Jacob Chalk

https://jacobchalk.github.io jacob.chalk@bristol.ac.uk +44 (0)7899346230
https://www.linkedin.com/in/jacob-chalk https://scholar.google.com/citations?user=FPugmicAAAAJ
https://github.com/jacobchalk Bristol, UK

Research Focus

My research focuses on 4D video understanding, aiming to develop systems that can perceive and reason about dynamic 3D scenes over time, including problems such as long-term 3D object tracking. Previously, my work centred on leveraging multimodal data for egocentric video understanding, including audio-visual learning and prediction of object interactions using eye-gaze and 3D annotations.

Current Role

University of Bristol Bristol, UK
Research Associate *January 2026 – Current*
Conducting research on 4D Video Understanding.

Previous Roles

NAVER LABS Europe Grenoble, France
Industrial Internship - Visual Representation Learning Team *February 2025 – January 2026*
*Carried out research on long-term 3D multi-object tracking, supervised by **Diane Larlus.***

Academic Qualifications

University of Bristol Bristol, UK
PhD in Computer Vision *September 2021 – September 2025*
*Thesis Title: Leveraging Multimodal Data for Egocentric Video Understanding, supervised by **Prof. Dima Damen.***

University of Bristol Bristol, UK
MEng in Computer Science - First Class Honours *September 2017 – September 2021*
*Dissertation Title: Video GANs for Human-Object Interactions, supervised by **Prof. Dima Damen.***

Teaching

University of Bristol Bristol, UK
Teaching Assistant *September 2019 – September 2024*
Software Product Engineering (2nd Year Module), Games Project (3rd Year Module), Computer Graphics (3rd Year Module)
Image Processing and Computer Vision (3rd Year Module), Applied Deep Learning (4th Year Module)

Publications

- M. Hatano*, S. Sinha*, **J. Chalk**, W. Li, H. Saito, D. Damen, "Prime and Reach: Synthesising Body Motion for Gaze-Primed Object Reach" arXiv preprint arXiv:2512.16456, 2025. [Page] | [Paper]
- T. Perrett*, A. Darkhalil*, S. Sinha*, O. Emara*, S. Pollard*, K. Parida*, K. Liu*, P. Gatti*, S. Bansal*, K. Flanagan*, **J. Chalk***, Z. Zhu*, R. Guerrier*, F. Abdelazim*, B. Zhu, D. Moltisanti, M. Wray, H. Doughty, and D. Damen, "HD-EPIC: A Highly-Detailed Egocentric Video Dataset" in CVPR 2025. [Page] | [Paper]
- C. Plizzari, S. Goel, T. Perrett, **J. Chalk**, A Kanazawa, D Damen, "Spatial Cognition from Egocentric Video: Out of Sight, Not Out of Mind" in 3DV 2025. [Page] | [Paper]
- **J. Chalk***, J. Huh*, E. Kazakos, A. Zisserman, and D. Damen, "TIM: A Time Interval Machine for Audio-Visual Action Recognition" in CVPR 2024. [Page] | [Paper]
- J. Huh*, **J Chalk***, E. Kazakos, D. Damen, and A. Zisserman, "EPIC-SOUNDS: A Large-Scale Dataset of Actions that Sound." in ICASSP 2023 & TPAMI 2025. [Page] | [Paper]

*: Equal Contribution

Datasets

HD-EPIC *Released February 2025*
Highly Detailed Egocentric Video Dataset *Paper Published in CVPR 2025*

HD-EPIC is an extensive dataset manually annotated with highly detailed and interconnected ground-truth labels covering recipe steps, fine-grained actions, ingredients with nutritional values, moving objects, and audio annotations. All annotations are grounded in 3D through digital twinning of the scene, fixtures, object locations, and primed with gaze. Footage is collected from unscripted recordings in diverse home environments, making HD-EPIC the first dataset collected in-the-wild but with detailed

annotations matching those in controlled lab environments. The dataset consists of 41 hours of video in 9 kitchens with digital twins of 413 kitchen fixtures, capturing 69 recipes, 59K fine-grained actions, 51K audio events, 20K object movements and 37K object masks lifted to 3D. On average, there are 263 annotations per minute across the unscripted videos. HD-EPIC also presents a challenging VQA benchmark of 26K questions assessing the capability to recognise recipes, ingredients, nutrition, fine-grained actions, 3D perception, object motion, and gaze direction.

Roles:

- Gathered audio annotations matching a similar process to EPIC-Sounds (see below)
- Computed and visualised eye-gaze priming statistics for all 3D objects start and end locations. This involved projecting the eye-gaze direction ray into the 3D environment and measuring when the ray intersected with the objects 3D-bounding box before it was picked-up, or the end locations bounding box when placed down.
- Created eye-gaze related VQA questions, including *"What is the camera wearer currently looking at?"*, which involved finding the intersection between the projected gaze ray and 3D environment, as well as *"What will the person interact with next?"* which involved using objects whose start locations were primed with eye-gaze before being picked up to indicate it was about to be moved by the camera wearer.
- Ran LLM inference of Llama 3.2 (Vision-Instruct-90B) on the VQA questions

EPIC-Sounds *Released January 2023*

Audio-Centric Annotated Dataset *Paper Published in ICASSP 2023 & TPAMI 2025*

EPIC-Sounds is an audio-only dataset, gathered across 100 hours of untrimmed audios from the videos of EPIC-KITCHENS. This dataset was motivated by the observation that audio-visual networks tend to use both temporally and semantically identical annotations for both modalities, yet audio and video can significantly differ even when describing the same action.

Roles:

- Post-processed the labelled segments from annotators, this included:
  - Manually correcting typos in the free-form text descriptions.
  - Removing erroneous/redundant annotations.
  - Categorising the free-form descriptions into the 44 classes in the dataset.
  - Decomposing challenging segments containing multiple overlapping sounds into distinct annotations.
  - Training binary classifiers on the head classes of the dataset to clean the training set.
- Trained and ran quantitative analysis on the baseline models.
- Visualisation of the class distribution for the dataset.
- Ran analyses on the gathered data, such as the duration distribution of all classes and interplay between the visual and audio labels.
- Set and currently manage two challenges on the dataset: Audio-Based Interaction Recognition and Audio-Based Interaction Detection

Honours and Awards

Outstanding Reviewer *October 2025*

International Conference on Computer Vision *IEEE/CVF Conference (ICCV)*

Awarded in recognition of meaningful peer review contributions.

Outstanding Reviewer *May 2025*

Conference on Computer Vision and Pattern Recognition *IEEE/CVF Conference (CVPR)*

Awarded in recognition of meaningful peer review contributions.

EgoVis 2022/23 Distinguished Paper Awards *June 2024*

First Joint Egocentric Vision Workshop *IEEE/CVF Conference (CVPR)*

Awarded to our 2023 ICASSP paper: "EPIC-Sounds: A large-scale dataset of actions that sound."

EPIC-KITCHENS Challenges Winner *June 2024*

First Joint Egocentric Vision Workshop *IEEE/CVF Conference (CVPR)*

Audio-Based Interaction Recognition (2nd), Audio-Based Interaction Detection (2nd), Action Detection (3rd)

Top 5 Third Year MEng Computer Science/Computer Science with Maths Student *October 2020*

Netcraft *University of Bristol*

Awarded to the Top-5 performing Computer Science and Computer Science with Maths students.

Research Activities

Presentations

- TIM: A Time-Interval Machine for Audio-Visual Action Recognition. Presented at the Sight & Sound Workshop (CVPR 2024) and Twelve Labs Multimodal Weekly #56.
- EPIC-KITCHENS Challenges. Presented at the First Joint Egocentric Vision Workshop (CVPR 2024).
- Oral Presentation for EPIC-Sounds. Presented at ICASSP 2023.

Reviewing

- 2026: CVPR, ECCV, ICPR, BMVC
- 2025: CVPR, NeurIPS, ICCV, IJCV, OJSP
- 2024: ECCV, TPAMI

Technical Knowledge and Programming

Programming Languages

- Python (Strongest)
- C++ (Intermediate)
- JavaScript (Intermediate)

Frameworks

- PyTorch (Strongest)
- TensorFlow (Intermediate)

Technical Fields

- Machine Learning, particularly Deep Learning
- Computer Vision
- Data Science