3758199>

**General rights**

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

**Disclaimer/Complaints regulations**

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.  
*UvA-DARE is a service provided by the library of the University of Amsterdam (<https://dare.uva.nl>)*



# The CASTLE 2024 Dataset: Advancing the Art of Multimodal Understanding

Luca Rossetto Dublin City University Dublin, Ireland	Werner Bailer JOANNEUM RESEARCH Graz, Austria	Duc-Tien Dang-Nguyen University of Bergen Bergen, Norway	Graham Healy Dublin City University Dublin, Ireland	Björn Þór Jónsson Reykjavik University Reykjavík, Iceland
Onanong Kongmeesub Dublin City University Dublin, Ireland	Hoang-Bao Le Dublin City University Dublin, Ireland	Stevan Rudinac University of Amsterdam Amsterdam Netherlands	Klaus Schöffmann Klagenfurt University Klagenfurt, Austria	Florian Spiess University of Basel Basel, Switzerland
Allie Tran Dublin City University Dublin, Ireland	Minh-Triet Tran VNU Ho Chi Minh University of Science Ho Chi Minh City Vietnam	Quang-Linh Tran Dublin City University Dublin, Ireland	Cathal Gurrin Dublin City University Dublin, Ireland	

## Abstract

Egocentric video has seen increased interest in recent years, as it is used in a range of areas. However, most existing datasets are limited to a single perspective. In this paper, we present the CASTLE 2024 dataset, a multimodal collection containing ego- and exo-centric (i.e., first- and third-person perspective) video and audio from 15 time-aligned sources, as well as other sensor streams and auxiliary data. The dataset was recorded by volunteer participants over four days in a common location and includes the point of view of 10 participants, with an additional 5 fixed cameras providing an exocentric perspective. The entire dataset contains over 600 hours of UHD video recorded at 50 frames per second. In contrast to other datasets, CASTLE 2024 does not contain any partial censoring, such as blurred faces or distorted audio. The dataset is available via <https://castle-dataset.github.io/>.

## CCS Concepts

• **Computing methodologies** → *Computer vision*; • **Information systems** → *Multimedia and multimodal retrieval*.

## Keywords

Dataset, Egocentric Vision, Multi-perspective Video, Lifelogging, Multimodal Understanding

### ACM Reference Format:

Luca Rossetto, Werner Bailer, Duc-Tien Dang-Nguyen, Graham Healy, Björn Þór Jónsson, Onanong Kongmeesub, Hoang-Bao Le, Stevan Rudinac, Klaus Schöffmann, Florian Spiess, Allie Tran, Minh-Triet Tran, Quang-Linh Tran, and Cathal Gurrin. 2025. The CASTLE 2024 Dataset: Advancing the Art of

Multimodal Understanding. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*, October 27–31, 2025, Dublin, Ireland. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3746027.3758199>

## 1 Introduction

Human interactions and everyday experiences are inherently complex, dynamic, and multifaceted. Understanding and analysing these interactions is critical for advancing research across numerous fields, including human-computer interaction, social dynamics, psychology, and linguistics. Although plenty of visual datasets capturing human activities have been created, many of them exhibit significant limitations. Third-person datasets, for example, often lack the subjective context crucial for interpreting human behaviour, while first-person datasets frequently limit either the recording duration or scope of activities. Multi-perspective datasets that combine first-person and third-person views are rare and typically include only a limited number of activities and do not last long enough to capture the full range of interactions and social dynamics characteristic of everyday life.

In this paper, we introduce the CASTLE 2024 dataset, a multimodal multi-perspective collection of ego-centric (first-person) and exo-centric (third-person) high-resolution video recordings, augmented with additional sensor streams, designed to capture the complexity of daily human experiences. The dataset captures the experience and daily interaction of ten volunteer participants over the course of four days. It shows a broad range of domestic and social activities, including *cooking*, *eating*, *cleaning*, *meeting* and *leisure activities*, capturing authentic interactions among participants. The main part of the dataset consists of time-aligned videos from 15 GoPro HERO10 cameras in UHD resolution (3840 × 2160 pixels) at 50 frames per second, capturing a total of over 600 hours of video and audio data. Ten cameras were worn by the participants, providing immersive ego-centric views, while five stationary



This work is licensed under a Creative Commons Attribution 4.0 International License. *MM '25, Dublin, Ireland*

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2035-2/2025/10

<https://doi.org/10.1145/3746027.3758199>



**Figure 1: Example from the dataset: six perspectives of a joint breakfast**

cameras offered broader contextual coverage with exo-centric perspectives. Figures 1, 2, and 3 show examples of different situations from multiple perspectives.

Additional metadata recorded by the cameras, such as inertial measurements (IMU) and GPS data is also included in the dataset, providing valuable context for detailed analysis. In addition, all participants were wearing heart rate monitors and took additional images and videos with additional recording devices, all of which are included as auxiliary data in the dataset. Moreover, the dataset is also uniquely self-documenting through the inclusion of workshop sessions, during which participants discussed the data generation process and planned further downstream applications. These workshop sessions were carried out in English and touched on a variety of related topics. Although most of the conversations recorded in the dataset are in English, the participants come from diverse ethnic and linguistic backgrounds, and the dataset also contains further conversations in the native languages of some of the participants, including German, Swiss German, and Vietnamese. Such multimodal and multilingual richness makes the dataset particularly interesting for studies in linguistics, social dynamics, and human-computer interaction.

The dataset is expected to be useful for a variety of multimedia analysis and understanding tasks and beyond, such as lifelog retrieval, object and action recognition (including understanding of long, complex and multi-person action sequences), social interaction analysis, and reconstruction and analysis of dynamic 3D scenes. Some of the auxiliary data can already provide metadata for these tasks, while others will require the creation of specific annotations. Furthermore, the combination of ego-centric and exo-centric perspectives can be used to study the interplay between individual and group activities, as well as the impact of different perspectives on the understanding of human activities and interactions. Given the rich and diverse nature of the dataset, it is also expected to be useful for the development and evaluation of new methods for multimodal data analysis, including methods for data fusion, multimodal retrieval, and multimodal summarization. By providing a large-scale dataset with a variety of perspectives and sensor streams, we aim to enable researchers to develop and evaluate new methods for understanding human activities and interactions in

everyday environments, as well as to explore new applications and use cases for multimodal data analysis.

The remainder of this paper is structured as follows: Section 2 discusses other datasets and their commonalities and differences to CASTLE 2024. Section 3 then presents an overview of the methods by which our dataset was constructed. Section 4 details the most important properties of the dataset. Section 5 discusses some possible downstream tasks that can be addressed with the dataset. Finally, Section 6 concludes the paper.

The dataset is available via <https://castle-dataset.github.io/> under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

## 2 Related Work

Visual data captured from a first-person perspective has received increased attention in recent years, and several collections have emerged that contain such data in various settings. An early example is the *LSC Dataset* [5], used for the annual Lifelog Search Challenge [11]. It contains a lifelog in the form of an egocentric image sequence representing the first-person view of a single person, covering multiple years of their life. The images were taken with a cadence of 2-3 images per minute, resulting in a sequence that can be considered a very low frame rate video.

Among egocentric video datasets, a prominent example is *EPIC Kitchens* [2], which contains 100 hours of video recorded using head mounted GoPro cameras. The dataset consists of unscripted activities recorded in 45 different kitchen environments. Similarly, *HD-EPIC* [8] contains 41 hours of egocentric video recorded in 9 kitchens. It augments the video with eyegaze data and provides full 3D reconstructions of the 9 kitchen environments.

*Ego4D* [3] substantially broadened the scope of egocentric video by collecting activity of daily living video with a combined duration of about 3670 hours. The dataset contains videos of different durations, captured by 931 different camera wearers at 74 locations in 9 different countries. *MultiEgoView* [6] extends the egocentric view beyond one camera by including recordings from six synchronized cameras worn at different locations on the body of the same person while performing different actions.



Figure 2: Example from the dataset: six perspectives of cooking activities

Other approaches for understanding scenes and human activities within them use multi-perspective third-person views. Examples of such datasets include *MM-Office* [14] which places four cameras and eight microphones at different locations in an office environment. *MEVA* [1] expands on this theme by compiling 9 300 hours of security camera footage from 38 RGB and thermal IR cameras from an access-controlled campus. The dataset features roughly 100 participants involved in predefined scenarios as well as the spontaneous background activities, recorded over three weeks.

A combination of egocentric and exocentric videos can be found in other datasets such as *Assembly101* [10], which contains multiple time-aligned videos of a single person performing specific assembly and disassembly tasks. The dataset consists of 513 hours of video recorded by four head-mounted cameras and eight fixed external cameras in a highly constrained environment. *Ego-Exo4D* [4] brings the combination of first- and third-person perspective video to a much broader range, providing 1286 total hours of video from a range of different activities in 123 different natural scene contexts. Individual videos are between 1 and 42 minutes long and are augmented with a range of additional sensor data, including multichannel audio, eye gaze, 3D point clouds, camera poses, and IMU data.

To go beyond single-person actions, multi-perspective multi-user recordings are necessary to capture the complexity of human interactions. Examples of data collections addressing this include *HoloAssist* [12], a large dataset including the recording of interactions between pairs of participants, collaboratively solving a set of predefined object-centric manipulation tasks. Every activity is jointly performed by two people with two distinct roles: the ‘performer’ wears a head-mounted recording device and performs a specific activity, while the ‘instructor’ watches and provides instructive information. The *Aria Everyday Activities Dataset* [7] also contains the perspective from multiple participants, but does not assign specific roles to them. It consists of 7.3 hours of video recorded by wearable cameras in 5 locations, with one to two wearers per location. The videos are augmented with eyegaze and IMU data.

While all of these datasets provide insights into human activities in multiple ways and some of them are augmented with multiple additional sensor streams, none of them provides a long-form

multi-perspective representation capturing the richness of everyday activity. The one dataset that is closest to ours in terms of addressing these challenges is the very recent *EgoLife* [13], which was created in a comparable setting to our dataset. It consists of recordings by 6 participants wearing recording glasses who lived together in the same location for 7 days. Their location was also equipped with 15 stationary cameras, which provided first- and third-person perspectives. The egocentric videos are recorded in a square  $1408 \times 1408$  pixel resolution at 20 frames per second and include audio, capturing conversations between the participants. The videos are augmented with additional eyegaze and IMU data, and the dataset contains additional information, such as a complete 3D reconstruction of the environment, as well as extensive annotations.

Compared to CASTLE 2024, *EgoLife* covers a longer time span (i.e. 7 vs. 4 days) and includes video from more stationary cameras (i.e. 15 vs 5). It also contains more explicit annotations of activities which have no direct equivalent in the initial release of CASTLE 2024. Our dataset does contain more egocentric perspectives (10 vs 6) and recorded video at a higher resolution and frame rate (UHD@50fps vs  $1408 \times 1408$  pixel @ 20fps). CASTLE 2024 also contains longer continuously recorded video and more video overall. In contrast to *EgoLife*, CASTLE 2024 does not blur faces nor partially anonymise any other content.

### 3 Data Collection

In the following, we discuss the data collection procedure.

#### 3.1 Recording Setup

The CASTLE 2024 dataset was recorded in early December 2024 in a vacation home at the west coast of Ireland, located in a remote area chosen to minimize interaction with third parties. The recording process included 12 volunteer participants, i.e., the authors of this paper<sup>1</sup> who had all previously agreed to share the data recorded during this time period. Most of the activities shown in the dataset took place in the common area of the house, schematically depicted in Figure 4, which includes a kitchen, a dining area, and various

<sup>1</sup>Two authors could not be physically present due to scheduling conflicts.



Figure 3: Example from the dataset: six perspectives of people playing a game

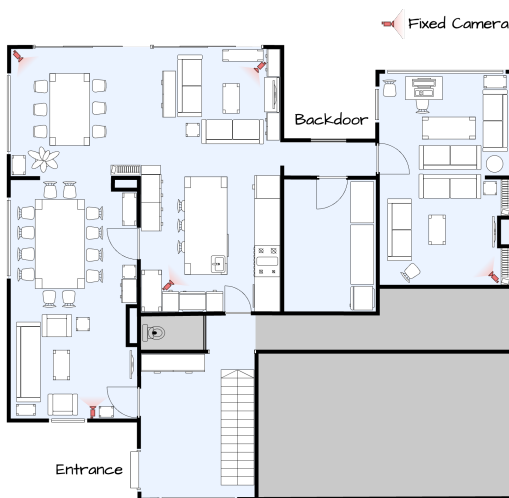


Figure 4: Schematic floor plan of the House (not to scale). Locations of static cameras are highlighted in red.

spaces for leisure activities. Individual bedrooms in the house, as well as a secondary house with additional bedrooms and a common space, served as camera-free zones for privacy. Some activities were also recorded outside the house, such as walks and visits to a nearby beach, including driving to and from the location. However, the majority of the data was recorded indoors due to the weather conditions at the time of the recording. The on-site team included ten active *data gatherers* and two *helpers*.<sup>2</sup>

Each data gatherer was wearing a head-mounted GoPro HERO10 Black, equipped with a 256GB SD-card and connected to an external 20,000mAh battery bank. The cameras were configured to record in UHD (3840 × 2160 pixels) resolution at 50 frames per second with minimal motion stabilization. Each data gatherer was also wearing a FitBit that continuously recorded the wearer’s heart rate.

Five stationary cameras of the same make and model as used by the data gatherers were placed at fixed locations in the house, as

<sup>2</sup>One helper acted as a data gatherer for the last day of the recording period, since another participant had to leave a day early.

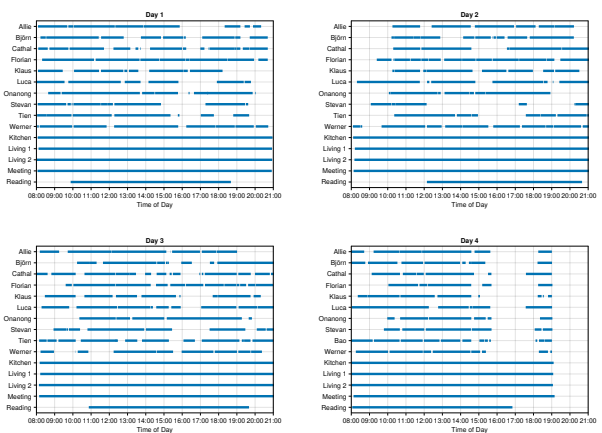


Figure 5: Daily coverage of the individual camera sources of the four days of the dataset

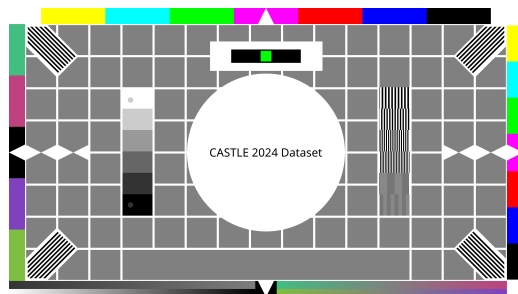
illustrated in Figure 4. These cameras were powered by an external mains-power adapter and equipped with a 512GB SD-card<sup>3</sup> to support continuous recording throughout the entire day. The cameras were configured to record in the same resolution and frame rate as the data gatherers’ cameras.

At the end of every day, SD cards and battery banks were collected for data copying and recharging. The FitBit devices were worn continuously by the data gatherers.

### 3.2 Activities

The participants were encouraged to participate in a variety of activities. A list of suggested activities was provided to participants, and was discussed, modified, and expanded on during the recording period. However, participants were free to participate in any activity they chose. Examples of such activities include cooking, eating, cleaning, reading, playing music instruments, painting and drawing, playing board games, watching television, and most importantly interacting with other participants.

<sup>3</sup>Except for the camera in the reading room, which had a 256GB SD-card.



**Figure 6: Placeholder used to pad videos when no recording data was available.**

Some social activities were planned in advance to ensure that multiple participants were present and interacting at the same time. For example, each day, the participants gathered for a workshop session to discuss the data collection process, plan further activities, and address any issues that arose during the recording. These sessions were recorded and are included in the dataset, acting as a form of self-documentation. Other social activities that were recorded include reading experiments conducted by one participant for a research project, where eyetracking data was collected using a Gazeport<sup>4</sup> GP3 HD eyetracker. ‘Happy Quiz’, as called by the participants, was also played most evenings, with a quizmaster asking questions to the other participants, who would press a buzzer to answer.

### 3.3 Privacy and Consent

One of the most important aspects of the dataset was that we did not wish to anonymise the participants, for example by blurring faces or distorting voices. Instead, we aimed to capture authentic experiences and interactions among the participants, including their facial expressions, body language, and tone of voice. This decision was made to preserve the richness and authenticity of the data, as well as to enable studies on social interactions, language use, and various other aspects of human behaviour, especially in the context of multimodal data analysis and understanding. However, we also wanted to ensure that the participants’ privacy was protected and that they felt comfortable and safe during the recording process.

Prior to their agreement to participate in the data collection, the participants were informed about the recording process, the intended use of the data, and the measures taken to protect their privacy—such as excluding bedrooms from the recording area and allowing them to deactivate the cameras at any time. Each participant signed a consent form detailing these aspects. Additionally, participants were given the opportunity to review the data recorded by their own camera and request the removal of specific segments, including conversations or activities they preferred to keep private, before the data was shared publicly.

Interactions with third parties were kept to a minimum and any third party instances that appeared in the recordings were removed to protect their privacy. When third parties were present, the helpers were responsible for ensuring that the cameras were deactivated to prevent recording as needed. Mostly, these occurred

<sup>4</sup><https://www.gazeport.com>

when some participants left the house and ventured onto the public road.

## 4 Dataset Properties

The CASTLE 2024 dataset consists of two sub-parts – the main part and the auxiliary part – totaling 8.22TB in size.

### 4.1 Main Part

All data in the main part is segmented into one-hour-long, time-aligned segments. It consists of 666 videos, including their audio streams, and additional files for the other aligned sensor data. All segments are exactly one hour long (i.e., 180 000 frames) and start on the hour, covering the time span from 8am to 9pm, if any recording is available for this time period. Figure 5 shows the ranges during which data was recorded for every stream. To ensure alignment, gaps in the recording are padded using a placeholder image based on a television broadcast test pattern,<sup>5</sup> as illustrated in Figure 6. Whenever there was no data available for the entire hour, it was omitted completely and replaced with an empty placeholder file to signify the absence of data. Next to the video files, other sensor streams are made available in CSV format with one file per data stream. For location data, an additional file in GPS Exchange Format (GPX) is added whenever positioning information was available. We also provide automatically generated transcriptions of all spoken dialog, generated using a *Whisper V3 Large* [9] model.

### 4.2 Auxiliary Part

The auxiliary part of the dataset contains additional data that is not structured in time-aligned one-hour chunks. This part contains the following:

- Heart-rate of all participants captured once per minute.
- Images and videos captured by the participants using their personal devices.
- Images captured at regular intervals by one participant using a Narrative Clip<sup>6</sup>
- Images captured sporadically using a thermal camera.
- Eyegaze data captured using a Gazeport GP3 HD during dedicated reading sessions for some participants.

## 5 Potential Applications

We believe that the CASTLE 2024 dataset has the potential to be used in a wide range of downstream applications. Understanding how researchers might leverage this dataset to address specific tasks provides valuable direction and encourages innovation. To this end, we outline three initial tasks included in a ACM Multimedia 2025 Grand Challenge: event instance search, object instance search, and question answering. These represent foundational tasks for advancing multimedia analytics and highlight the dataset’s strengths and diversity. Then, we discuss some further potential applications of the dataset.

<sup>5</sup><https://github.com/edent/SVGtestcard>

<sup>6</sup><https://getnarrative.com/> PoV wearable camera

## 5.1 Event Instance Search

*Motivation.* Being able to search for specific events in a large-scale multimodal dataset is a fundamental task in multimedia retrieval. In the context of the CASTLE 2024 dataset, event search can be used to find specific activities or interactions among the participants. For example, one might want to find all instances of a specific action, such as someone making coffee, or all instances of a specific interaction, such as someone telling a joke and the others laughing. Being able to search for such events can be the first step towards more complex tasks and applications.

*Task Definition.* Given a textual query describing an event in natural language, such as ‘someone making coffee’ or ‘someone telling a joke and the others laughing’, the task is to retrieve all video segments in the dataset that contain the relevant event. The events are to be identified by the time range and video ID. The task can be evaluated using standard information retrieval metrics, such as mean average precision (mAP) or recall at  $k$ .

## 5.2 Object Instance Search

*Motivation.* Object instance search is another fundamental task in multimedia retrieval, where the goal is to find all instances of a specific object in a large-scale multimodal dataset. This specific object can be described by a textual query or an image of the object. The CASTLE 2024 dataset contains a wide variety of objects, such as kitchen utensils, food items, books, and small decorative items which were moved around by the participants during the recording period. This task can be challenging due to the large number of objects in the dataset, the diversity of the scenes in which they appear, and the sheer volume of data.

*Task Definition.* Given a natural language query text describing an object, such as ‘a cookie cutter shaped like a star’, or a reference image of the object, the task is to retrieve all occurrences of that object in any of the video streams. Similar to event search, the task can be evaluated using standard information retrieval metrics.

## 5.3 Question Answering

*Motivation.* It is human nature to ask questions about the world around us. Recent advances in natural language processing have enabled the development of question answering systems that can answer questions about text, images, and videos. Video question answering is a challenging task that requires understanding of both the visual and textual content of the video. Participants in the CASTLE 2024 dataset engaged in a wide range of activities and interactions, providing a rich source of data for video question answering. The long-form nature of the videos and the multiple perspectives captured by the cameras make this task particularly challenging and interesting.

In contrast to how most video question answering tasks are defined, where a video is provided with a question, we aim to elevate the challenge to a more general format, similar to the question answering task in the Lifelog Search Challenge [11]. This means that the question is unbound to any specific video and can refer to anything that happened during the recording period and was captured in visual or audio channel of at least one camera. As

such, the task is more challenging as the answer must be found by searching through the entire dataset.

*Task Definition.* Given a question formulated in natural language, the task is to find an answer to the question. Answers are to be provided in natural language and include references to sensor streams and time intervals to provide evidence. The task can be evaluated using standard question answering metrics, such as accuracy or F1 score.

## 5.4 Other Potential Applications

Beyond the tasks outlined above, the richness of the CASTLE 2024 dataset can be useful for numerous other research directions. Its multi-perspective visual coverage is particularly well-suited for training and evaluating computer vision models for tasks such as activity recognition, object detection and multi-perspective object tracking, or scene reconstruction. Similarly, the synchronised multimodal data streams offer opportunities to develop advanced techniques for multimodal data analysis, including data fusion, cross-modal retrieval, multimedia forensics, and multimodal summarisation. The detailed capture of social interactions and activities can be a valuable resource for investigating social dynamics, language use, and the development of sophisticated human-computer interaction methodologies. This also applies to AI-based linguistic research, where the dataset’s multilingual conversational data can be used to develop and evaluate models for automatic speech recognition, multilingual language modelling, speaker identification, dialect analysis, and sentiment (or humour) detection across languages and conversational contexts.

## 6 Conclusion

In this paper, we presented the CASTLE 2024 dataset, a multimodal, multi-perspective collection comprising synchronised high-resolution ego-centric and exo-centric video recordings, complemented by additional sensor data, aimed at comprehensively capturing daily human experiences. This data set documents authentic interactions and diverse domestic and social activities – including cooking, eating, cleaning, and leisure activities – among ten volunteer participants over four days.

Although the initial release of the dataset does not include in-depth semantic annotations, we foresee that CASTLE 2024 might still prove useful in a range of applications. In particular, the presented dataset goes beyond currently available alternatives in terms of content diversity and concurrent points of view. It also increases authenticity and representativeness by forgoing any partial censoring of recordings.

## Acknowledgments

This work was partially supported by an ACM SIGMM Special Initiatives Grant. Individual contributors were further supported by: the Swiss National Science Foundation through project MediaGraph (contract no. 202125) and project “Participatory Knowledge Practices in Analog and Digital Image Archives” (contract no. 193788); by Research Ireland under grant number 13/RC/2106 P2; and by the Icelandic Research Fund grant 239772-051.

## References

- [1] Kellie Corona, Katie Osterdahl, Roderic Collins, and Anthony Hoogs. 2021. MEVA: A Large-Scale Multiview, Multimodal Video Dataset for Activity Detection. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3–8, 2021*. IEEE, 1059–1067. doi:10.1109/WACV48630.2021.00110
- [2] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2022. Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision (IJCV)* 130 (2022), 33–55. <https://doi.org/10.1007/s11263-021-01531-2>
- [3] 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, Abrahm Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolár, Satwik Kotler, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Kartikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Meredy Ramazanov, 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 A. 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. 2022. Ego4D: Around the World in 3, 000 Hours of Egocentric Video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18–24, 2022*. IEEE, 18973–18990. doi:10.1109/CVPR52688.2022.01842
- [4] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zachary Chavis, Joya Chen, Feng Cheng, Fugen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, María Escobar, Cristhian Forigua, Abrahm Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Dutt Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J. Liang, Jia-Wei Liu, Sagnik Majumder, Yongsun Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina González, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Meredy Ramazanov, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbeláez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard A. Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shout, and Michael Wray. 2024. Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16–22, 2024*. IEEE, 19383–19400. doi:10.1109/CVPR52733.2024.01834
- [5] Cathal Gurrin, Klaus Schoeffmann, Hideo Joho, Bernd Münzer, Rami Albalat, Frank Hopfgartner, Liting Zhou, and Duc-Tien Dang-Nguyen. 2019. A Test Collection for Interactive Lifelog Retrieval. In *MultiMedia Modeling - 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8–11, 2019, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 11295)*, Ioannis Kompatsiaris, Benoit Huet, Vasileios Mezaris, Cathal Gurrin, Wen-Huang Cheng, and Stefanos Vrochidis (Eds.). Springer, 312–324. doi:10.1007/978-3-030-05710-7\_26
- [6] Dominik Hollidt, Paul Strelci, Jiayi Jiang, Yasaman Haghighi, Changlin Qian, Xintong Liu, and Christian Holz. 2024. EgoSim: An Egocentric Multi-view Simulator and Real Dataset for Body-worn Cameras during Motion and Activity. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10–15, 2024*. Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakob M. Tomczak, and Cheng Zhang (Eds.). [http://papers.nips.cc/paper\\_files/paper/2024/hash/c1017d0a006d31dfbfd4cf1e9189d747-Abstract-Datasets\\_and\\_Benchmarks\\_Track.html](http://papers.nips.cc/paper_files/paper/2024/hash/c1017d0a006d31dfbfd4cf1e9189d747-Abstract-Datasets_and_Benchmarks_Track.html)
- [7] Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, Kiran Somasundaram, Luis Pesqueira, Mark Schwesinger, Omkar M. Parkhi, Qiao Gu, Renzo De Nardi, Shangyi Cheng, Steve Saarinen, Vijay Baiyya, Yuyang Zou, Richard A. Newcombe, Jakob Julian Engel, Xiaqing Pan, and Carl Yuheng Ren. 2024. Aria Everyday Activities Dataset. *CoRR* abs/2402.13349 (2024). doi:10.48550/ARXIV.2402.13349 arXiv:2402.13349
- [8] Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, Jacob Chalk, Zhifan Zhu, Rhodri Guerrier, Fahd Abdelazim, Bin Zhu, Davide Moltisanti, Michael Wray, Hazel Doughty, and Dima Damen. 2025. HD-EPIC: A Highly-Detailed Egocentric Video Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [9] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *International Conference on Machine Learning, ICML 2023, 23–29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 28492–28518. <https://proceedings.mlr.press/v202/radford23a.html>
- [10] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. 2022. Assembly101: A Large-Scale Multi-View Video Dataset for Understanding Procedural Activities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18–24, 2022*. IEEE, 21064–21074. doi:10.1109/CVPR52688.2022.02042
- [11] Ly-Duyen Tran, Manh-Duy Nguyen, Duc-Tien Dang-Nguyen, Silvan Heller, Florian Spiess, Jakub Lokoc, Ladislav Peska, Thao-Nhu Nguyen, Omar Shahbaz Khan, Aaron Duane, Björn Pór Jónsson, Luca Rossetto, An-Zi Yen, Ahmed Alateeq, Naushad Alam, Minh-Triet Tran, Graham Healy, Klaus Schoeffmann, and Cathal Gurrin. 2023. Comparing Interactive Retrieval Approaches at the Lifelog Search Challenge 2023. *IEEE Access* 11 (2023), 30982–30995. doi:10.1109/ACCESS.2023.3248284
- [12] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrisc, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. 2023. HoloAssist: an Egocentric Human Interaction Dataset for Interactive AI Assistants in the Real World. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1–6, 2023*. IEEE, 20213–20224. doi:10.1109/ICCV51070.2023.01854
- [13] Jingkan Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, Bo Li, and Ziwei Liu. 2025. EgoLife: Towards Egocentric Life Assistant. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [14] Masahiro Yasuda, Yasunori Ohishi, Shoichiro Saito, and Noboru Harada. 2022. Multi-View And Multi-Modal Event Detection Utilizing Transformer-Based Multi-Sensor Fusion. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23–27 May 2022*. IEEE, 4638–4642. doi:10.1109/ICASSP43922.2022.9746006