Video Understanding

University of BRISTOL

Dima Damen
PAISS 2025

Definitions…

***Ego**… a person's sense of self-esteem or self-importance*

Definitions…

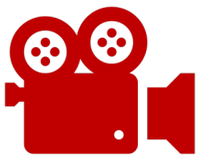
***Ego**… a person's sense of self-esteem or self-importance*

***Egocentric vision**… the wearer serves as the central reference point in the study of interesting entities: objects, actions, interactions and intentions*

In today's tutorial

Motivation and Datasets in Egocentric Video Understanding

Video Understanding Out of the Frame



Video Understanding: Data and Tasks

Teaser: The Wizard of Oz & Genie 3

Videos are Multimodal

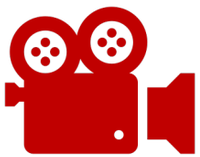
Outlook into the Future of Egocentric Vision

Connected Videos of One's Life

Conclusion

In today's tutorial

Motivation and Datasets in Egocentric Video Understanding



Video Understanding: Data and Tasks

Videos are Multimodal

Connected Videos of One's Life

Video Understanding Out of the Frame

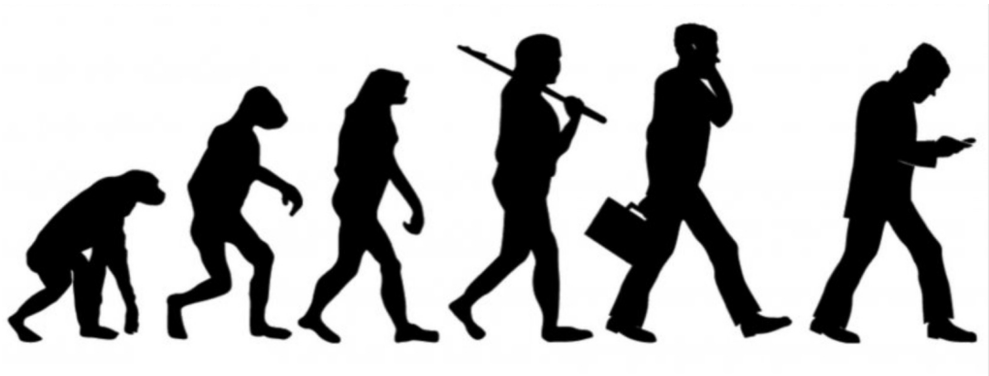
Teaser: The Wizard of Oz & Genie 3

Outlook into the Future of Egocentric Vision

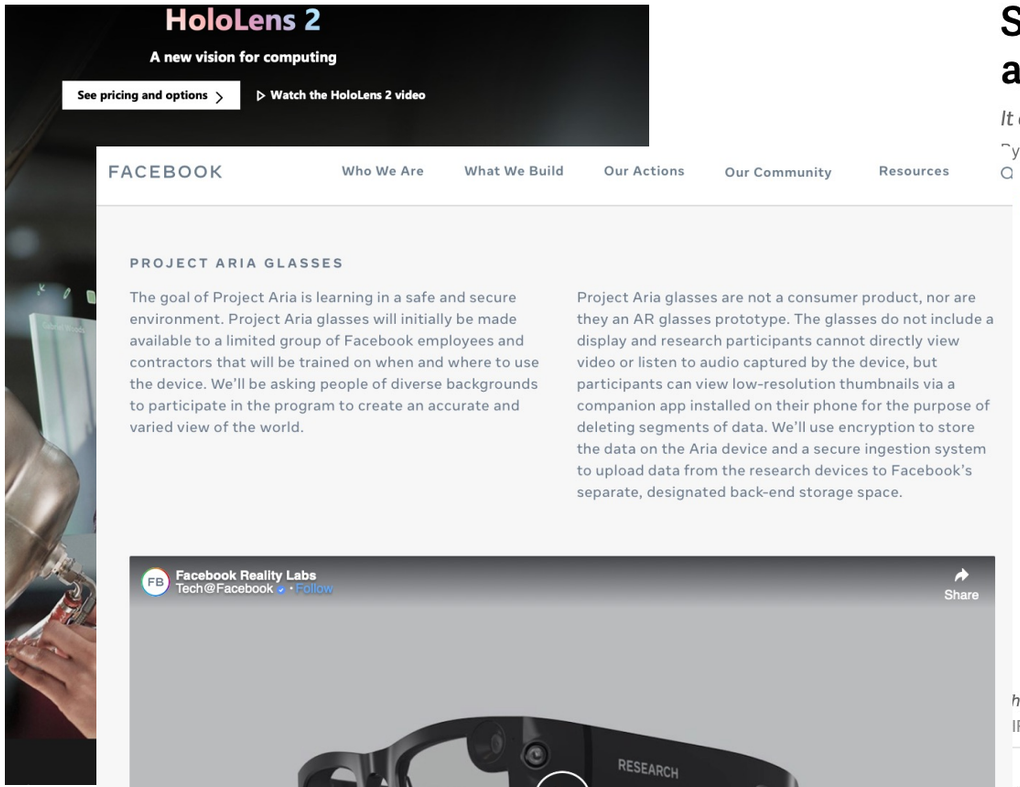
Conclusion

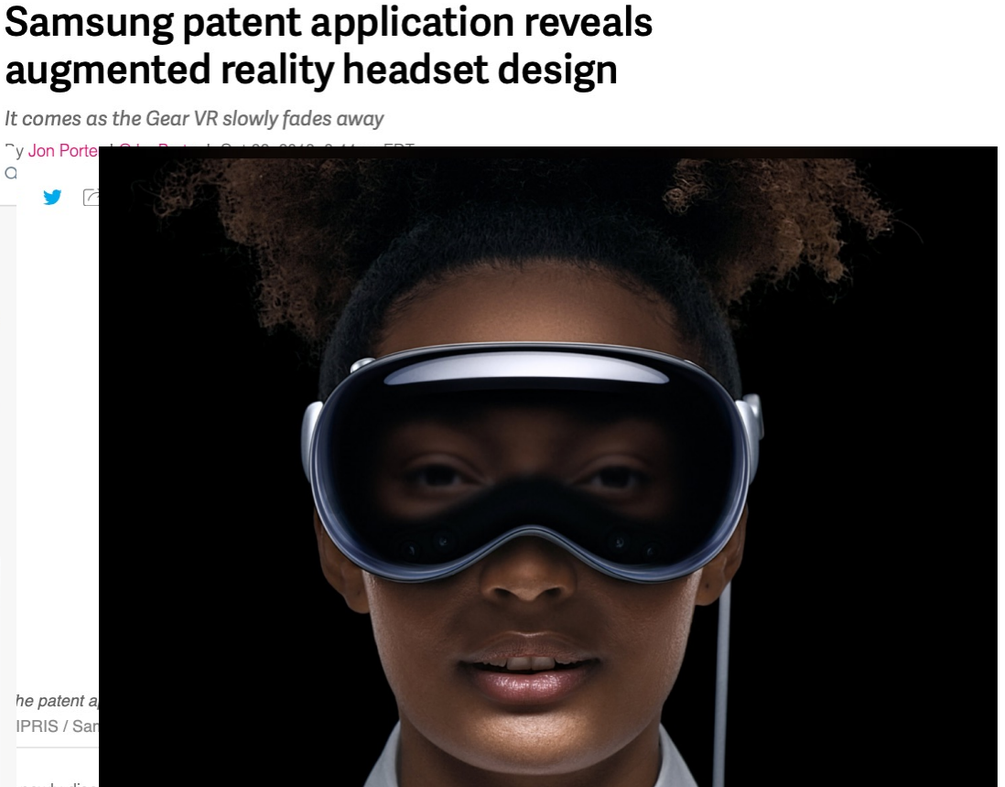
The present…

Photo *Illustration* by Pelle Cass



The future…


HoloLens 2
A new vision for computing
See pricing and options
Watch the HoloLens 2 video
FACEBOOK
Who We Are
What We Build
Our Actions
Our Community
Resources
PROJECT ARIA GLASSES
The goal of Project Aria is learning in a safe and secure environment. Project Aria glasses will initially be made available to a limited group of Facebook employees and contractors that will be trained on when and where to use the device. We'll be asking people of diverse backgrounds to participate in the program to create an accurate and varied view of the world.
Project Aria glasses are not a consumer product, nor are they an AR glasses prototype. The glasses do not include a display and research participants cannot directly view video or listen to audio captured by the device, but participants can view low-resolution thumbnails via a companion app installed on their phone for the purpose of deleting segments of data. We'll use encryption to store the data on the Aria device and a secure ingestion system to upload data from the research devices to Facebook's separate, designated back-end storage space.
Facebook Reality Labs
Tech@Facebook
Follow
Share
RESEARCH

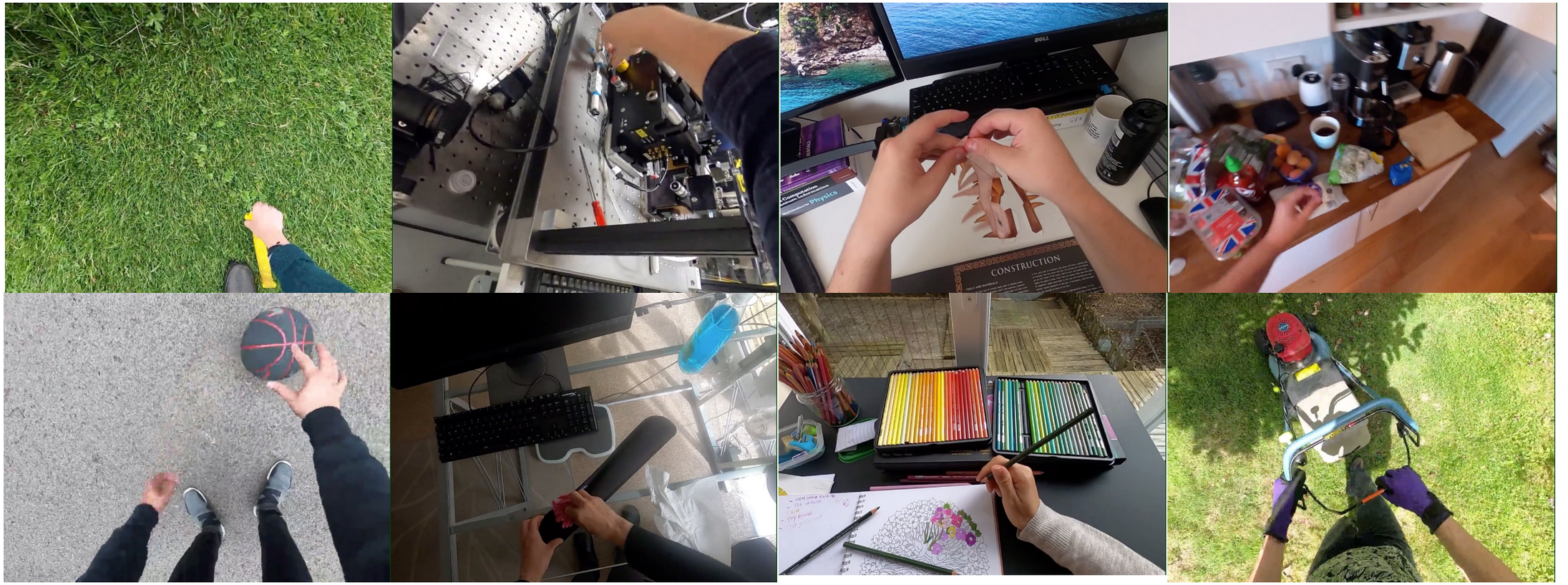

Samsung patent application reveals augmented reality headset design
It comes as the Gear VR slowly fades away
By Jon Porter
he patent a
IPRIS / San
newly disc
ugmented r
pplication,
hows a cab
e cable is j

The future is here…

The future can be imagined…

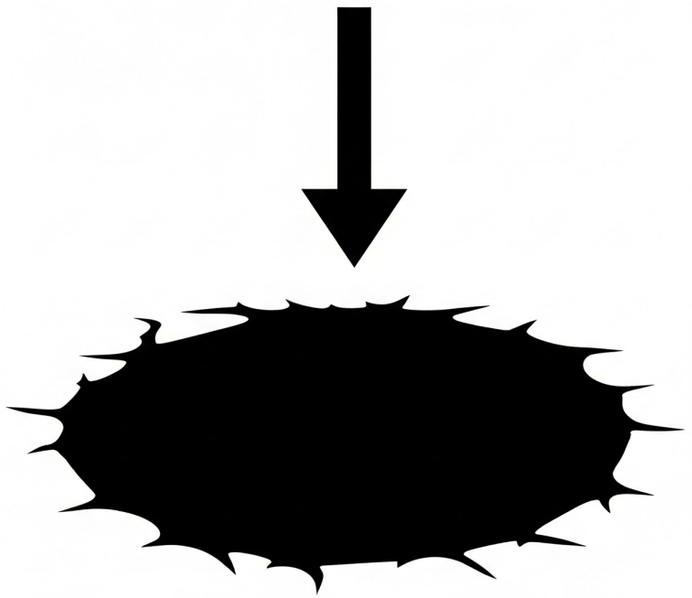
University of BRISTOL

Egocentric Videos?

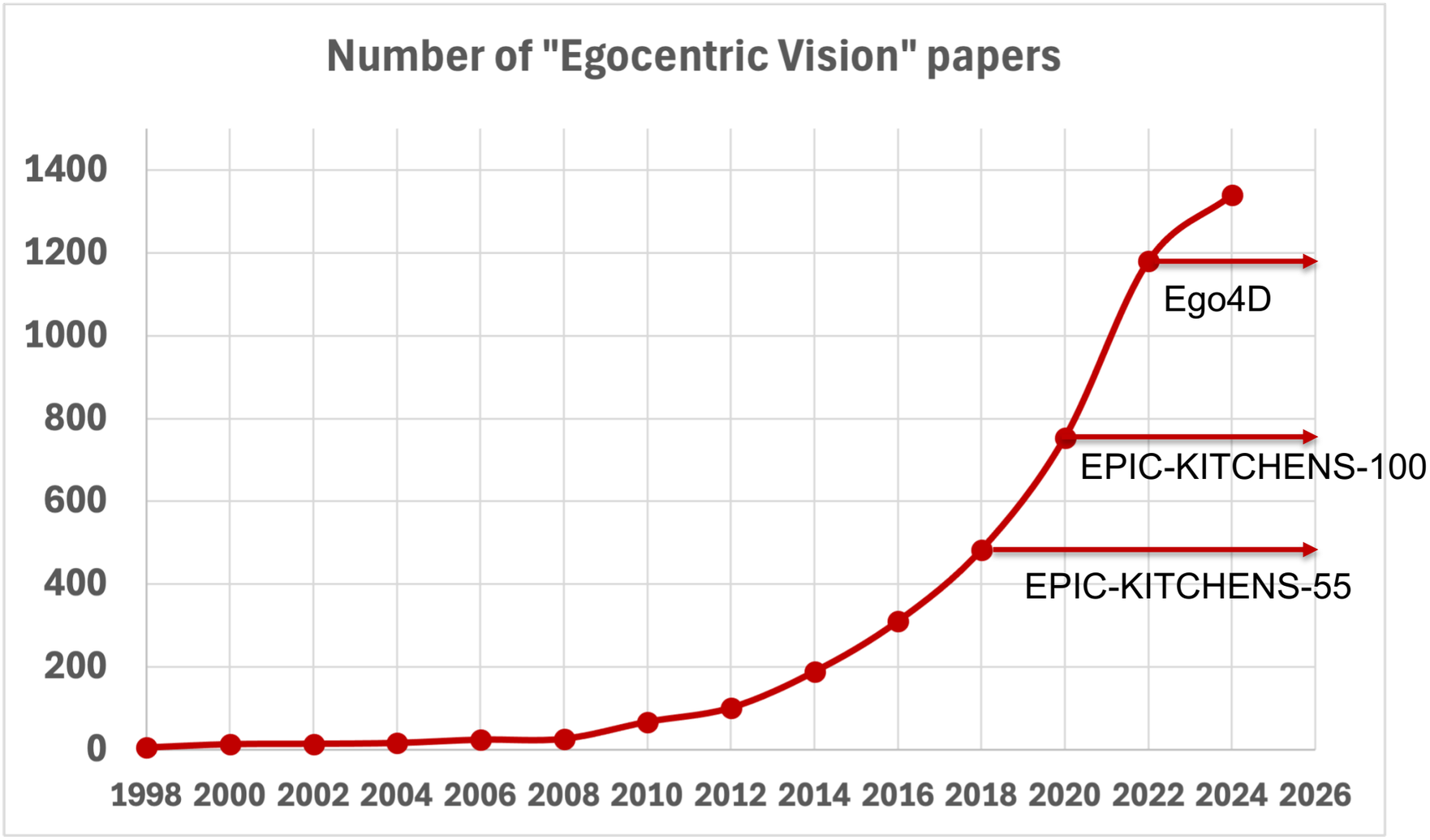


Machine Learning in Practice

no data
no machine
learning research



Is research catching up?

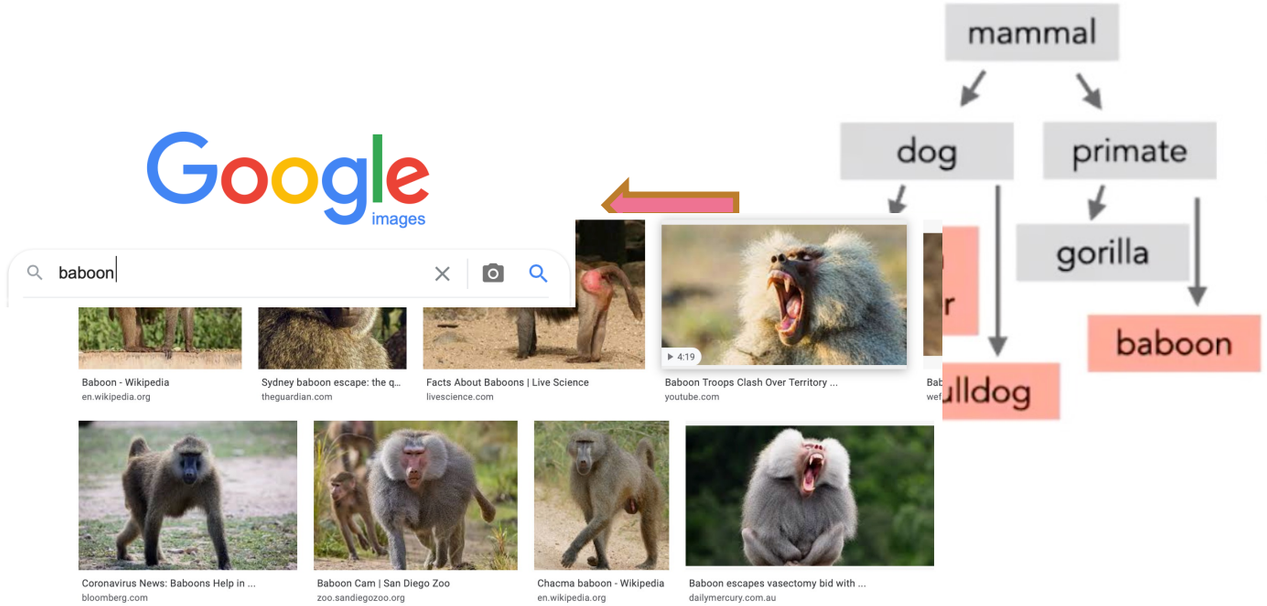


Statistics from Google Scholar – June 2025

Machine Learning in Practice

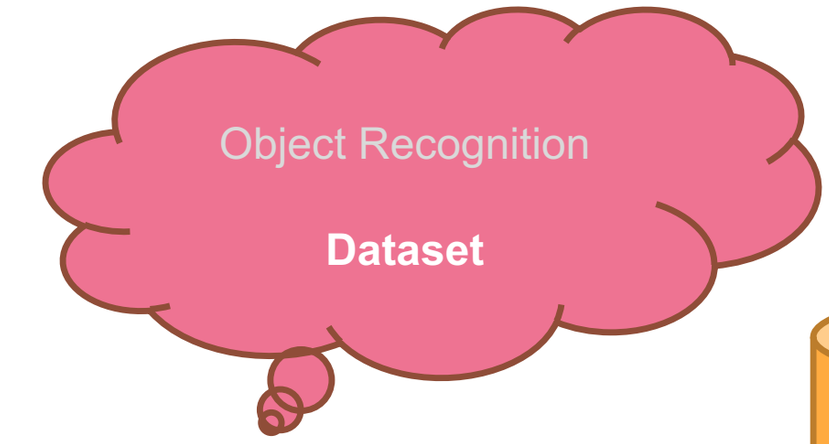
Real World Task….

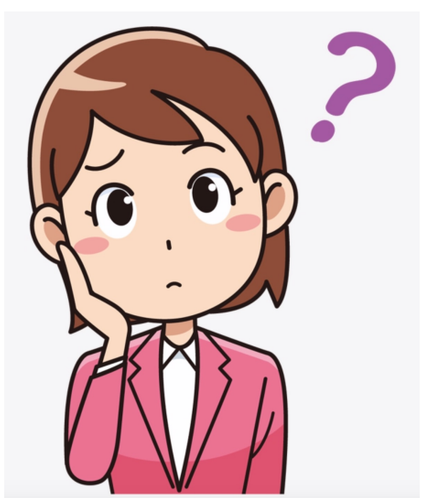
Object Recognition

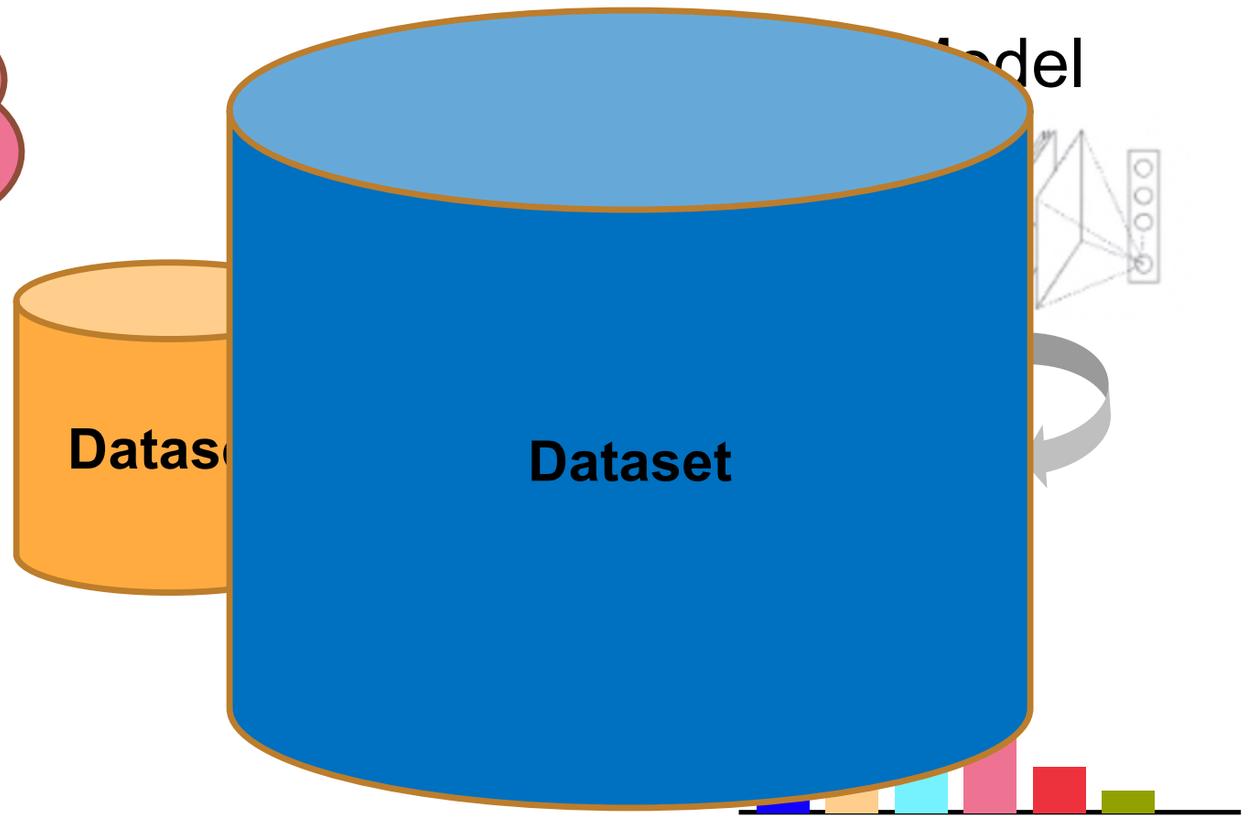


Dataset

Machine Learning in Practice







Machine Learning in Practice

- Applies to Most ML research at the moment
  - Object Recognition (Pascal, ImageNet, Places, …)
  - Action Recognition (Kinetics-400, -600, -700, AVA, SS, …)
  - …
- Datasets:
  - Methods overfit to the dataset
  - useful for **one** task
  - unnaturally balanced (or nearly balanced) – unrelated to priors outside the dataset itself

One Exception

Machine Learning in Practice

- Autonomous Driving…

Welcome to the KITTI Vision Benchmark Suite!

We take advantage of our autonomous driving platform Annieway to develop novel challenging real-world computer vision benchmarks. Our tasks of interest are: stereo, optical flow, visual odometry, 3D object detection and 3D tracking. For this purpose, we equipped a standard station wagon with two high-resolution color and grayscale video cameras. Accurate ground truth is provided by a Velodyne laser scanner and a GPS localization system. Our datsets are captured by driving around the mid-size city of Karlsruhe, in rural areas and on highways. Up to 15 cars and 30 pedestrians are visible per image. Besides providing all data in raw format, we extract benchmarks for each task. For each of our benchmarks, we also provide an evaluation metric and this evaluation website. Preliminary experiments show that methods ranking high on established benchmarks such as Middlebury perform below average when being moved outside the laboratory to the real world. Our goal is to reduce this bias and complement existing benchmarks by providing real-world benchmarks with novel difficulties to the community.

Share

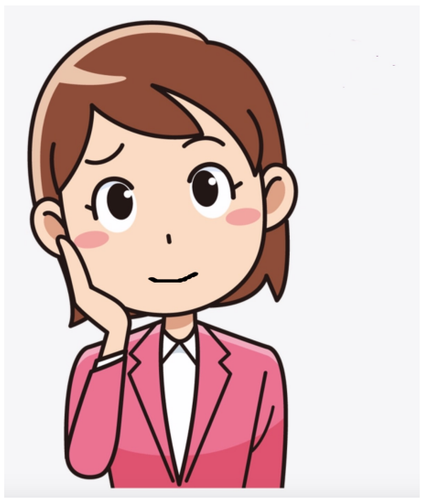
To get started, grab a cup of your favorite beverage and watch our video trailer (5 minutes):

stereo flow sceneflow depth odometry object tracking road semantics raw data

Machine Learning in Practice

~~Object Recognition~~

Let's collect Data!



EPIC KITCHENS-100

EPIC-KITCHENS

EPIC-KITCHENS


open oven
put spoon on counter
put on glove
pick up fork

Ego4D

with: Kristen Grauman
+83 authors

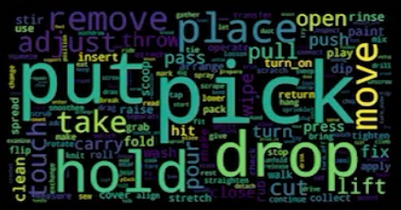
Narration

C: camera wearer

#C C scraps off wood filler from one putty knife with the other putty knife
#C C picks up another putty knife from the white board

13.2 sentences/min
3.8 M sentences

1,772 verbs



4,336 nouns

Data Collection Exercise

EPIC KITCHENS

2017 - now

100 hours
45 kitchens
4 countries
Long-term recording
Kitchen-based activities

EGO4D

2020 - now

13 universities
3670 hours
923 participants
74 locations
9 countries
Short-term recording
All daily activities

Data Collection Exercise

EGO-EXO4D

2024 - now

1286 hours
740 camera wearers
Skilled activities

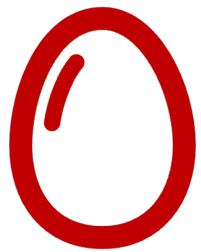
HD-EPIC

2025 - now

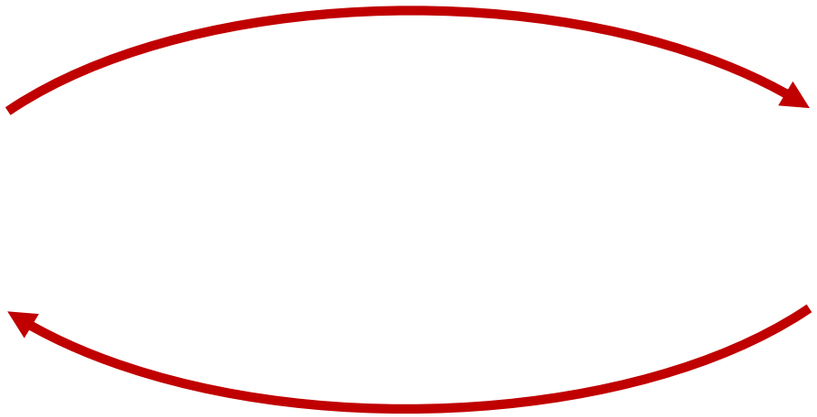
Validation dataset
41 hours
9 participants
Highly-Detailed
Digital Twin

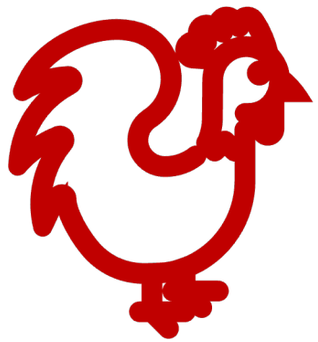
Data Collection Exercise

Pascal VOC
ImageNet
Kinetics
Something-Something



Labels

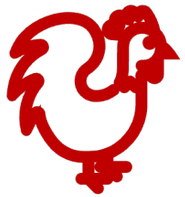




Data

EPIC-KITCHENS
Ego4D
Ego-Exo4D
HD-EPIC
…
KITTI

The chicken or the egg...

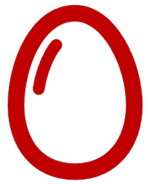
Data 

Naturally unbalanced

Harder to label (exposes ambiguity)

Closer to application

Multiple tasks

Labels 

Unnaturally balanced (or nearly)

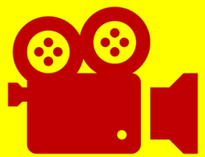
Easier to label (hides ambiguity)

Can be expanded

Single task

In today's tutorial

Motivation and Datasets in Egocentric Video Understanding



Video Understanding: Data and Tasks

Videos are Multimodal

Connected Videos of One's Life

Video Understanding Out of the Frame

Teaser: The Wizard of Oz & Genie 3

Outlook into the Future of Egocentric Vision

Conclusion

The history of VIDEO

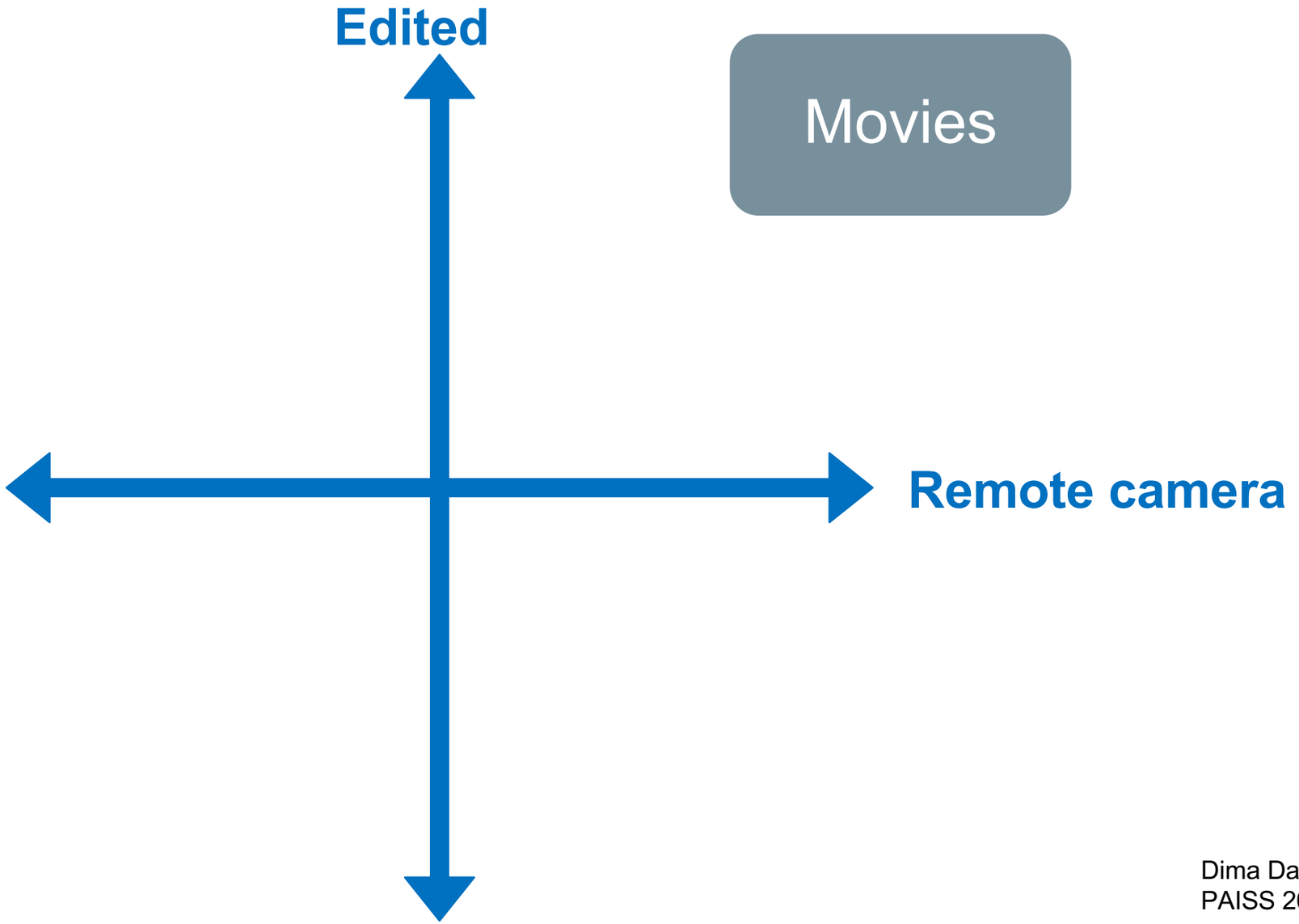
Arrival of a Train at La Ciotat (The Lumière Brothers, 1895)

The history of VIDEO

University of BRISTOL

Dima Damen
PAISS 2025

The history of VIDEO Understanding



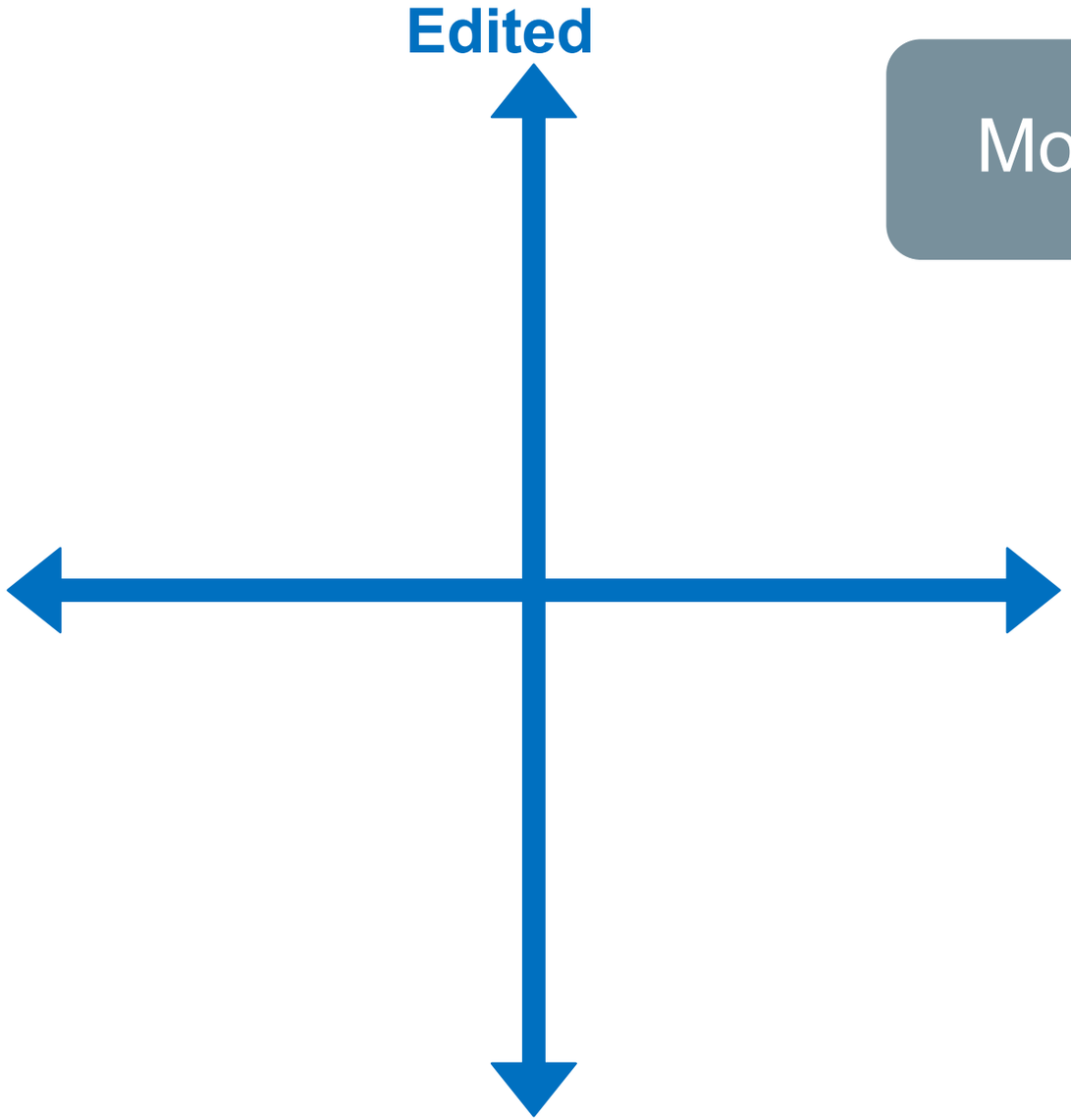
The history of VIDEO Understanding

Figure 1. Examples of two action classes (drinking and smoking) from the movie "Coffee and Cigarettes". Note the high within-

Laptev and Perez (2007)

The history of VIDEO Understanding

The history of VIDEO Understanding



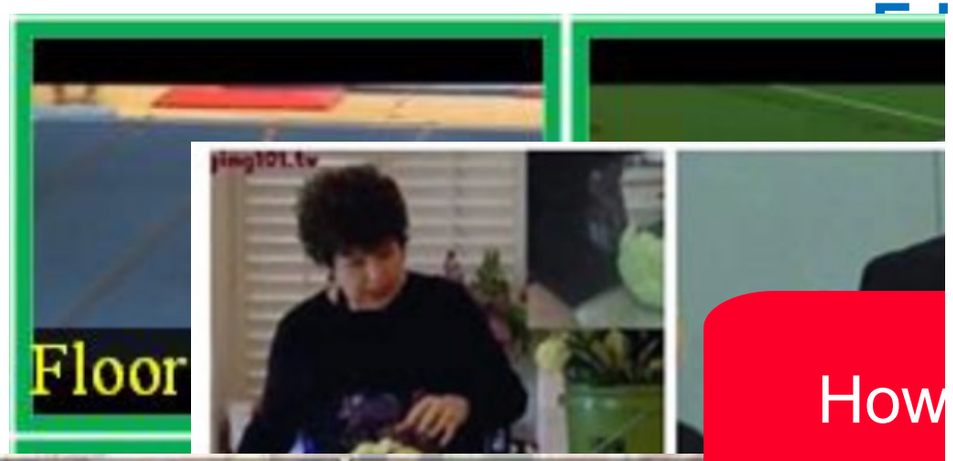
Movies

CCTV

The history of VIDEO Understanding

Damen and Hogg (2009). Recognizing linked events: Searching the space of feasible explanations. CVPR1

The history of VIDEO Understanding

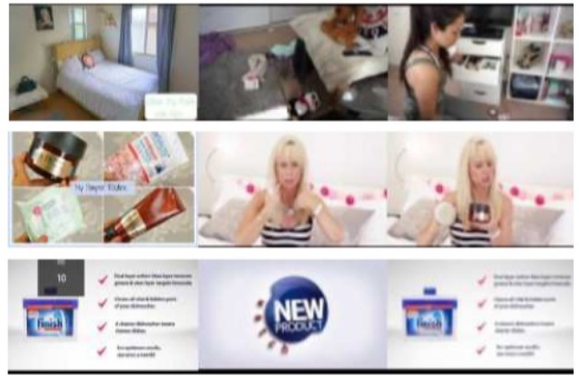


How

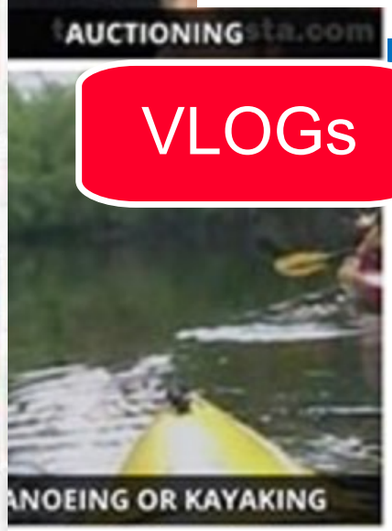
Templated, Multilingual Domain Queries:

"Morning routine", "realistic ditl 2015", "mijn realistische routine", "Ma routine d'apres-midi", …

216K Video Candidates (2.5 Years)
Low *Video-level* Purity



Egocentric unscripted



VLOGs

YouTube Videos

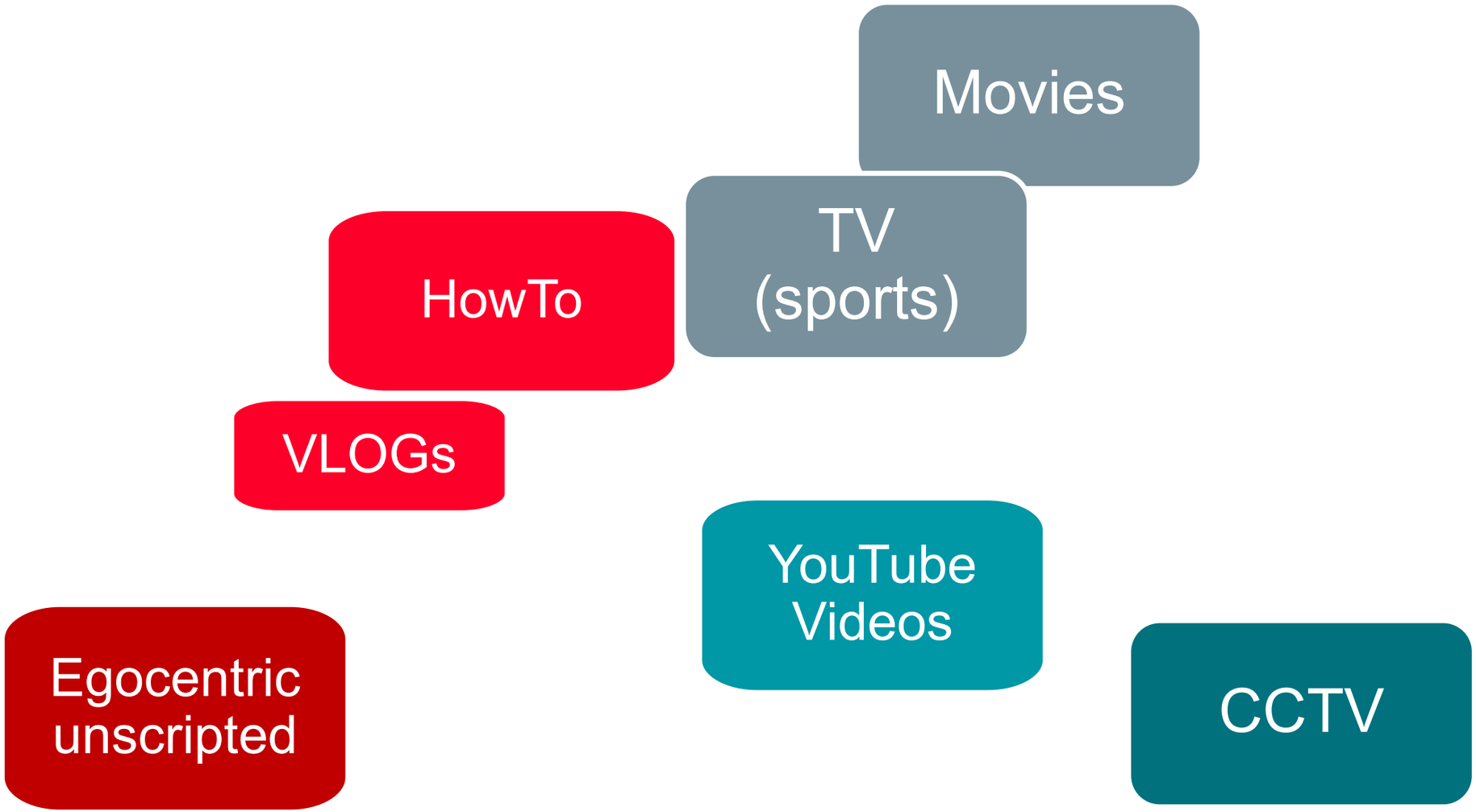
Remote camera

CCTV

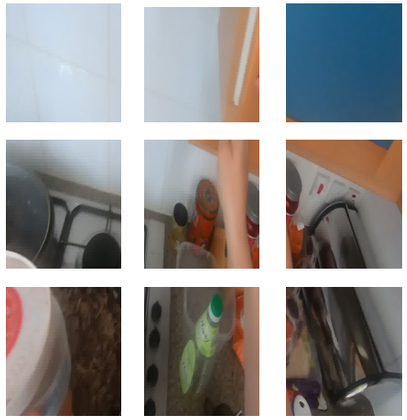
Edited

Dima Damen
PAISS 2025

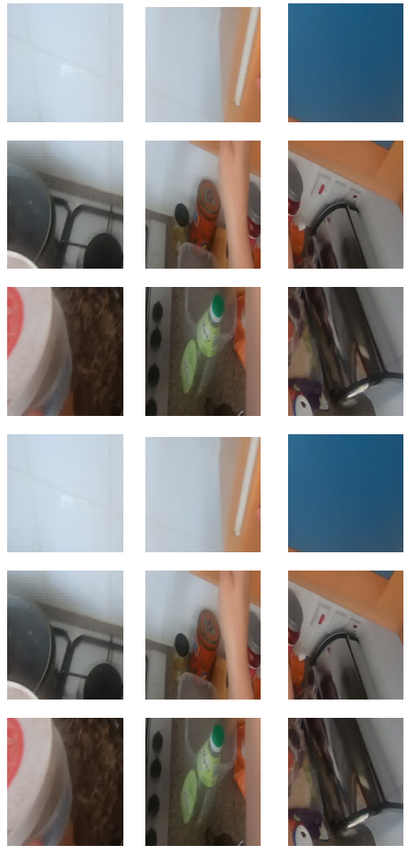
The history of VIDEO Understanding



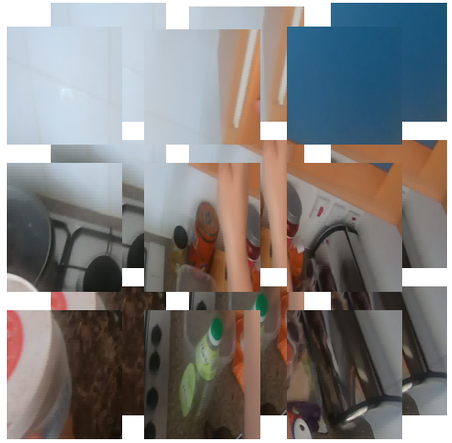
How to Patch-ify a Video?



H x W x C

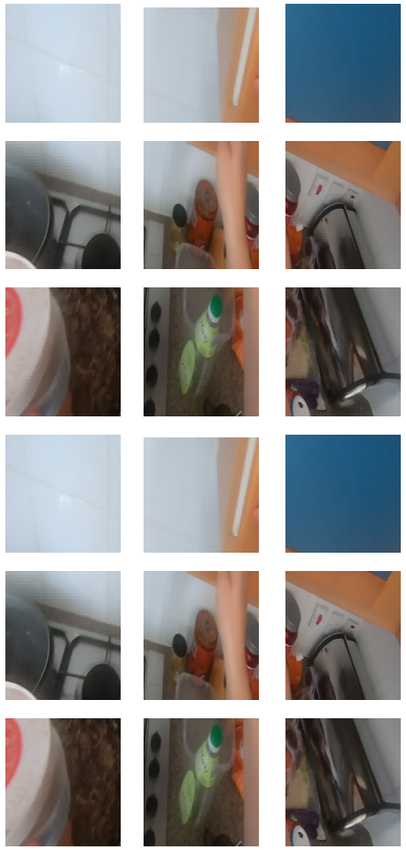


[TH] x W x C



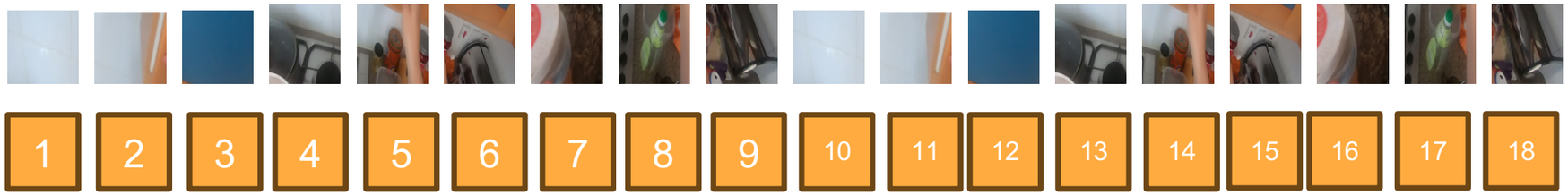
H x W x [CT]

How to Patch-ify a Video?

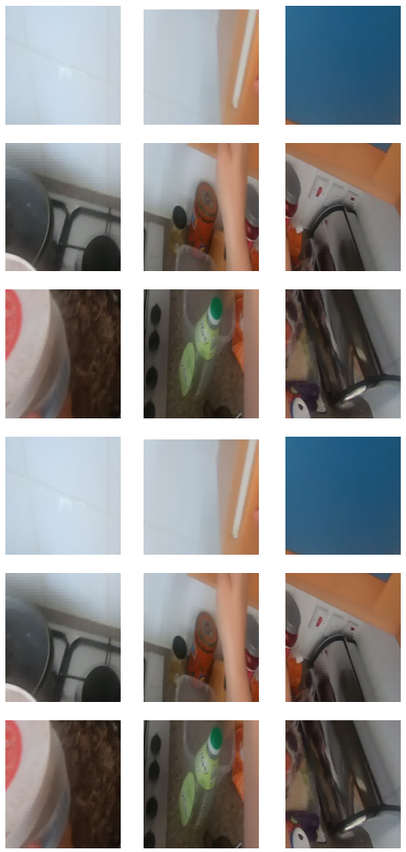


[TH] x W x C

Flatten

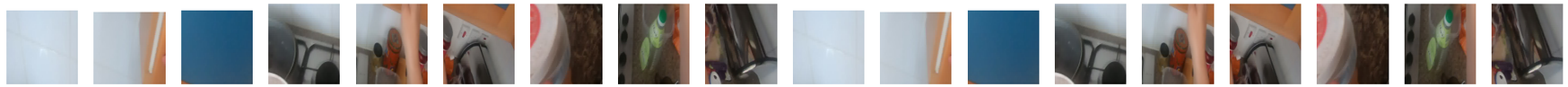


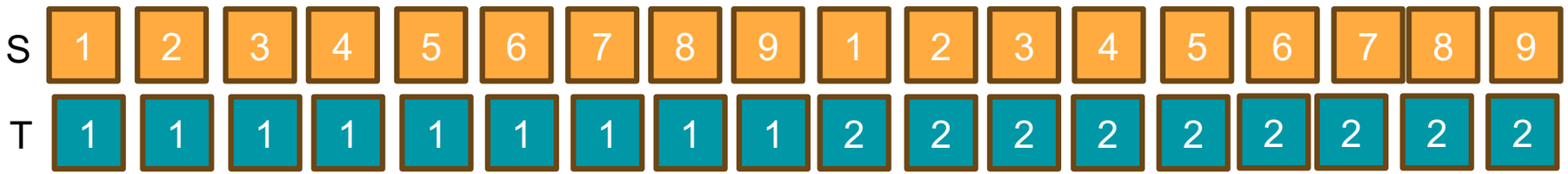
How to Patch-ify a Video?



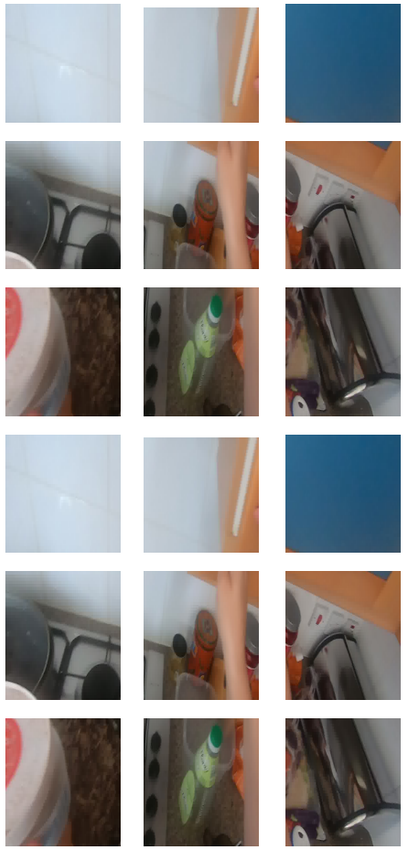
[TH] x W x C

Flatten

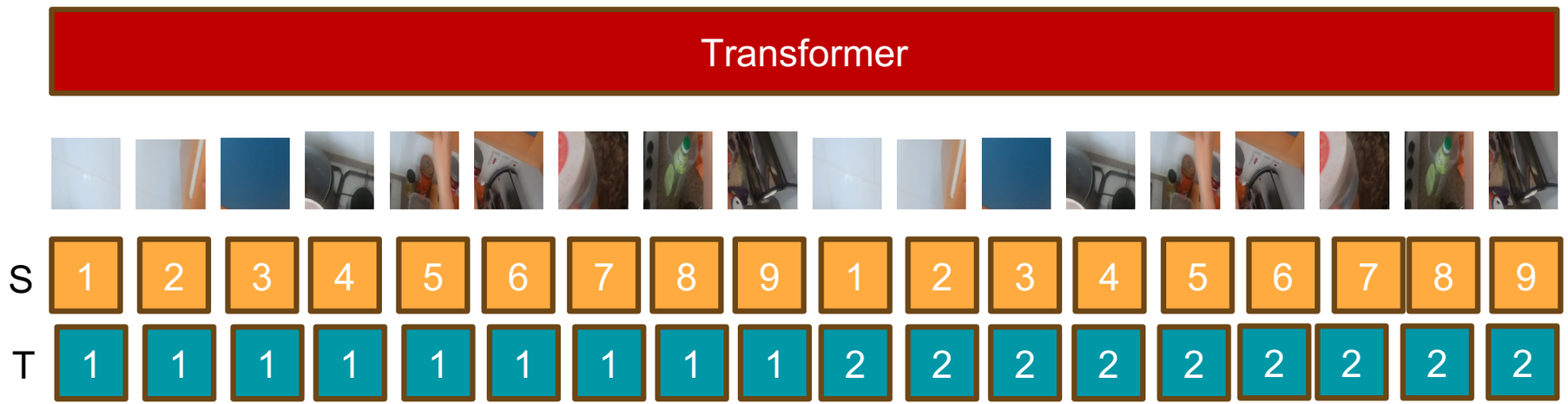




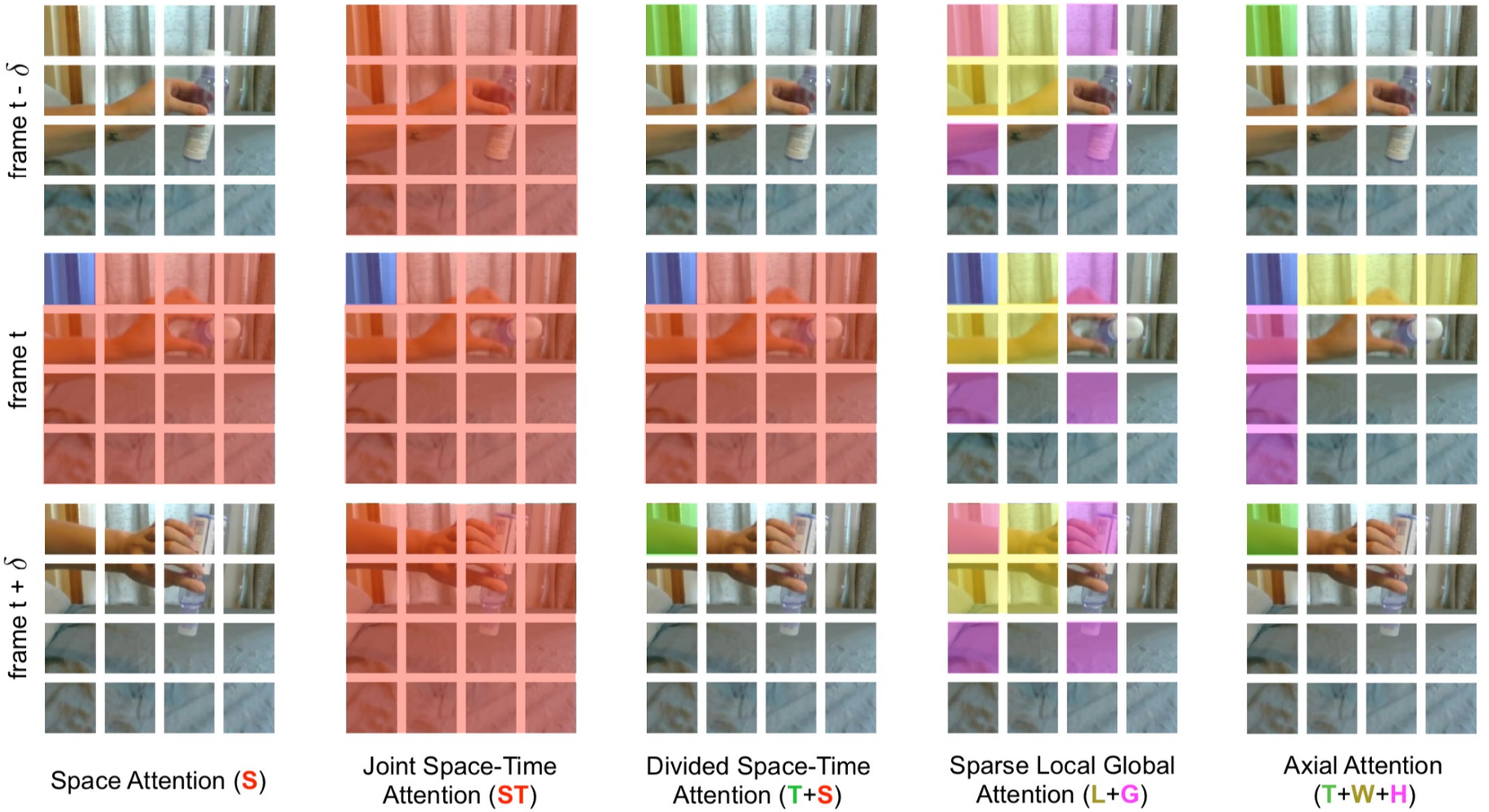
How to Patch-ify a Video?



[TH] x W x C

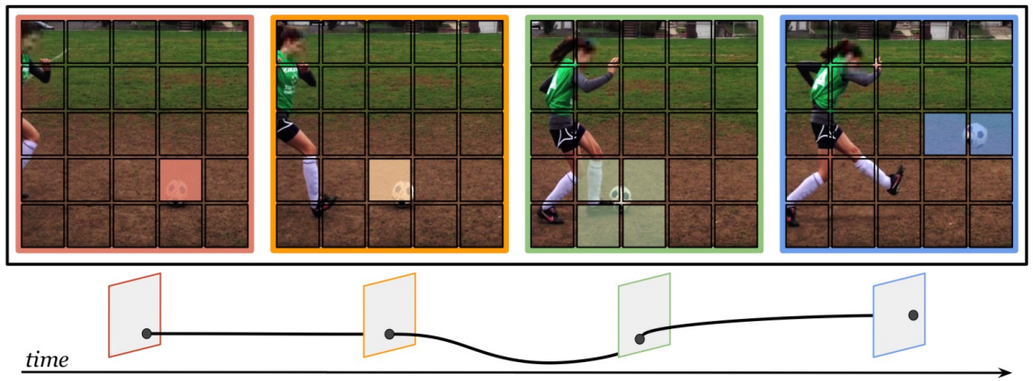


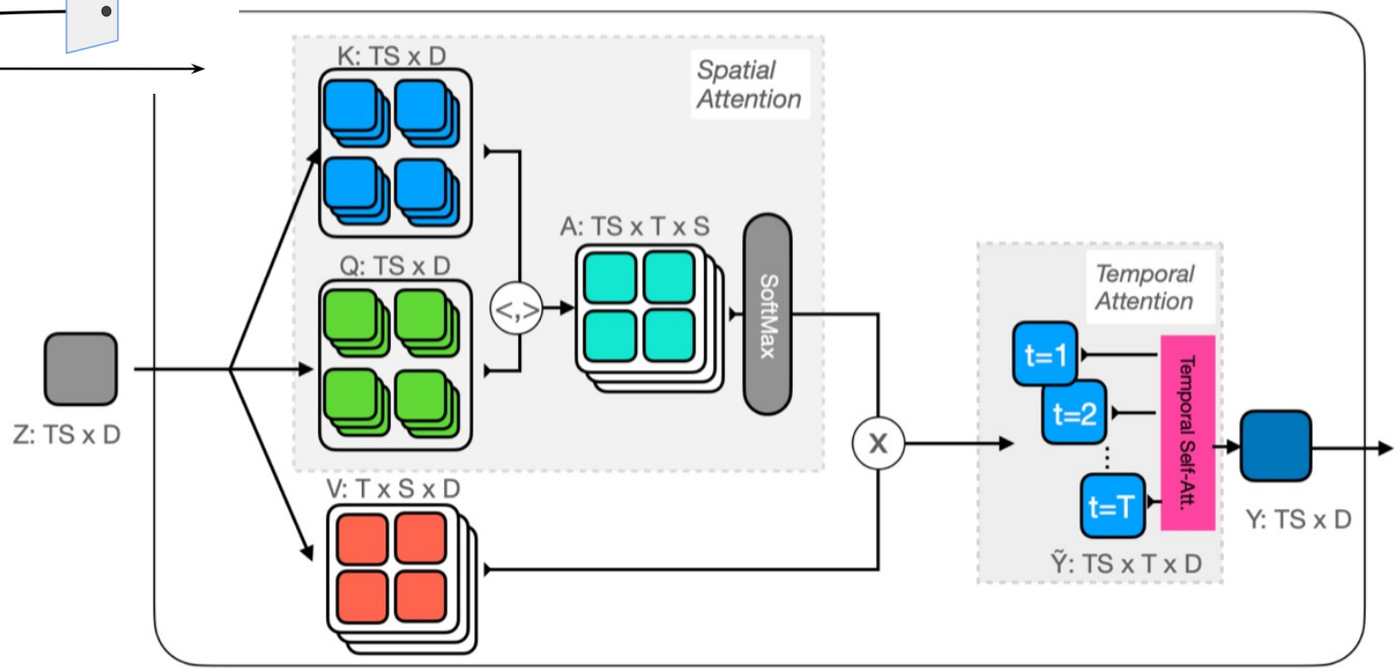
TimeSFormer



Bertasius et al (2021). Is Space-Time Attention All You Need for Video Understanding? ArXiv

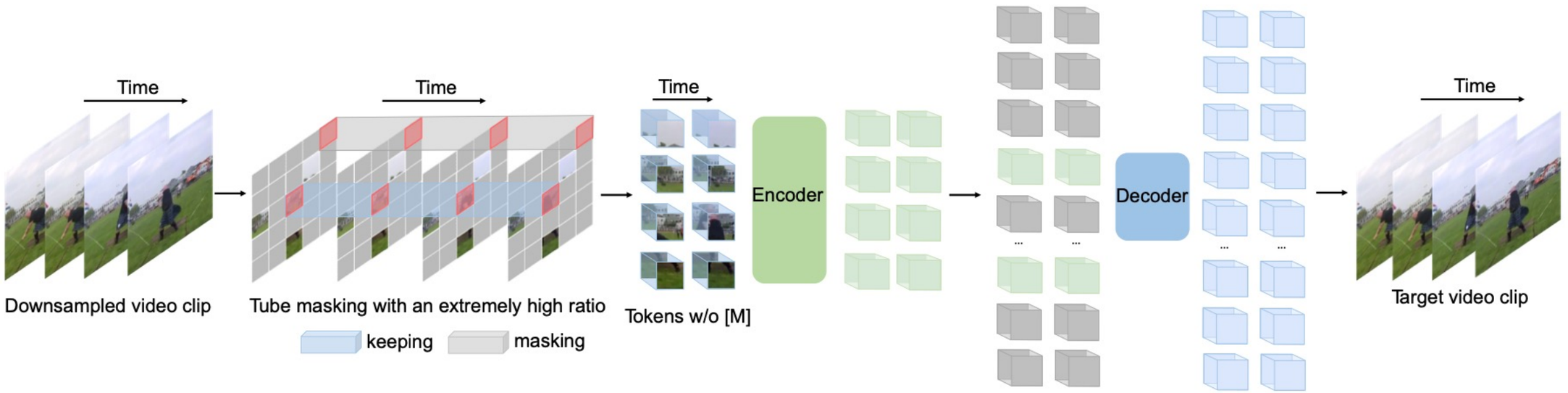
MotionFormer





Patrick et al (2021). Keeping Your Eye on the Ball: Trajectory Attention in Video Transformers. NeurIPS

VideoMAE



Tong et al (2022). VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. NeurIPS

OmniVore

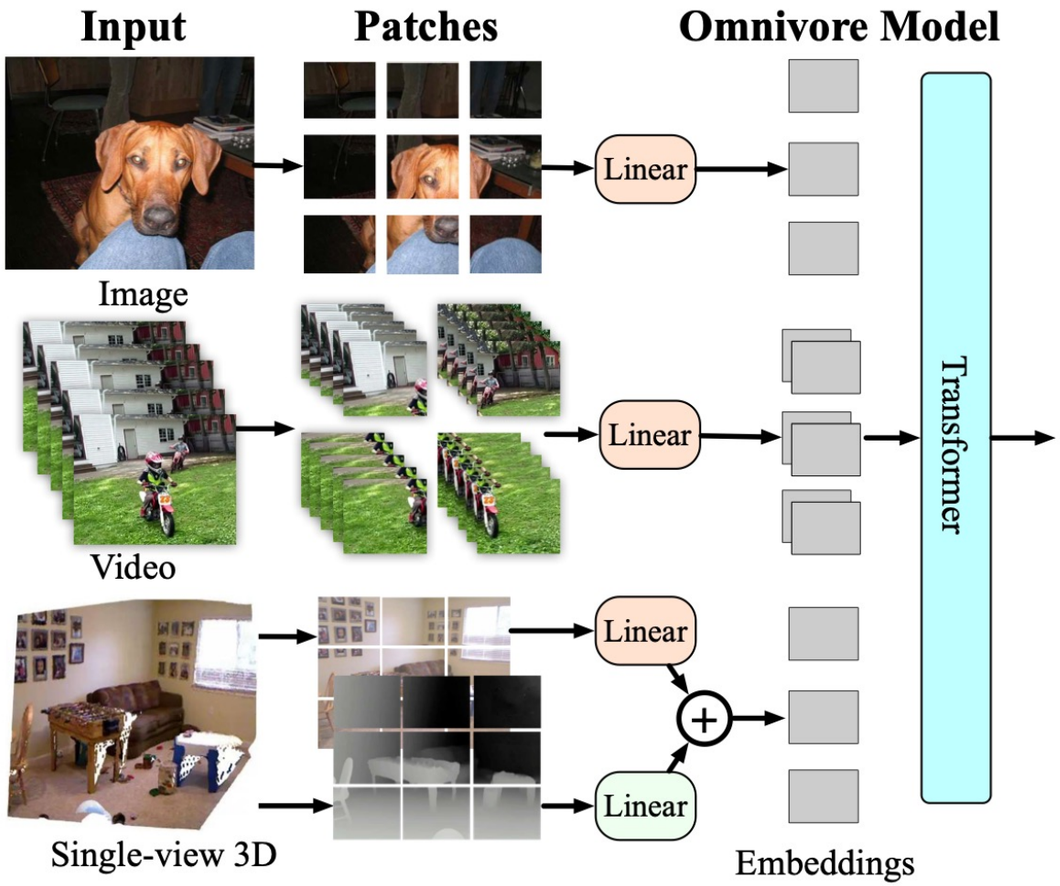
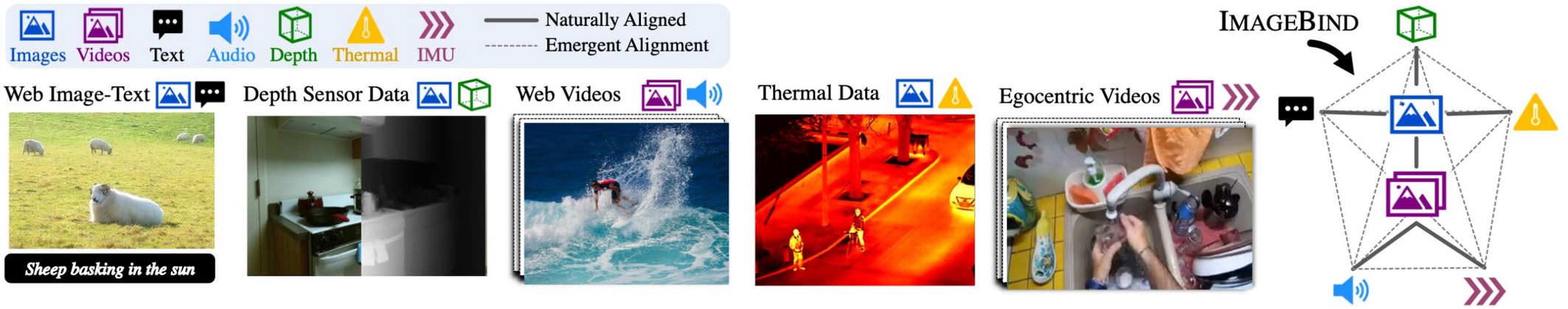


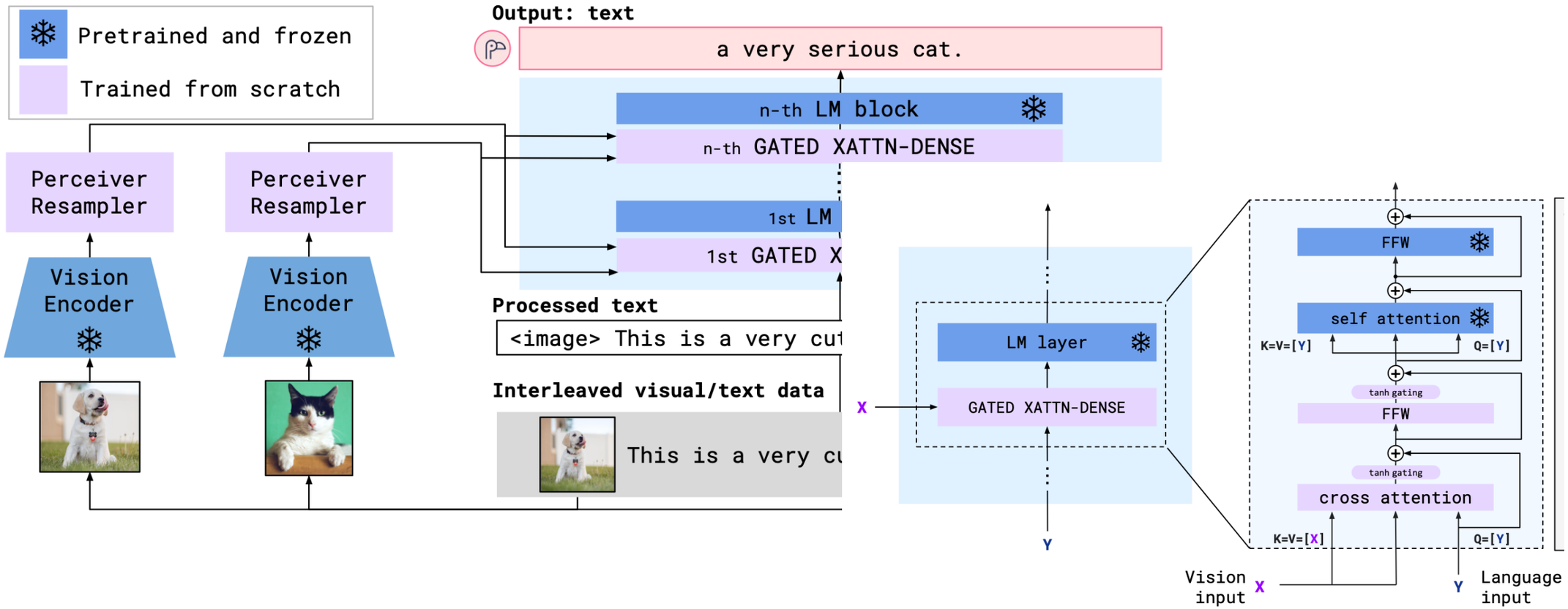
Figure 2. Multiple visual modalities in the OMNIVORE model.

ImageBind



$$L_{\mathcal{I},\mathcal{M}} = -\log \frac{\exp(\mathbf{q}_i^\intercal \mathbf{k}_i/\tau)}{\exp(\mathbf{q}_i^\intercal \mathbf{k}_i/\tau) + \sum_{j\neq i} \exp(\mathbf{q}_i^\intercal \mathbf{k}_j/\tau)}$$

Flamingo


Pretrained and frozen
Trained from scratch
Perceiver Resampler
Perceiver Resampler
Vision Encoder
Vision Encoder
Output: text
a very serious cat.
n-th LM block
n-th GATED XATTN-DENSE
1st LM
1st GATED X
Processed text
<image> This is a very cu
Interleaved visual/text data
This is a very cu
X
LM layer
GATED XATTN-DENSE
Y
FFW
self attention
K=V=[Y]
Q=[Y]
tanh gating
FFW
tanh gating
cross attention
K=V=[X]
Q=[Y]
Vision input
X
Y
Language input

InternVideo

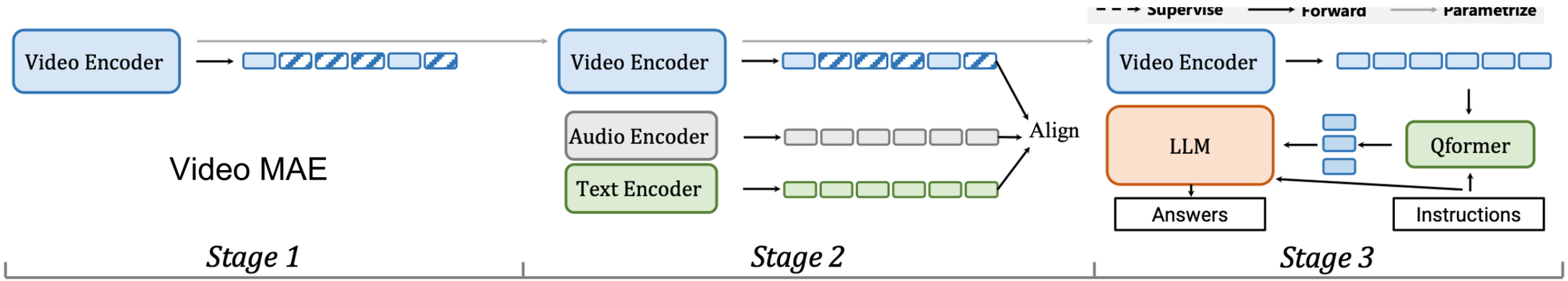
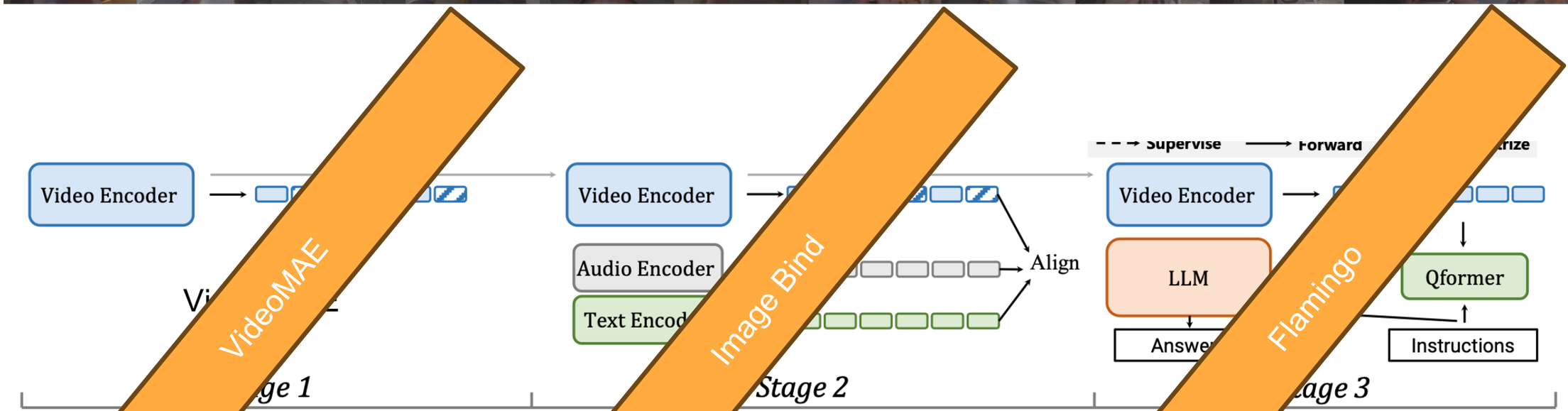


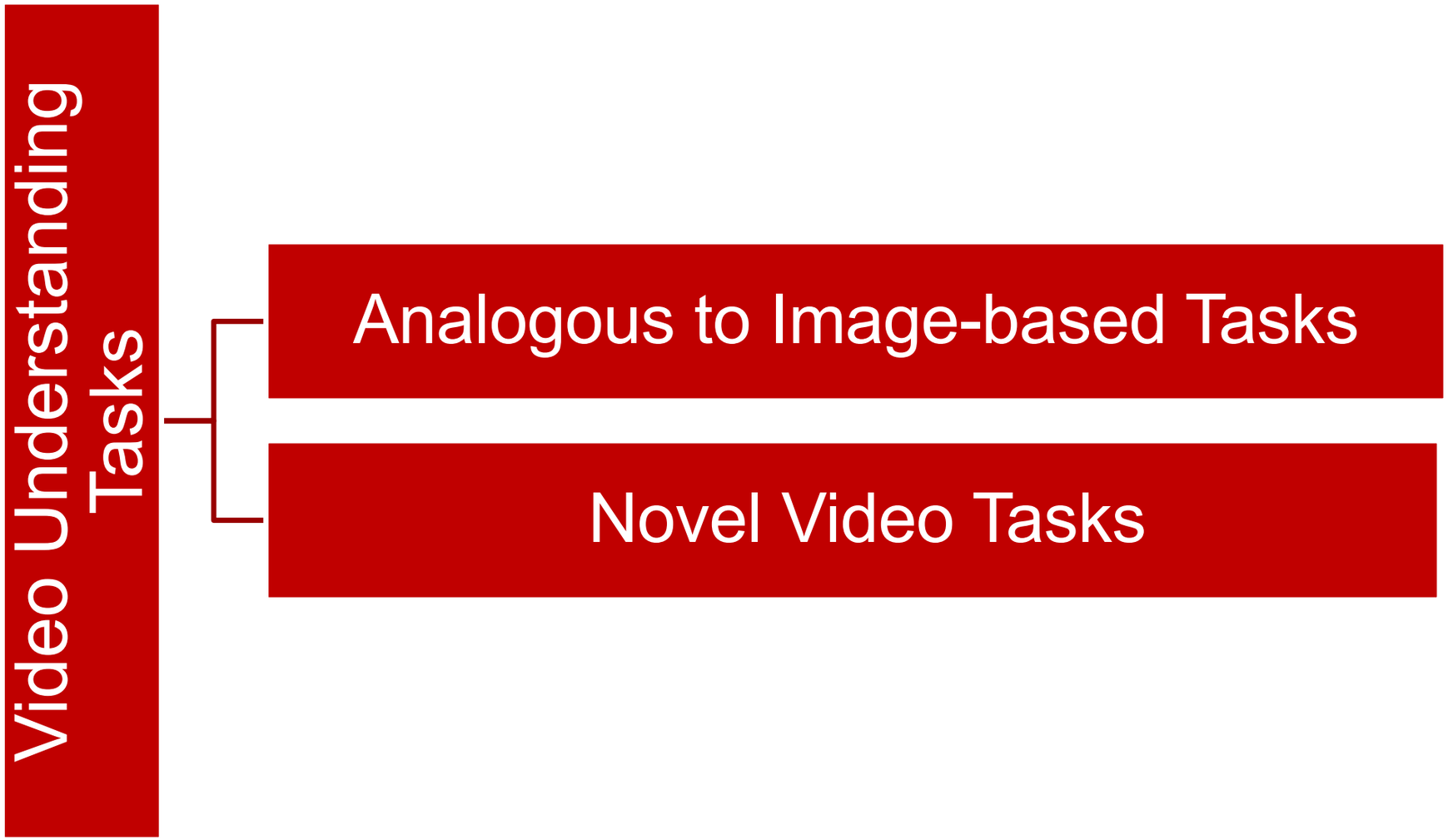
Figure 2: Framework of **InternVideo2**. It consists of three consecutive training phases: unmasked video token reconstruction, multimodal contrastive learning, and next token prediction. In stage 1, the video encoder is trained from scratch, while in stages 2 and 3, it is initialized from the version used in the previous stage.

InternVideo



Figu amework of **InternVideo2.** sts of three consecutive trainin unmasked video token re on, multimodal contrastive le nd next token prediction. In stage deo encoder is trained from s while in stages 2 and 3, it is in from the version used in the previ ge.

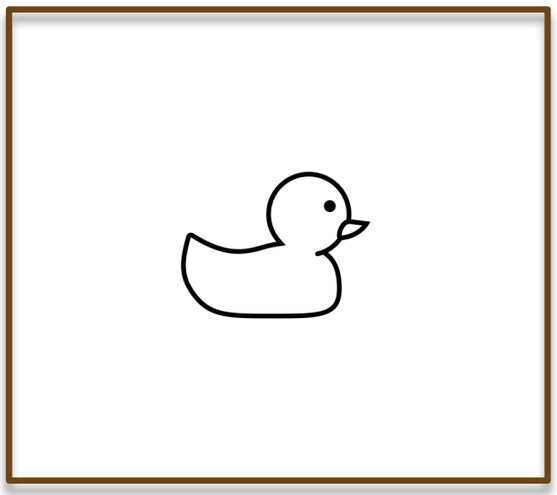
Two types of video understanding tasks



Analogous Tasks

Image

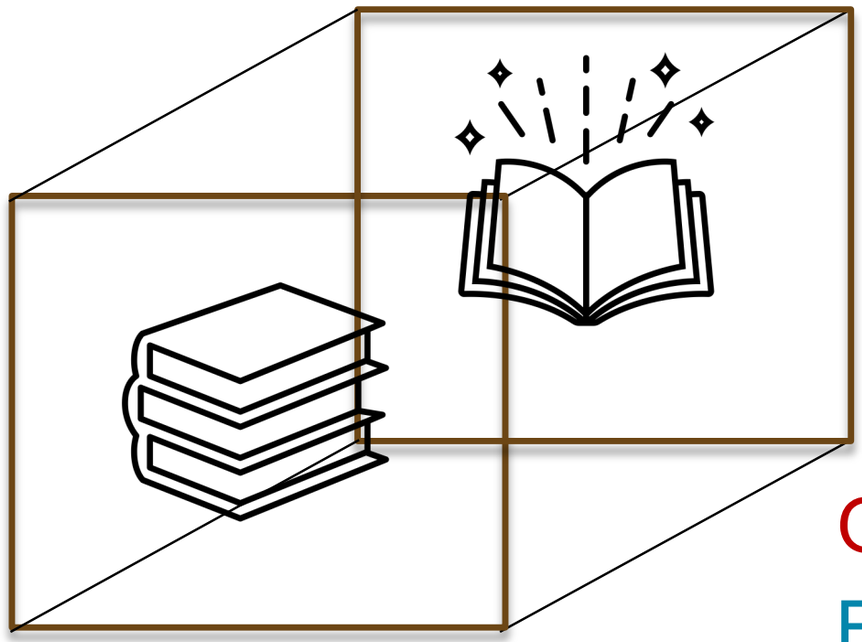
- Object Recognition



Duck

Video

- Action Recognition

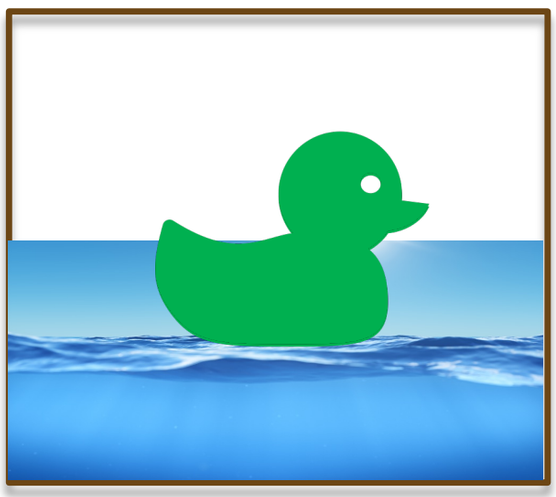


Open Book

Analogous Tasks

Image

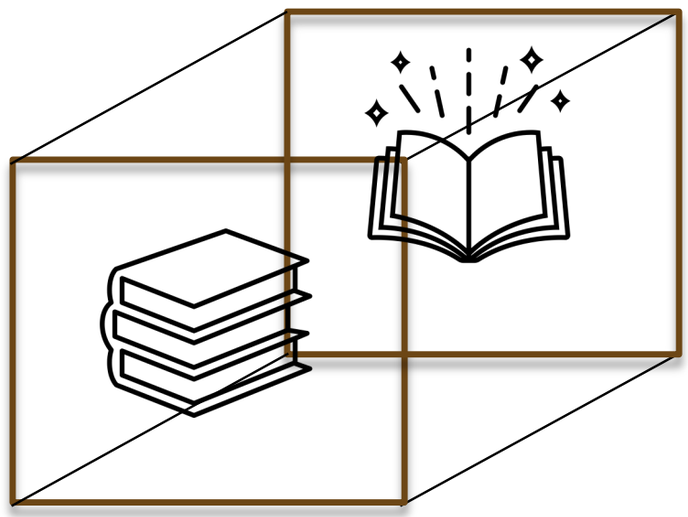
- Image Captioning



A green duck swimming
In clear water

Video

- Video Captioning

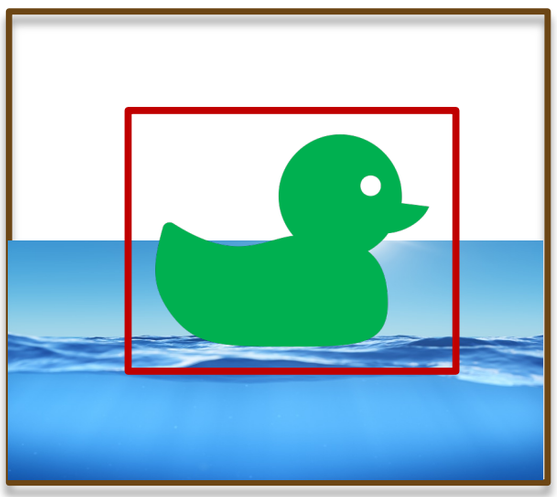


A book picked from top of the pile
and opened to a page in the middle

Analogous Tasks

Image

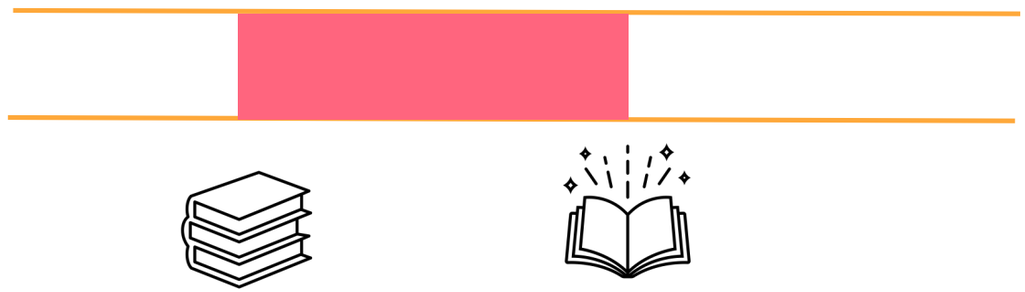
- Object Detection



Duck

Video

- Action Detection

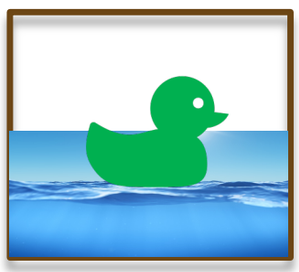


Open Book

Analogous Tasks

Image

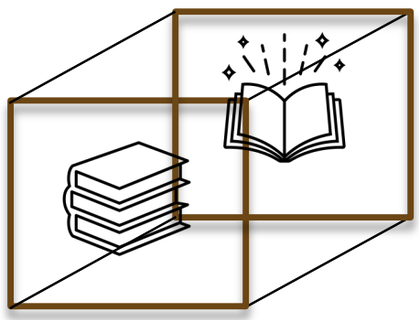
- Image Retrieval



Duck

Video

- Video Retrieval

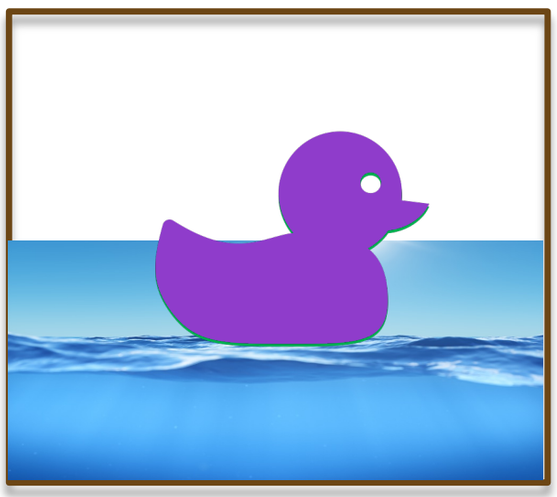


Open Book

Analogous Tasks

Image

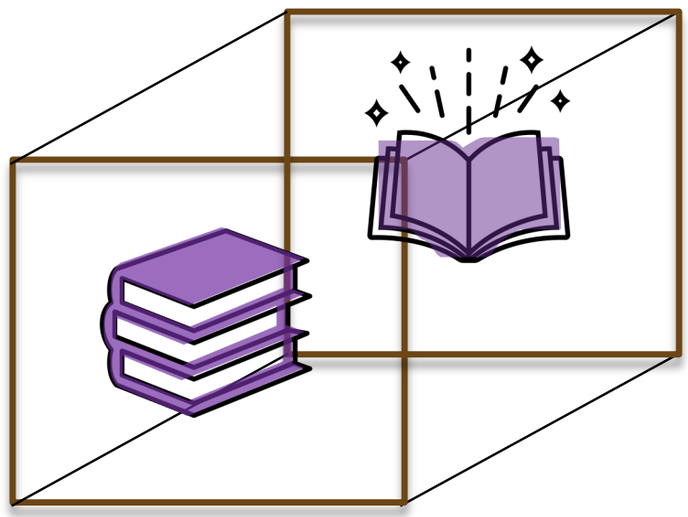
- Object Segmentation



Duck

Video

- Video Object Segmentation



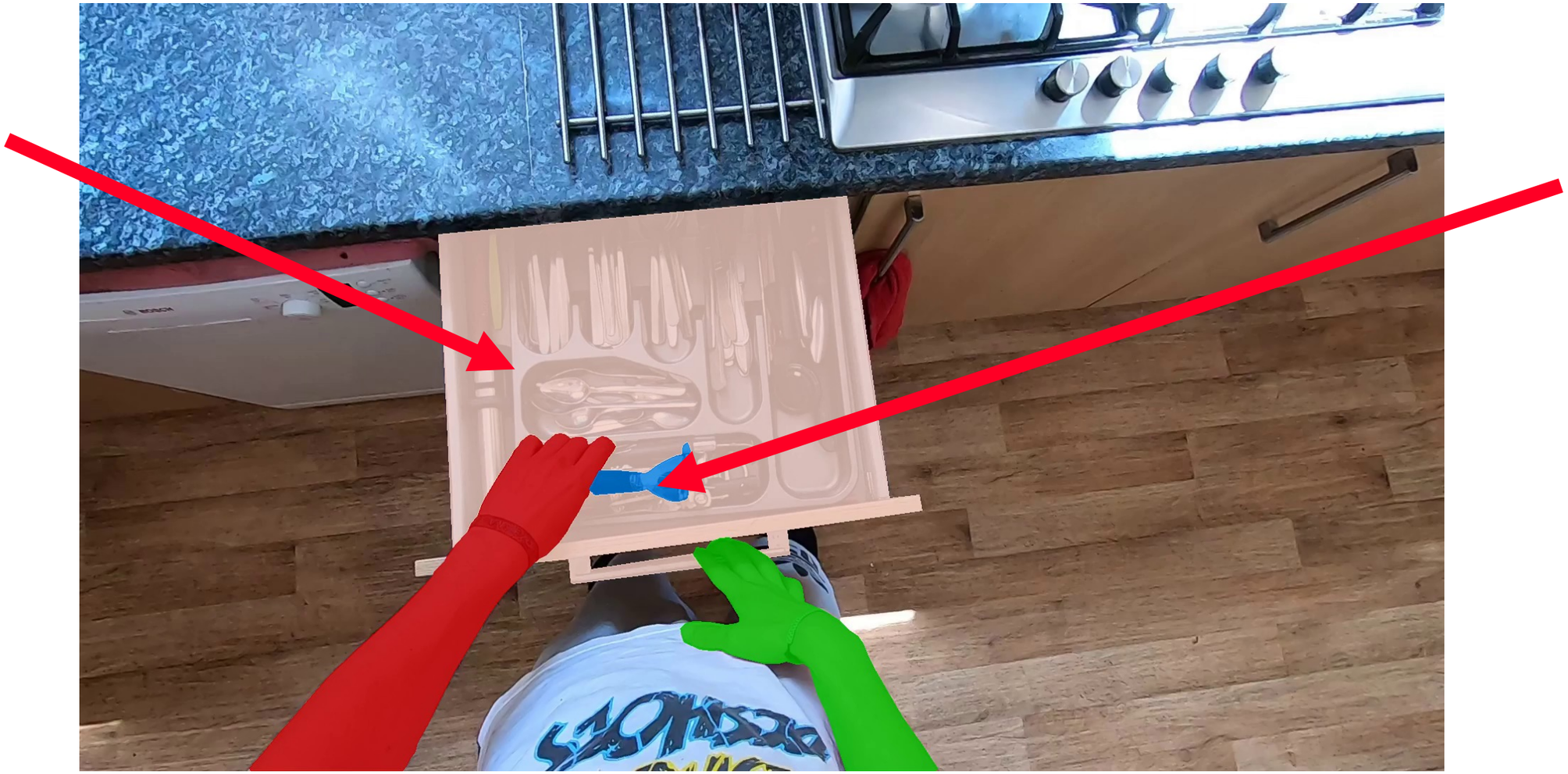
Book

EPIC-KITCHENS VISOR

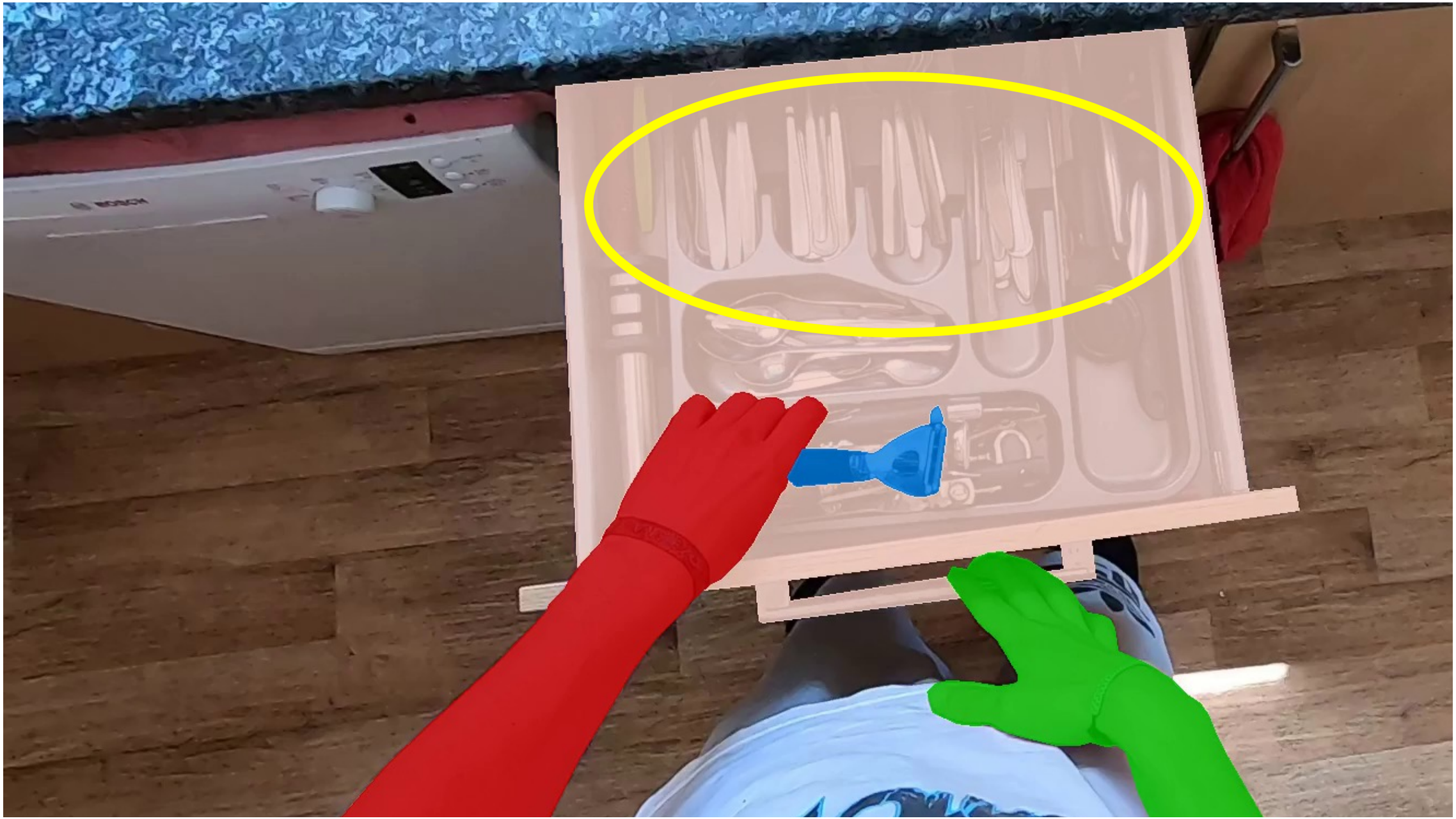
with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma,
Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler

EPIC-KITCHENS VISOR

with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler

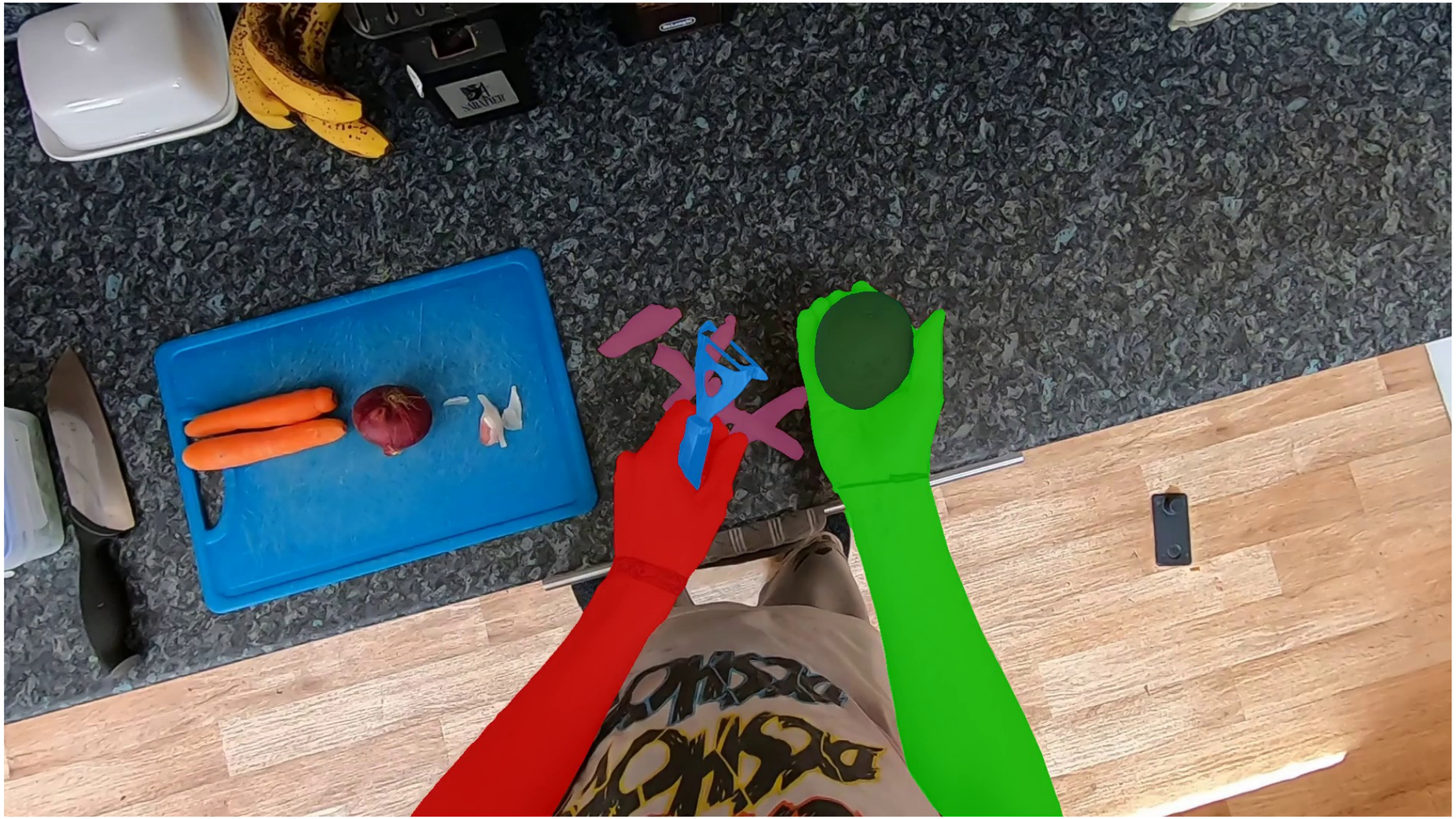


A Darkhalil et al (2022). EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. *NeurIPS*



EPIC-KITCHENS VISOR

with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler



University of BRISTOL

A Darkhalil et al (2022). EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. *NeurIPS*

EPIC-KITCHENS VISOR

with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler

A Darkhalil et al (2022). EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. *NeurIPS*

EPIC-KITCHENS VISOR

with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma,
Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler

A Darkhalil et al (2022). EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. *NeurIPS*

EPIC-KITCHENS VISOR

with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler

A Darkhalil et al (2022). EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. *NeurIPS*

EPIC-KITCHENS VISOR

with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler

A Darkhalil et al (2022). EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. *NeurIPS*

EPIC-KITCHENS VISOR

with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler

A Darkhalil et al (2022). EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. *NeurIPS*

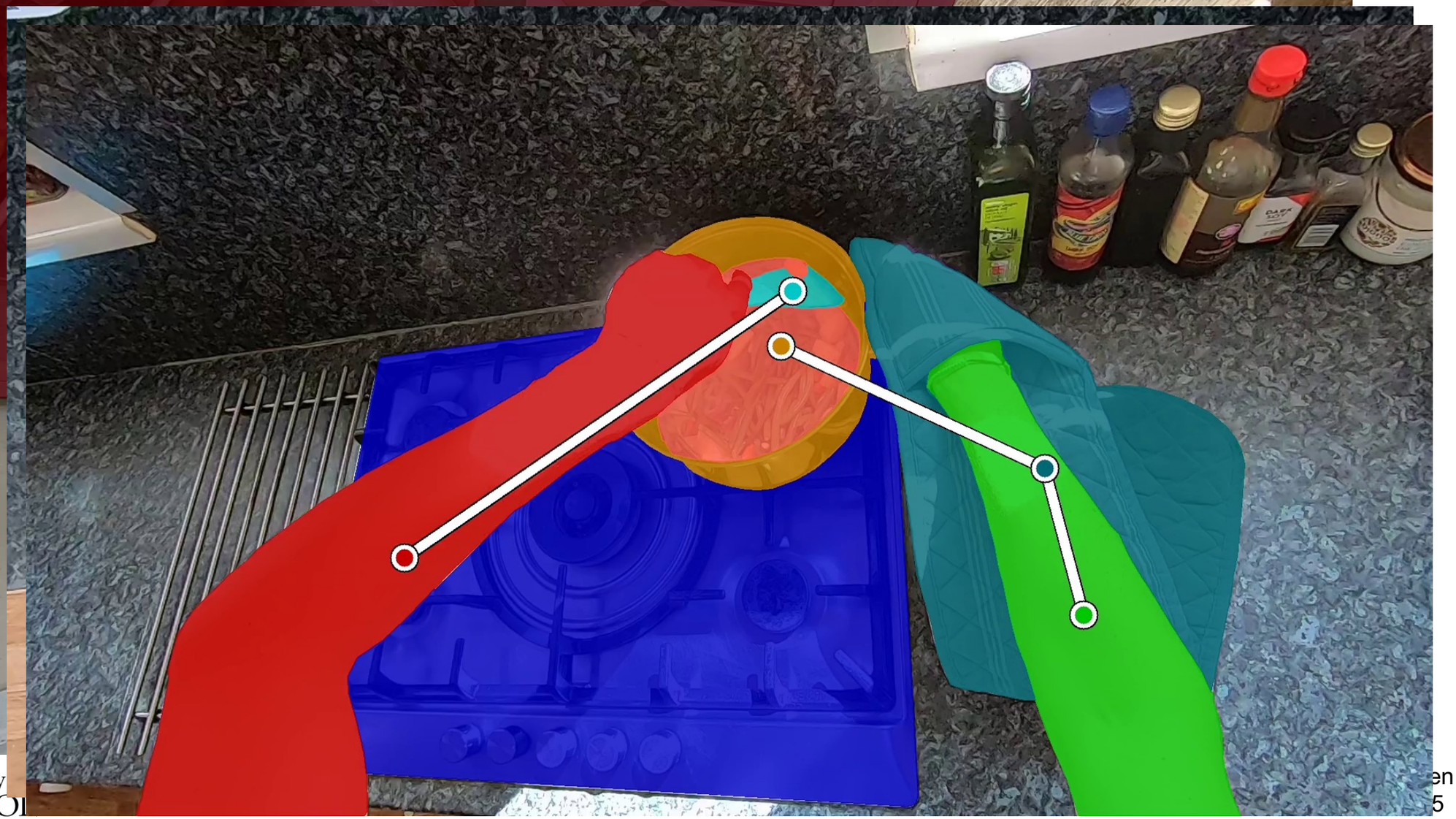
EPIC-KITCHENS VISOR

with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler

A Darkhalil et al (2022). EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. *NeurIPS*

VISOR Relations

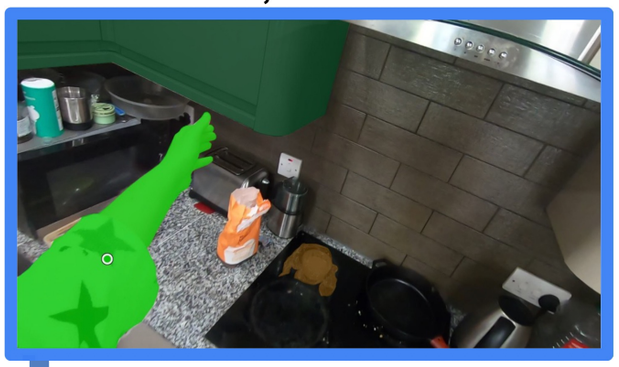
with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler, Dima Damen



Object relation stats

with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler, Dima Damen

1 Hand, No Contact



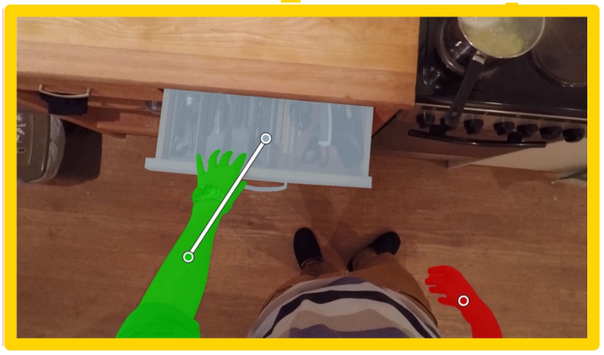
2 Hands, No Contact

1 Hand, In Contact

2.7% 41.5% 0.7% 19.4% 27.2% 8.5%

2 Hands, 2 Obj Contacts

2 Hands, Same Contact



2 Hands, 1 In Contact

A Darkhalil et al (2022). EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. *NeurIPS*

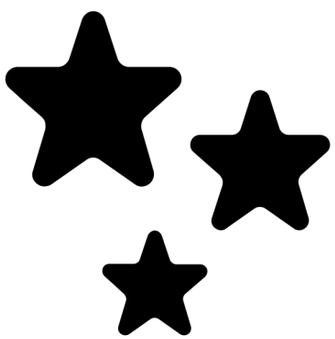
EPIC-KITCHENS VISOR

with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler

Analogous Tasks

Image

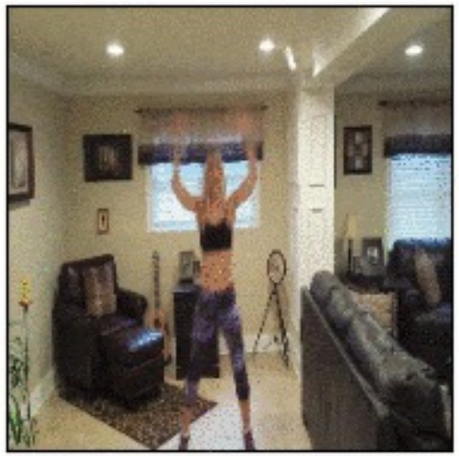
- Object Counting

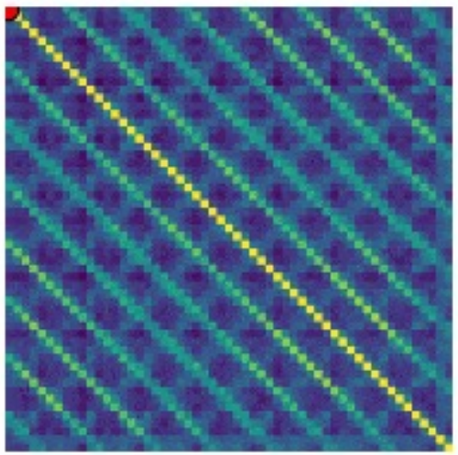
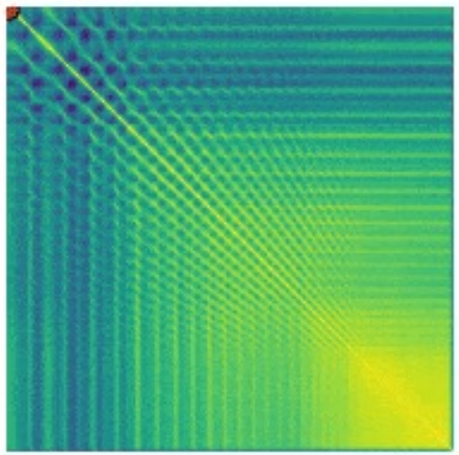
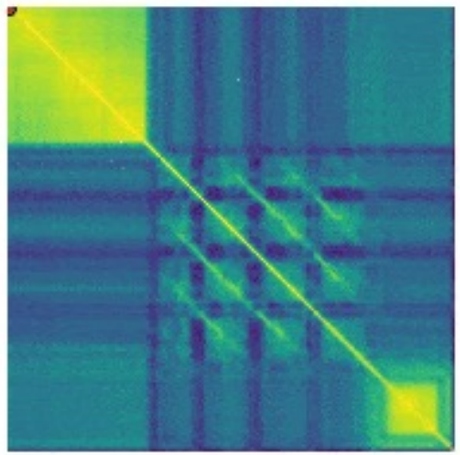


Video

- Action Counting

Countix



D Dwibedi et al (2020). Counting Out Time: Class Agnostic Video Repetition Counting in the Wild. CVPR

Every Shot Counts

with: Saptarshi Sinha
Alexandros Stergiou

RepCount

GT:6

Pred:6

Countix

GT:9

Pred:9

RepCount

GT:32

Pred:32

Analogous Tasks

Image

- Text-to-image Generation

Stable Diffusion

Video

- Text-to-Video Generation

SORA

Text-to-Video Generation

Prompt: A grandmother with neatly combed grey hair stands behind a colorful birthday cake with numerous candles at a wood dining room table, expression is one of pure joy and happiness, with a happy glow in her eye. She leans forward and blows out...

Gemini Veo 3 – May 2025

Gemini Veo 3 – May 2025

Dima Damen
PAISS 2025

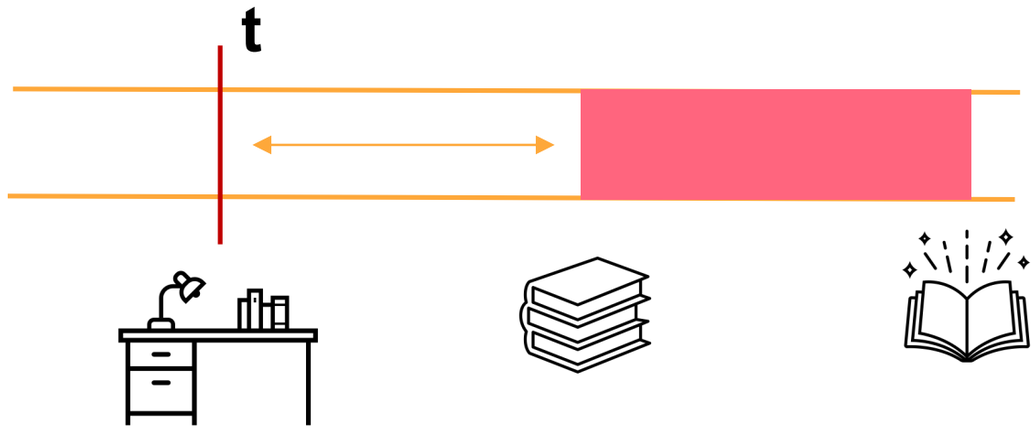
Non-Analogous Tasks

Image

Video

- Action Anticipation

What will happen after 1 second?



Open Book

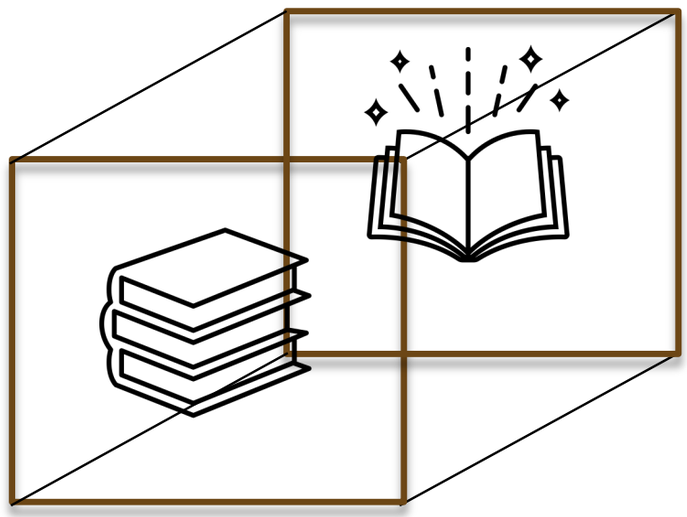
Non-Analogous Tasks

Image

Video

- Manner of action

How did you open the book?



How?

with: Hazel Doughty
Ivan Laptev
Walterio Mayol-Cuevas

... if you **turn** the bowl upside down **slowly** they won't come out ...

... mix it well until it is **completely dissolved** ...

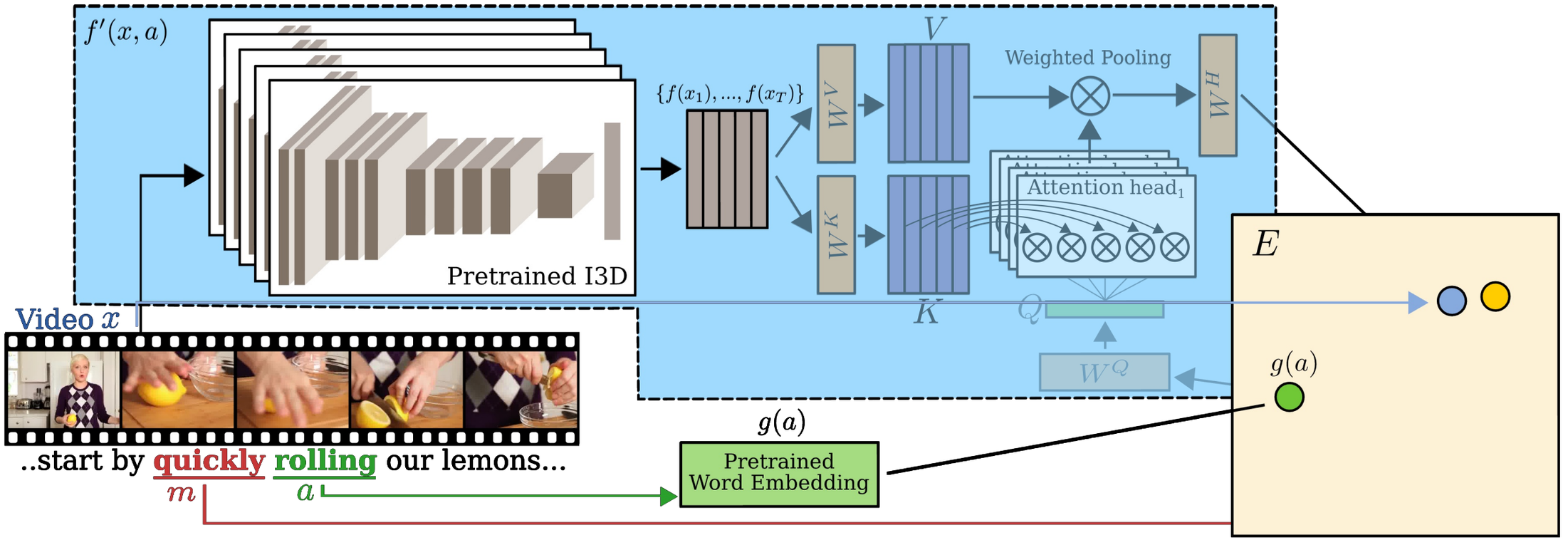
... you want to make sure you **fill** it up **partially** ...

... you want to **dice** it **finely**...

-10 seconds | timestamp | +10 seconds

How?

with: Hazel Doughty
Ivan Laptev
Walterio Mayol-Cuevas



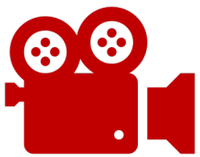
How?

with: Hazel Doughty
Ivan Laptev
Walterio Mayol-Cuevas

... we're going to mix these up real **quick**...

In today's tutorial

Motivation and Datasets in Egocentric Video Understanding



Video Understanding: Data and Tasks

Videos are Multimodal

Connected Videos of One's Life

Video Understanding Out of the Frame

Teaser: The Wizard of Oz & Genie 3

Outlook into the Future of Egocentric Vision

Conclusion


Videos are multi-modal

Multi-modal learning…

with: Vangelis Kazakos Jaesung Huh
Arsha Nagrani. Jacob Chalk
Andrew Zisserman

- The magic of audio-visual understanding…
- Object-Object interactions

Multi-modal learning…

with: Vangelis Kazakos Jaesung Huh
Arsha Nagrani. Jacob Chalk
Andrew Zisserman

- The magic of audio-visual understanding…
- Object-Object interactions
- Material sounds

Multi-modal learning…

with: Vangelis Kazakos Jaesung Huh
Arsha Nagrani. Jacob Chalk
Andrew Zisserman

- The magic of audio-visual understanding…
- Object-Object interactions
- Material sounds
- Sound-emitting objects

ICASSP 2023
4 - 10 JUNE,
RHODES ISLAND, GREECE

EPIC-Sounds: A Large-scale Dataset of Actions That Sound

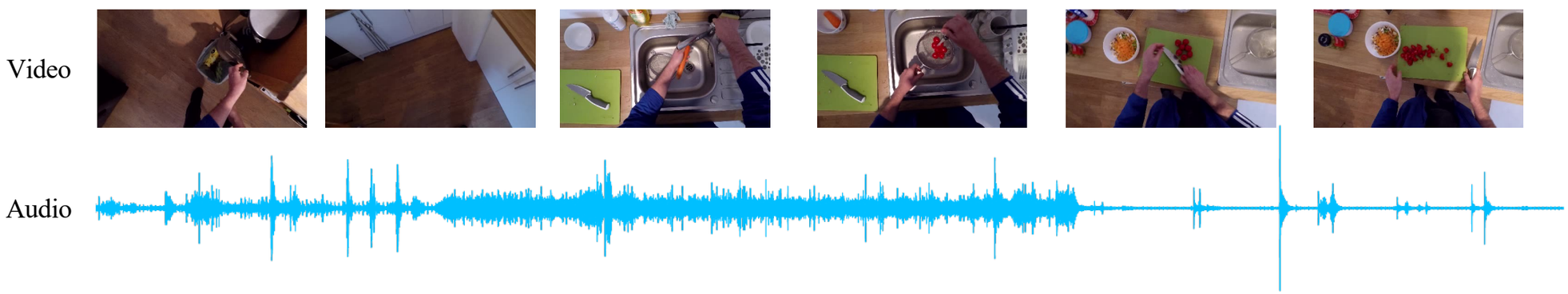
Jaesung Huh*, Jacob Chalk*, Evangelos Kazakos, Dima Damen, Andrew Zisserman
* : Equal contribution

VGG
University of Oxford

University of BRISTOL

Motivation

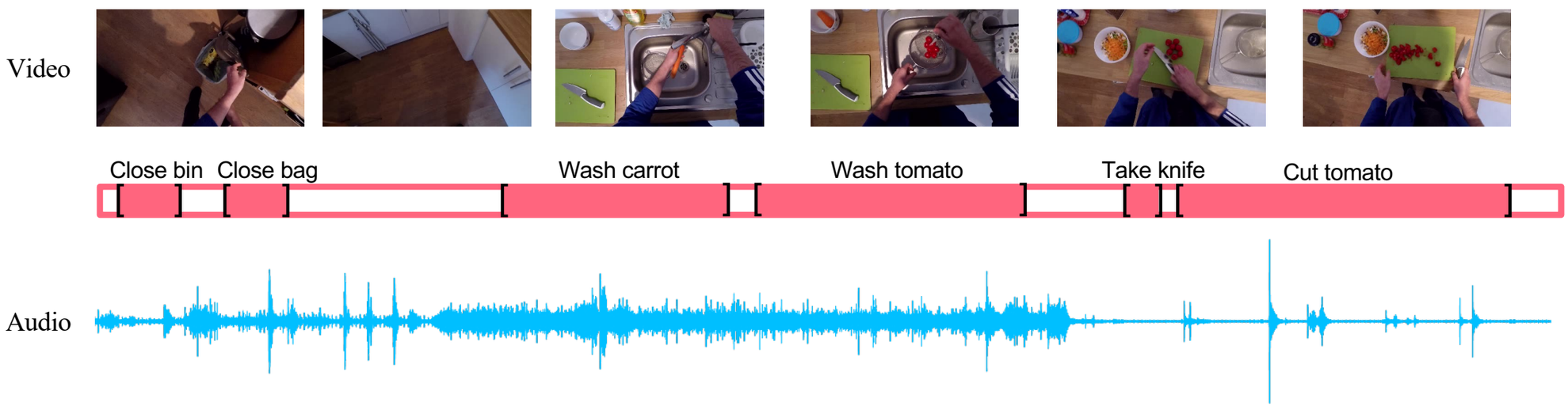
with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman



J Huh*, J Chalk* E Kazakos, D Damen, A Zisserman (2023). EPIC-Sounds: A Large-scale Dataset of Actions That Sound. ICASSP

Motivation

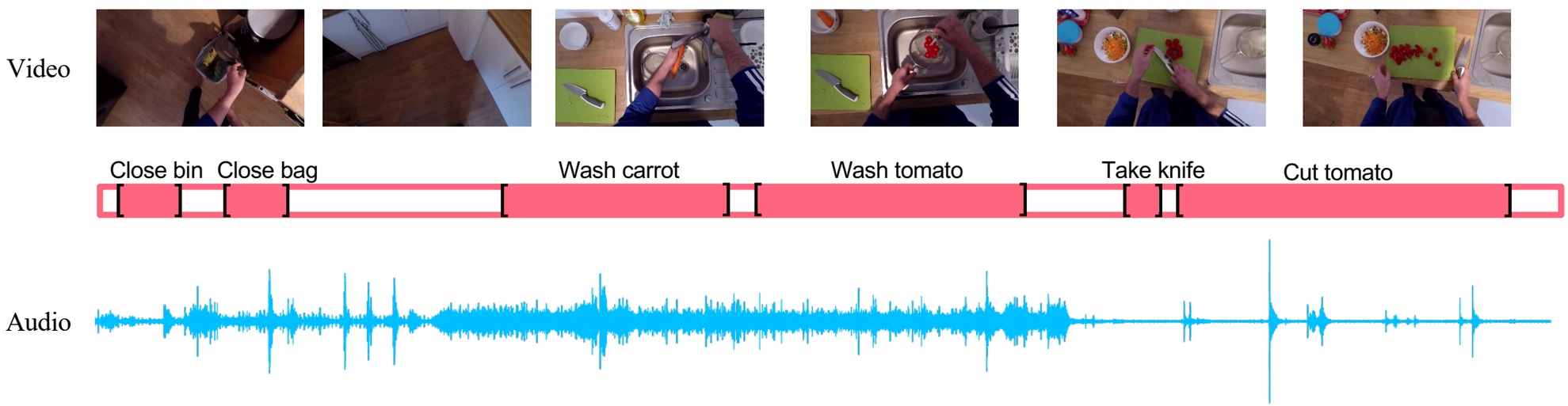
with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman



J Huh*, J Chalk* E Kazakos, D Damen, A Zisserman (2023). EPIC-Sounds: A Large-scale Dataset of Actions That Sound. ICASSP

Motivation

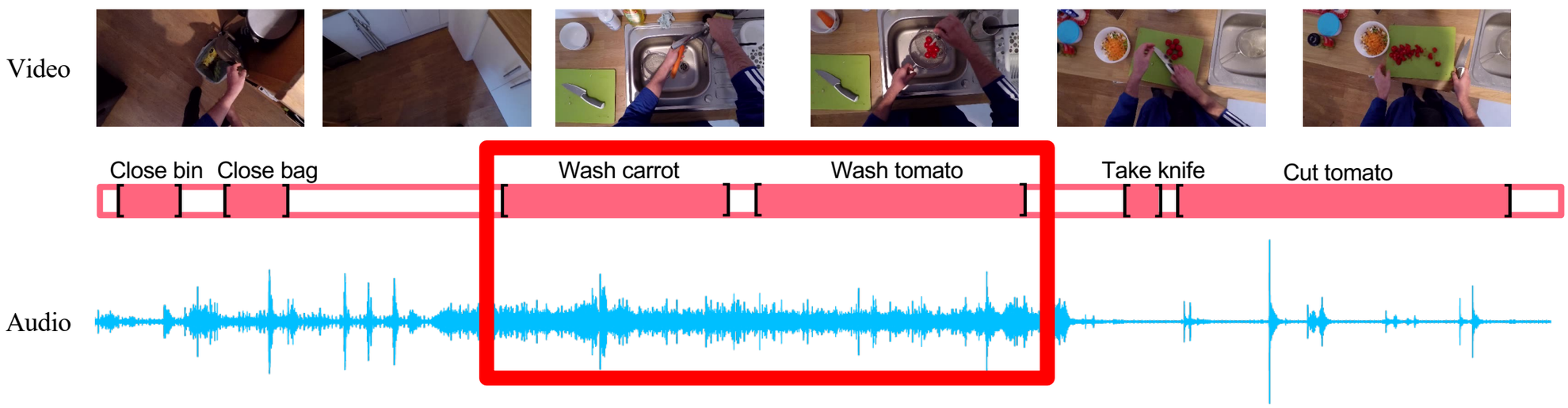
with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman



J Huh*, J Chalk* E Kazakos, D Damen, A Zisserman (2023). EPIC-Sounds: A Large-scale Dataset of Actions That Sound. ICASSP

Motivation

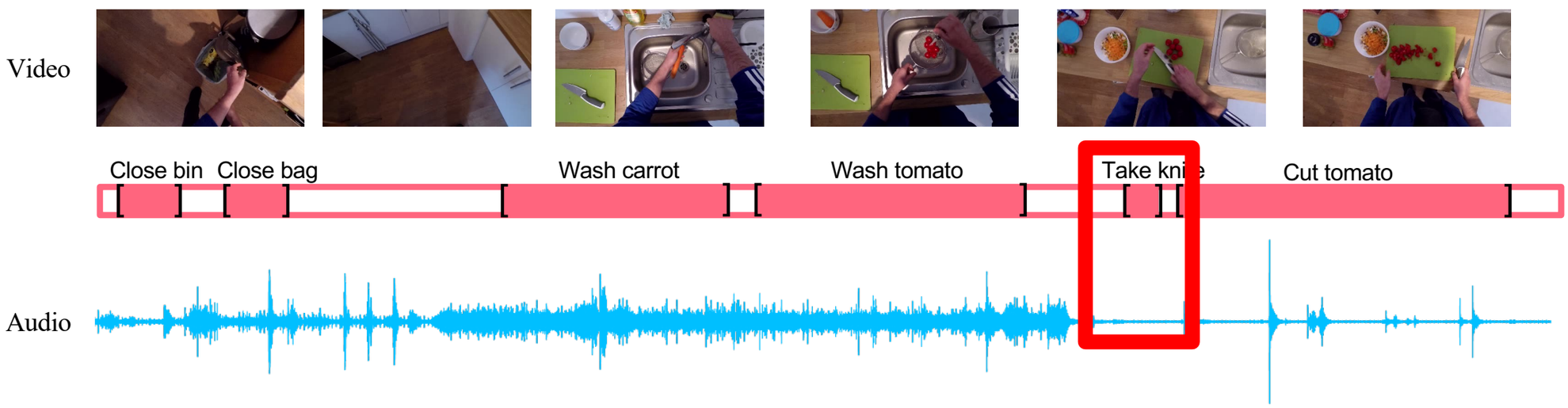
with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman



J Huh*, J Chalk* E Kazakos, D Damen, A Zisserman (2023). EPIC-Sounds: A Large-scale Dataset of Actions That Sound. ICASSP

Motivation

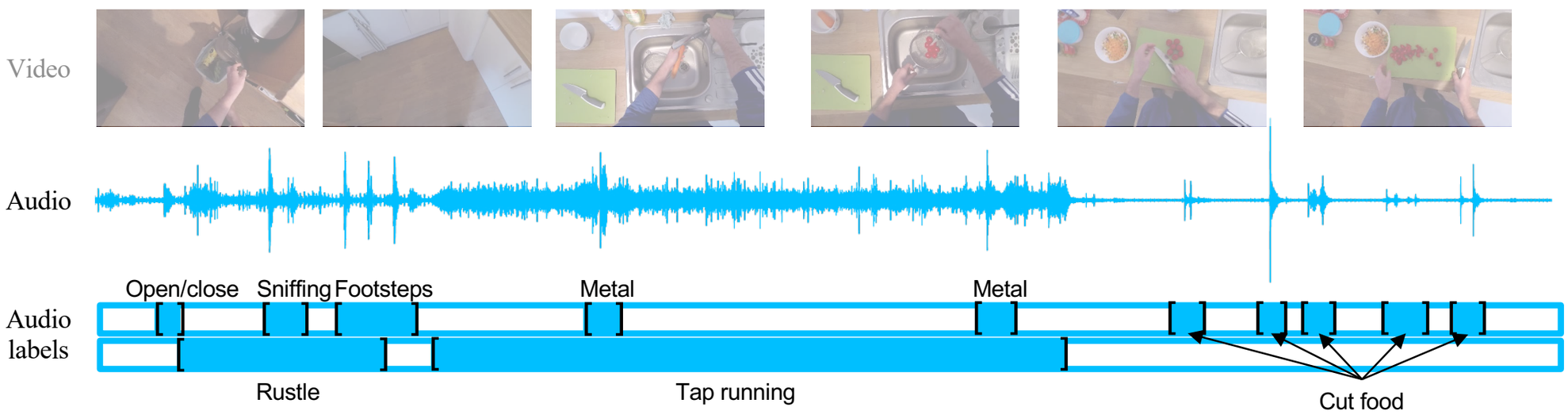
with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman



J Huh*, J Chalk* E Kazakos, D Damen, A Zisserman (2023). EPIC-Sounds: A Large-scale Dataset of Actions That Sound. ICASSP

Motivation

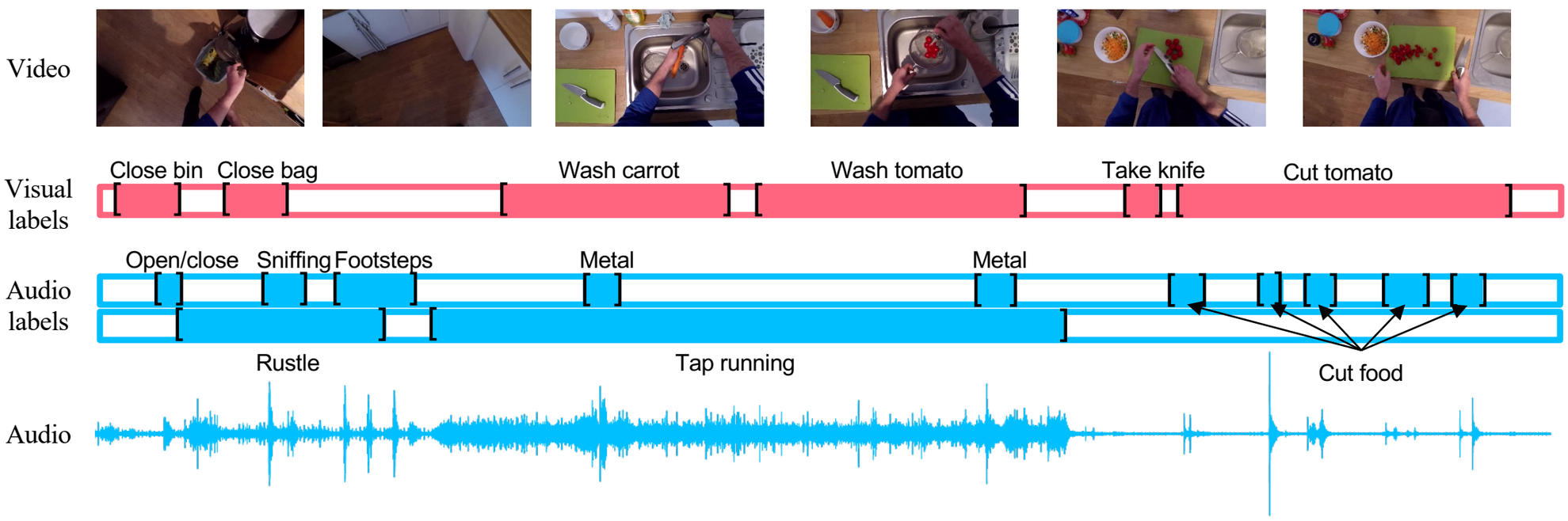
with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman



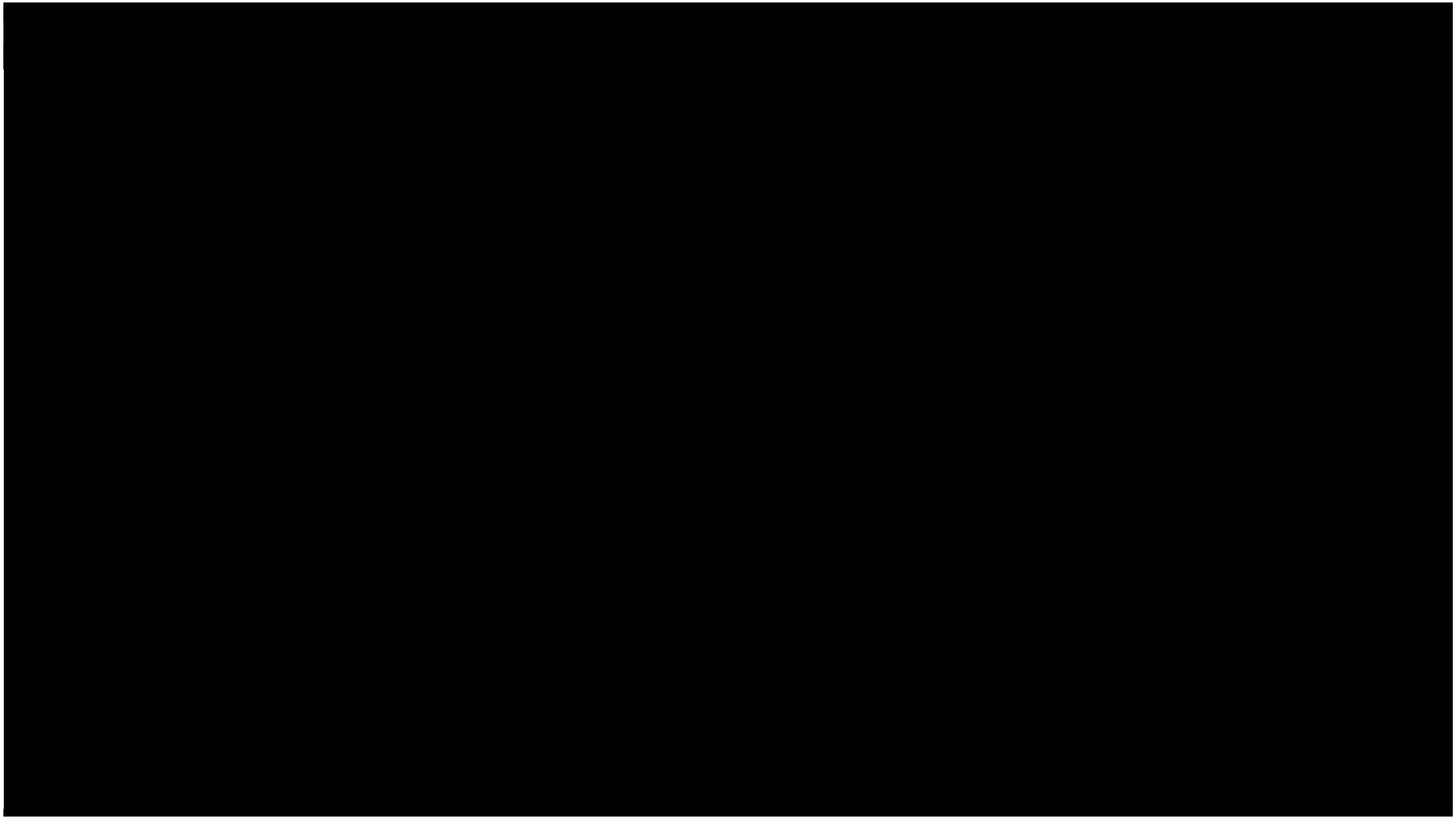
J Huh*, J Chalk* E Kazakos, D Damen, A Zisserman (2023). EPIC-Sounds: A Large-scale Dataset of Actions That Sound. ICASSP

Motivation

with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman



J Huh*, J Chalk* E Kazakos, D Damen, A Zisserman (2023). EPIC-Sounds: A Large-scale Dataset of Actions That Sound. ICASSP



EPIC-SOUNDS

with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman

EPIC-KITCHENS VIDEOS
100 hours
45 kitchens

Visual Action Annotations
90K visual actions
97 verb classes
300 noun classes

EPIC-Sounds
Audio-Based Annotations
79K categorised audio events
44 sound categories
39K uncategorised events

J Huh*, J Chalk* E Kazakos, D Damen, A Zisserman (2023). EPIC-Sounds: A Large-scale Dataset of Actions That Sound. ICASSP


spray

CVPR
JUNE 17-21, 2024
SEATTLE, WA

TIM: A Time Interval Machine for Audio-Visual Action Recognition

Jacob Chalk*, Jaesung Huh*, Evangelos Kazakos, Andrew Zisserman, Dima Damen
* : Equal contribution

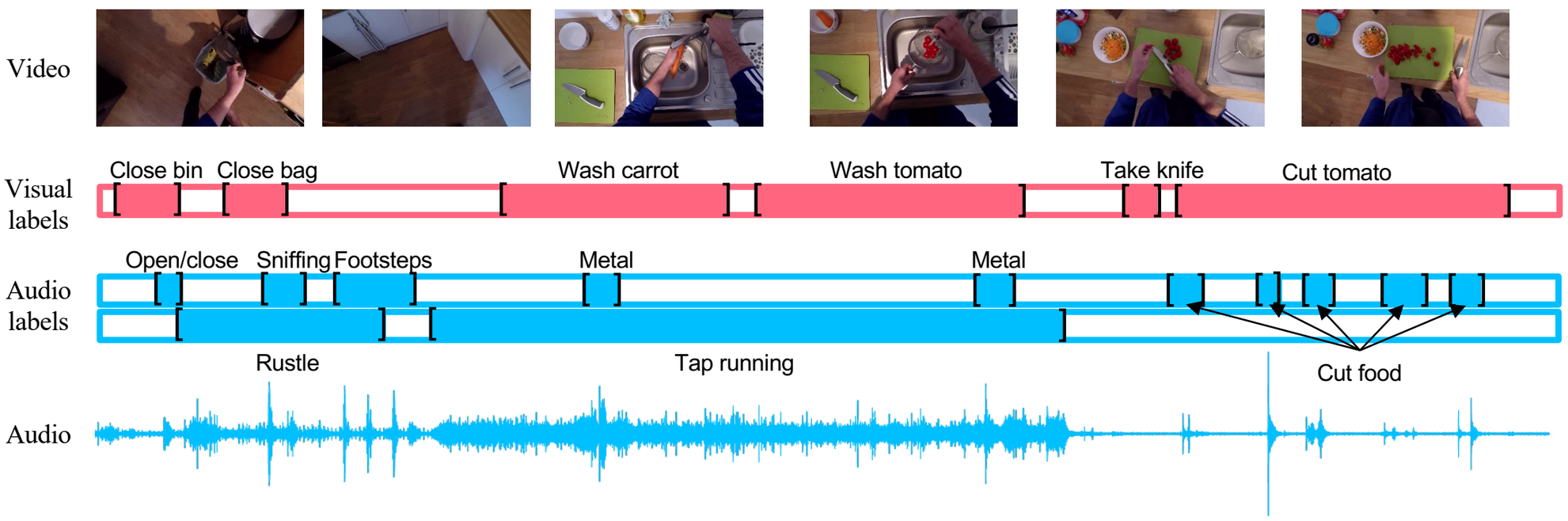
University of BRISTOL

VGG
University of Oxford

CZECH TECHNICAL UNIVERSITY IN PRAGUE

Multi-Modal Long-Form Dataset

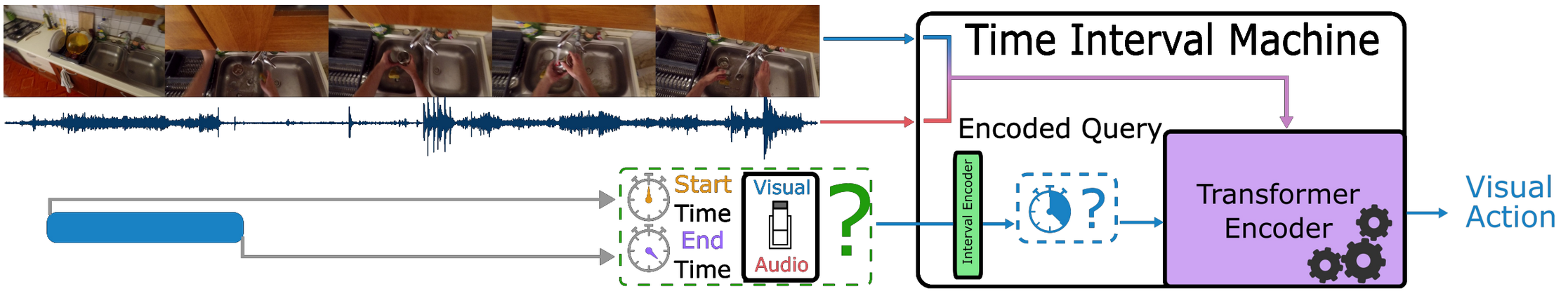
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



J Chalk*, J Huh*, E Kazakos, A Zisserman, D Damen (2024). TIM: A Time Interval Machine for Audio-Visual Action Recognition. CVPR

TIM: A Time-Interval Audio-Visual Machine

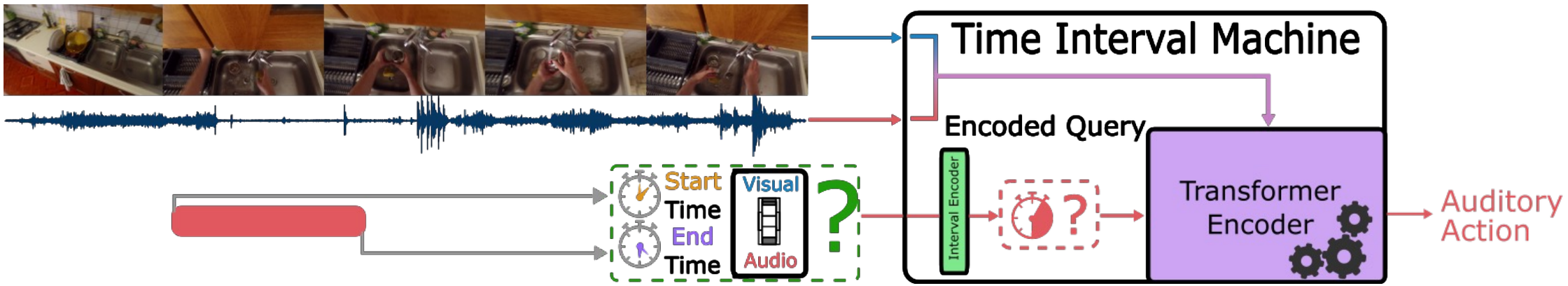
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



J Chalk*, J Huh*, E Kazakos, A Zisserman, D Damen (2024). TIM: A Time Interval Machine for Audio-Visual Action Recognition. CVPR

TIM: A Time-Interval Audio-Visual Machine

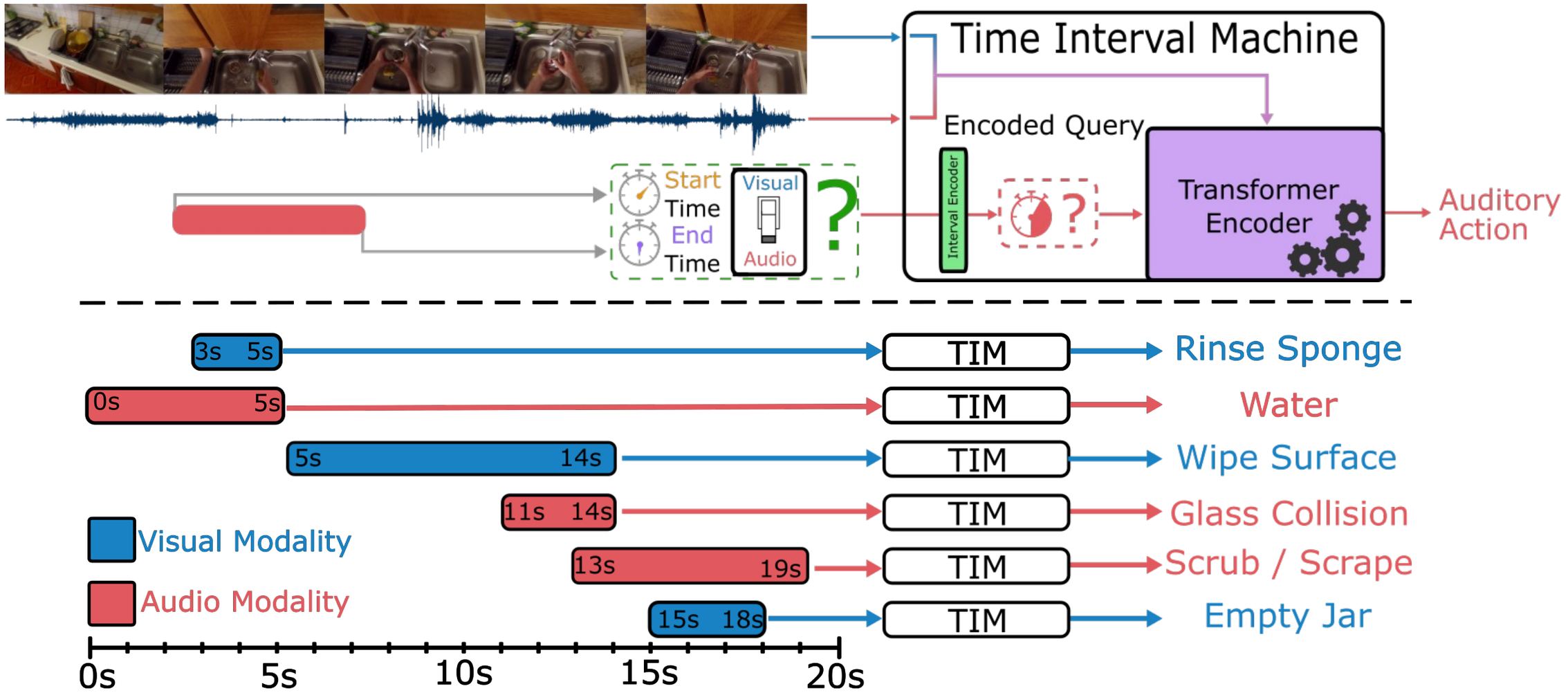
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



J Chalk*, J Huh*, E Kazakos, A Zisserman, D Damen (2024). TIM: A Time Interval Machine for Audio-Visual Action Recognition. CVPR

TIM: A Time-Interval Audio-Visual Machine

with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman

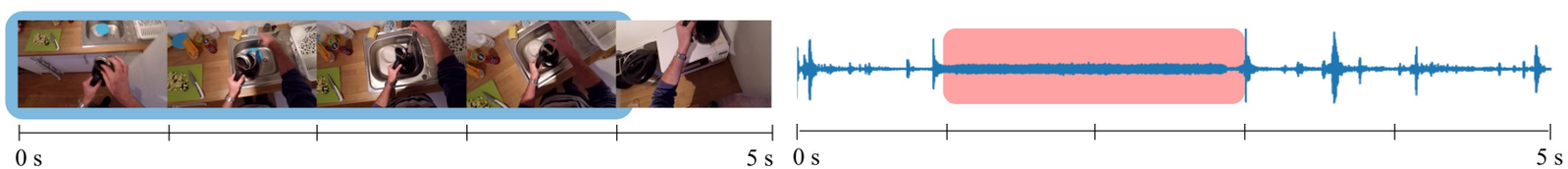


University of BRISTOL

J Chalk*, J Huh*, E Kazakos, A Zisserman, D Damen (2024). TIM: A Time Interval Machine for Audio-Visual Action Recognition. CVPR

TIM: A Time-Interval Audio-Visual Machine

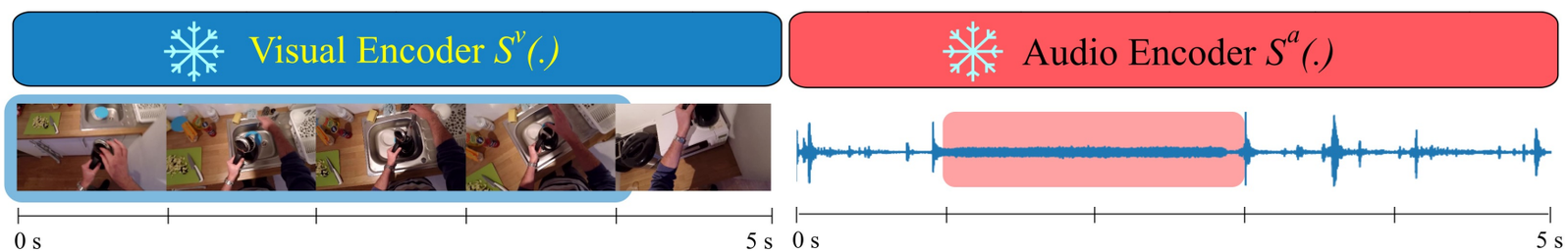
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



J Chalk*, J Huh*, E Kazakos, A Zisserman, D Damen (2024). TIM: A Time Interval Machine for Audio-Visual Action Recognition. CVPR

TIM: A Time-Interval Audio-Visual Machine

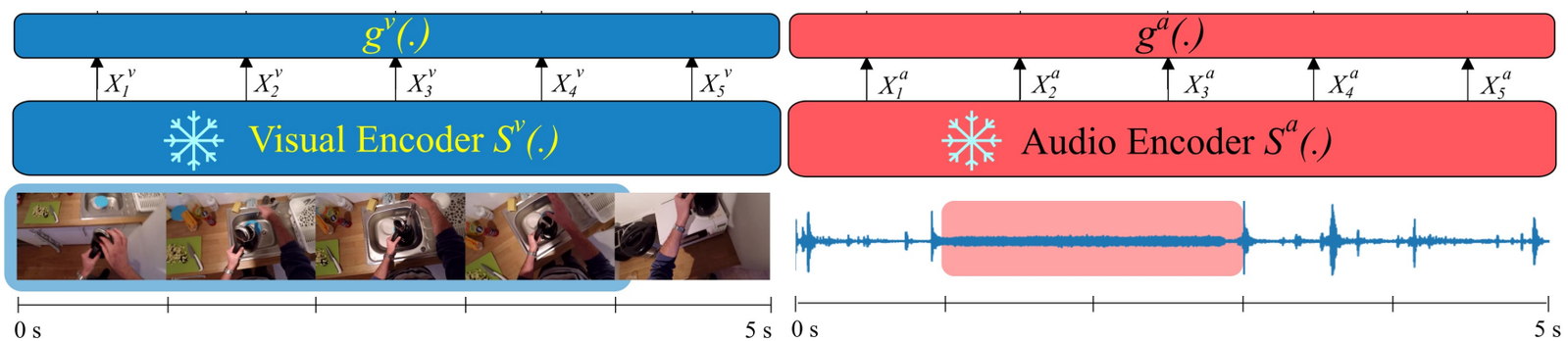
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



J Chalk*, J Huh*, E Kazakos, A Zisserman, D Damen (2024). TIM: A Time Interval Machine for Audio-Visual Action Recognition. CVPR

TIM: A Time-Interval Audio-Visual Machine

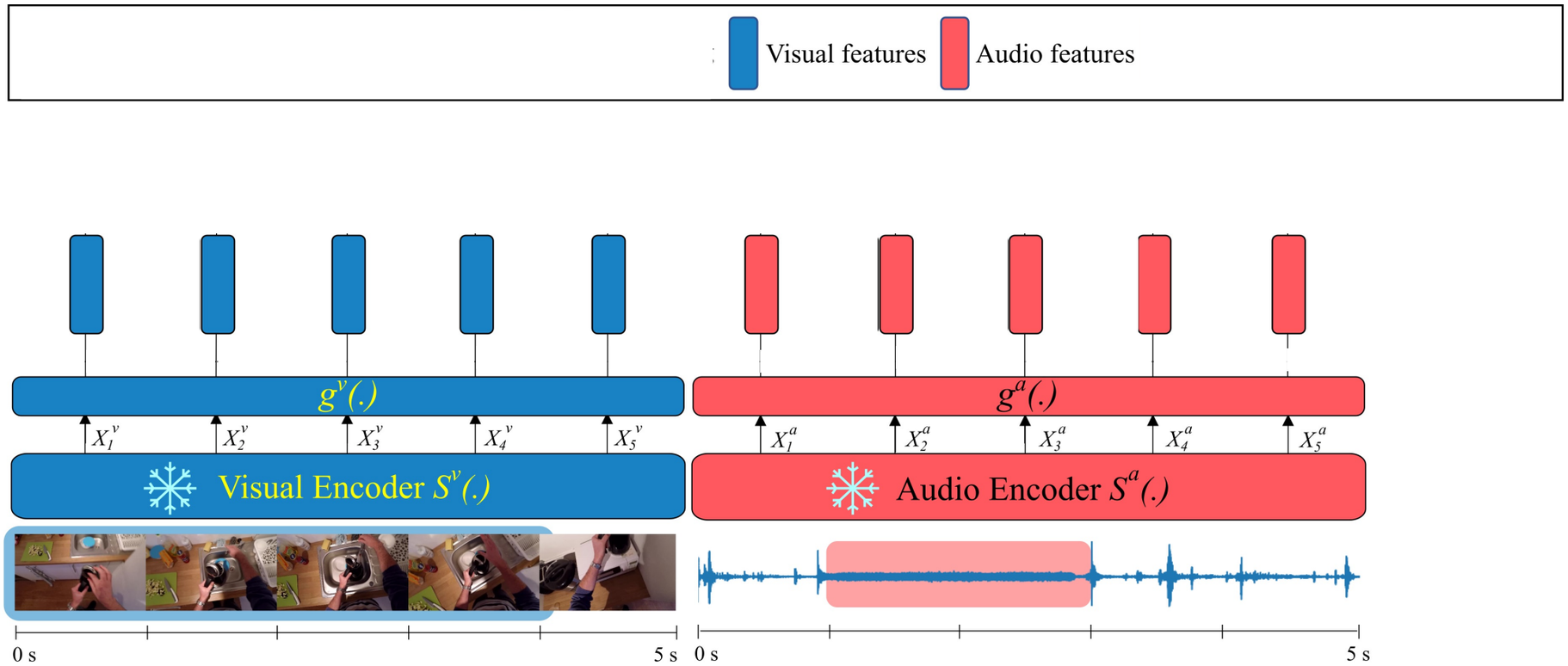
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



J Chalk*, J Huh*, E Kazakos, A Zisserman, D Damen (2024). TIM: A Time Interval Machine for Audio-Visual Action Recognition. CVPR

TIM: A Time-Interval Audio-Visual Machine

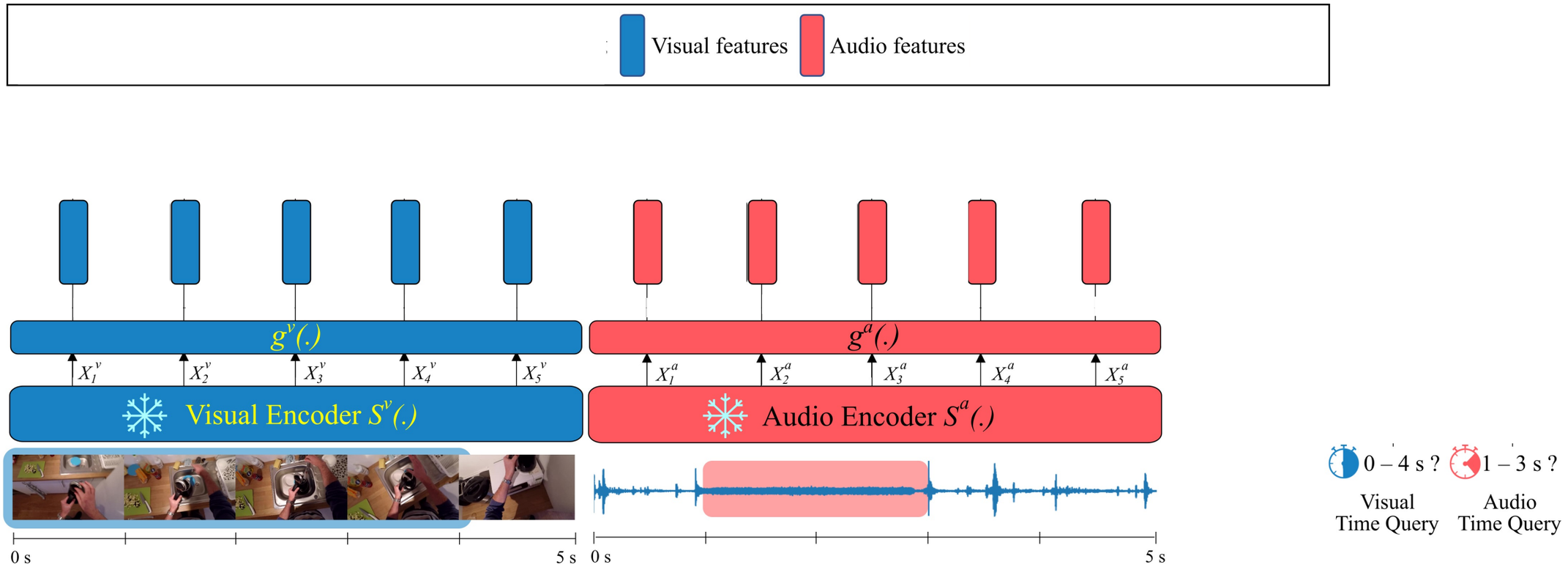
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



J Chalk*, J Huh*, E Kazakos, A Zisserman, D Damen (2024). TIM: A Time Interval Machine for Audio-Visual Action Recognition. CVPR

TIM: A Time-Interval Audio-Visual Machine

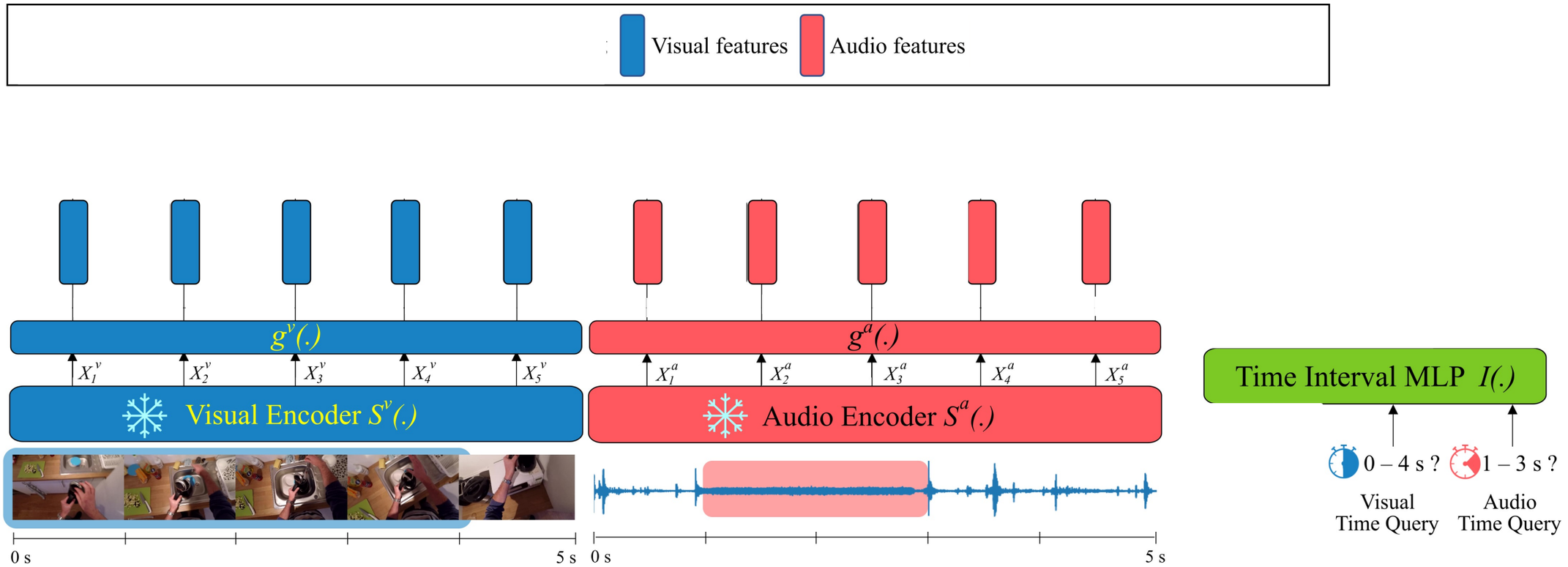
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



J Chalk*, J Huh*, E Kazakos, A Zisserman, D Damen (2024). TIM: A Time Interval Machine for Audio-Visual Action Recognition. CVPR

TIM: A Time-Interval Audio-Visual Machine

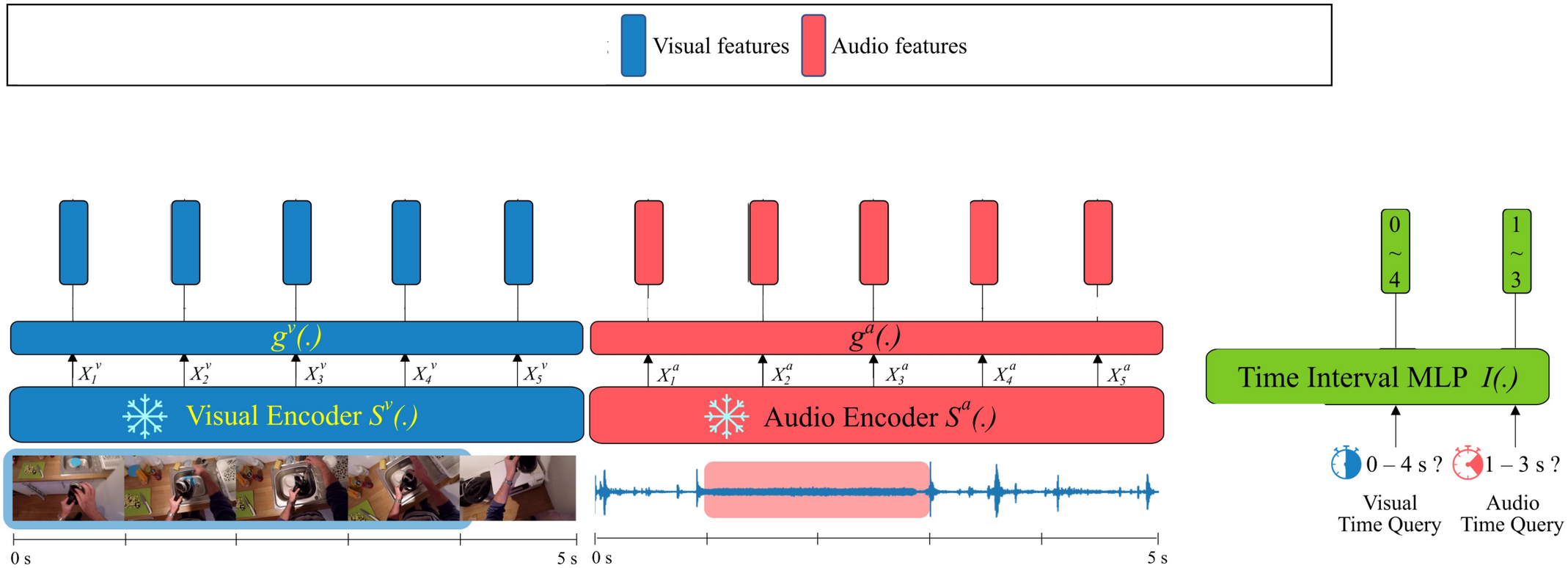
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



J Chalk*, J Huh*, E Kazakos, A Zisserman, D Damen (2024). TIM: A Time Interval Machine for Audio-Visual Action Recognition. CVPR

TIM: A Time-Interval Audio-Visual Machine

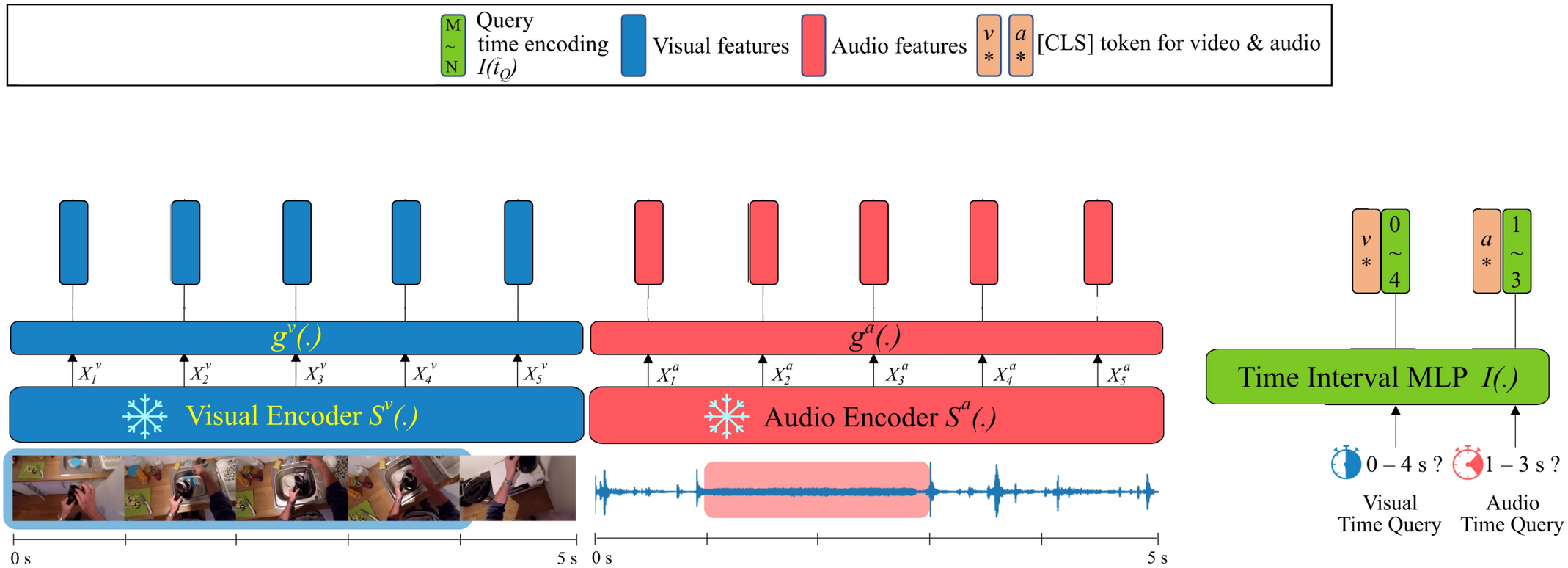
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



J Chalk*, J Huh*, E Kazakos, A Zisserman, D Damen (2024). TIM: A Time Interval Machine for Audio-Visual Action Recognition. CVPR

TIM: A Time-Interval Audio-Visual Machine

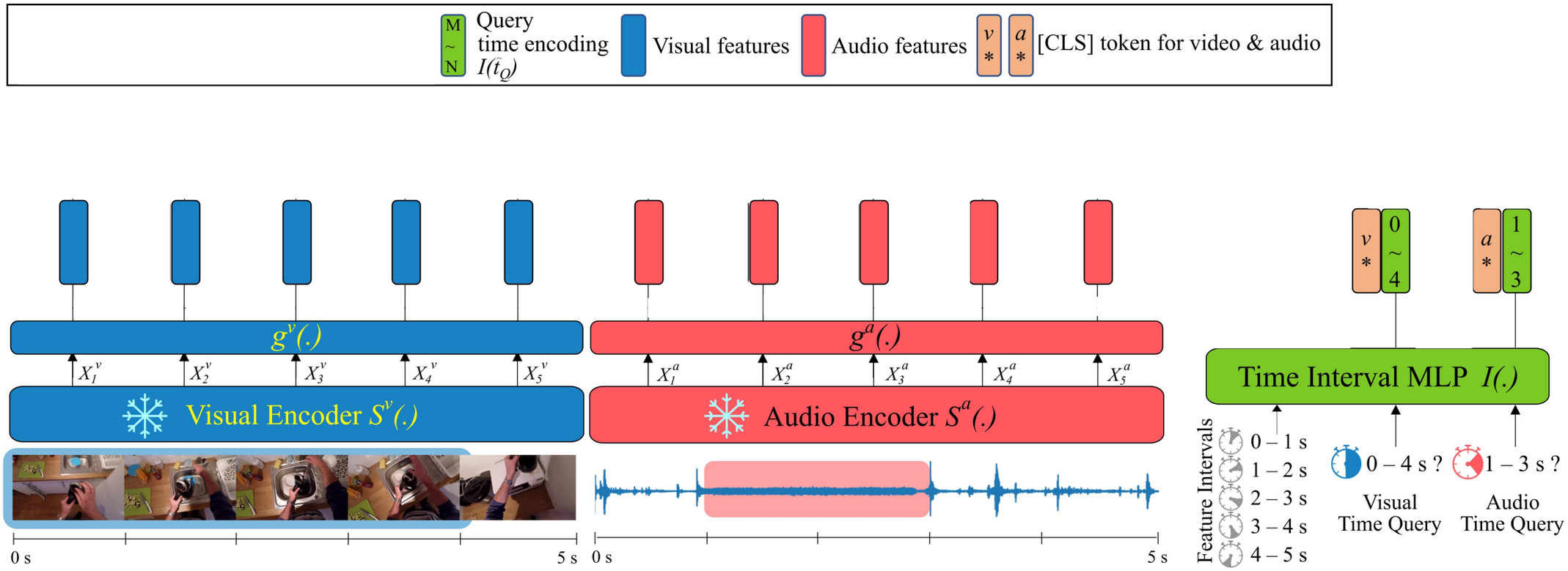
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



J Chalk*, J Huh*, E Kazakos, A Zisserman, D Damen (2024). TIM: A Time Interval Machine for Audio-Visual Action Recognition. CVPR

TIM: A Time-Interval Audio-Visual Machine

with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman

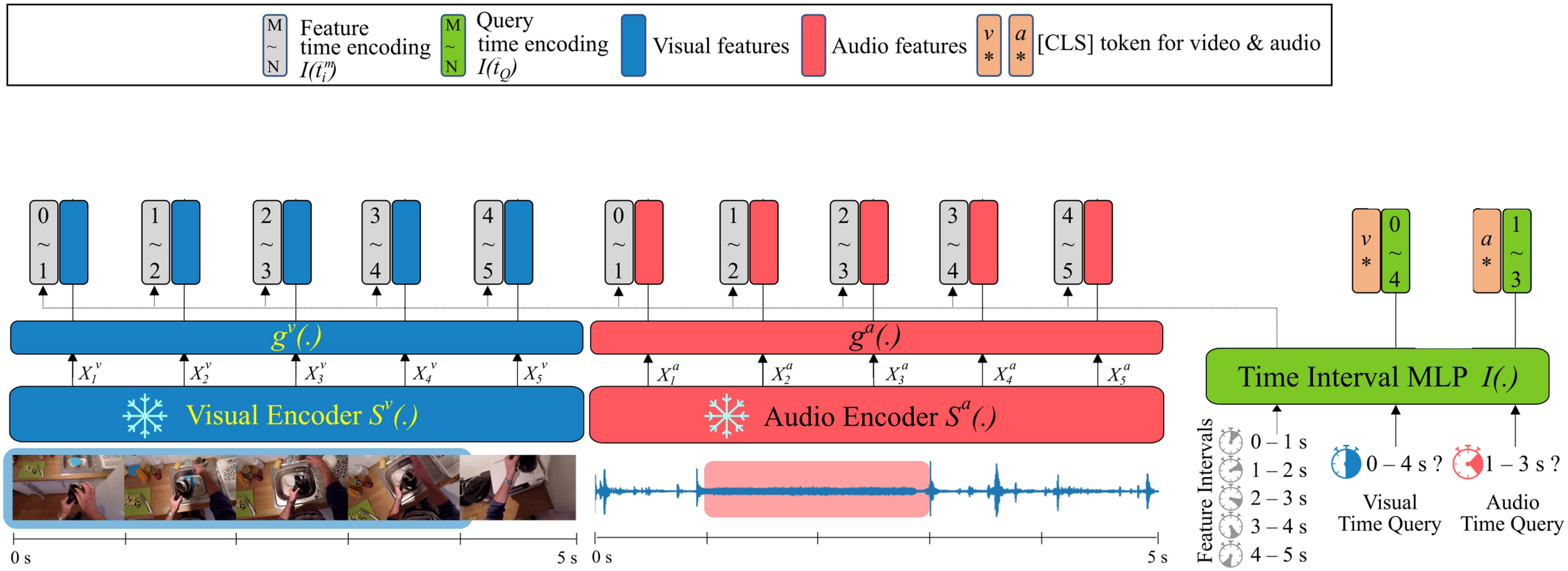


University of BRISTOL

J Chalk*, J Huh*, E Kazakos, A Zisserman, D Damen (2024). TIM: A Time Interval Machine for Audio-Visual Action Recognition. CVPR

TIM: A Time-Interval Audio-Visual Machine

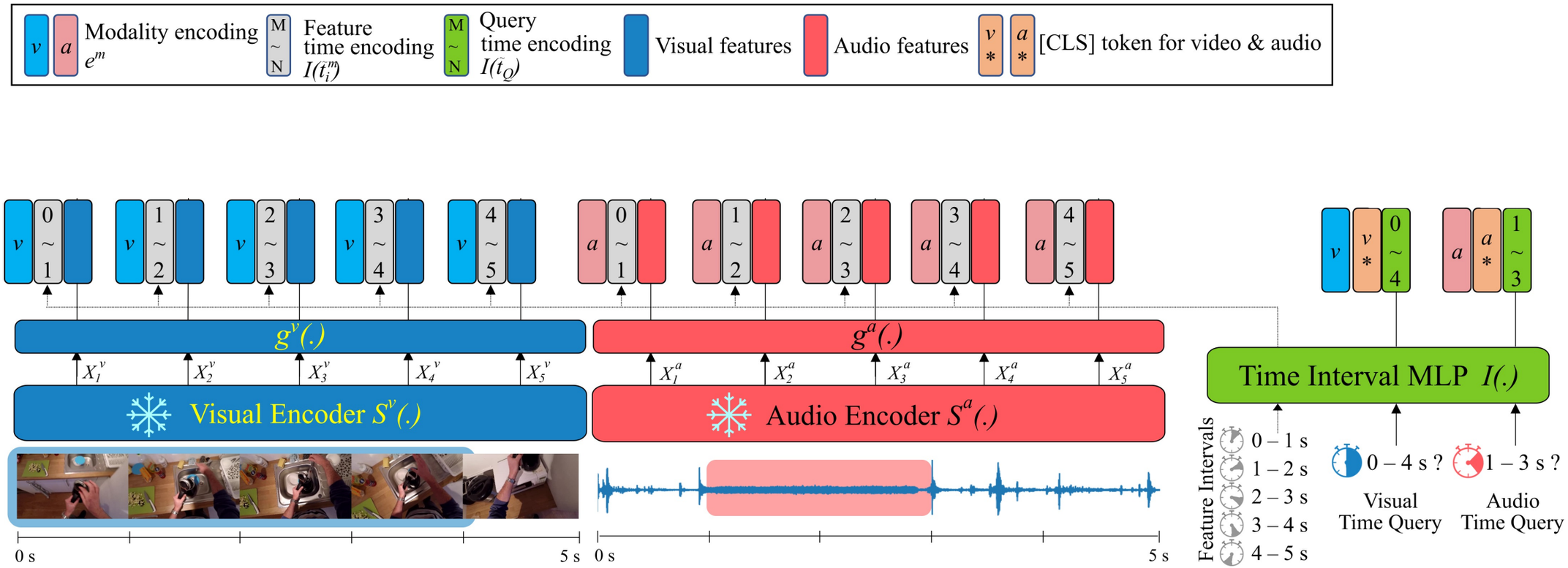
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



J Chalk*, J Huh*, E Kazakos, A Zisserman, D Damen (2024). TIM: A Time Interval Machine for Audio-Visual Action Recognition. CVPR

TIM: A Time-Interval Audio-Visual Machine

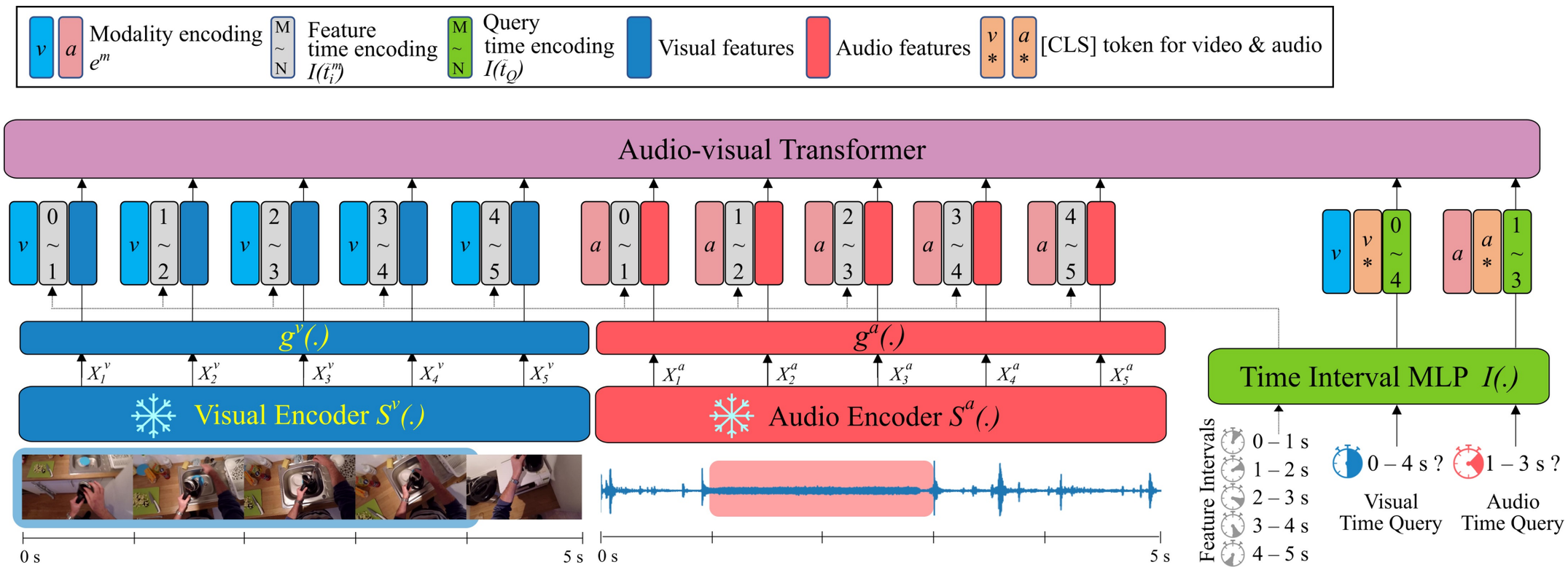
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



J Chalk*, J Huh*, E Kazakos, A Zisserman, D Damen (2024). TIM: A Time Interval Machine for Audio-Visual Action Recognition. CVPR

TIM: A Time-Interval Audio-Visual Machine

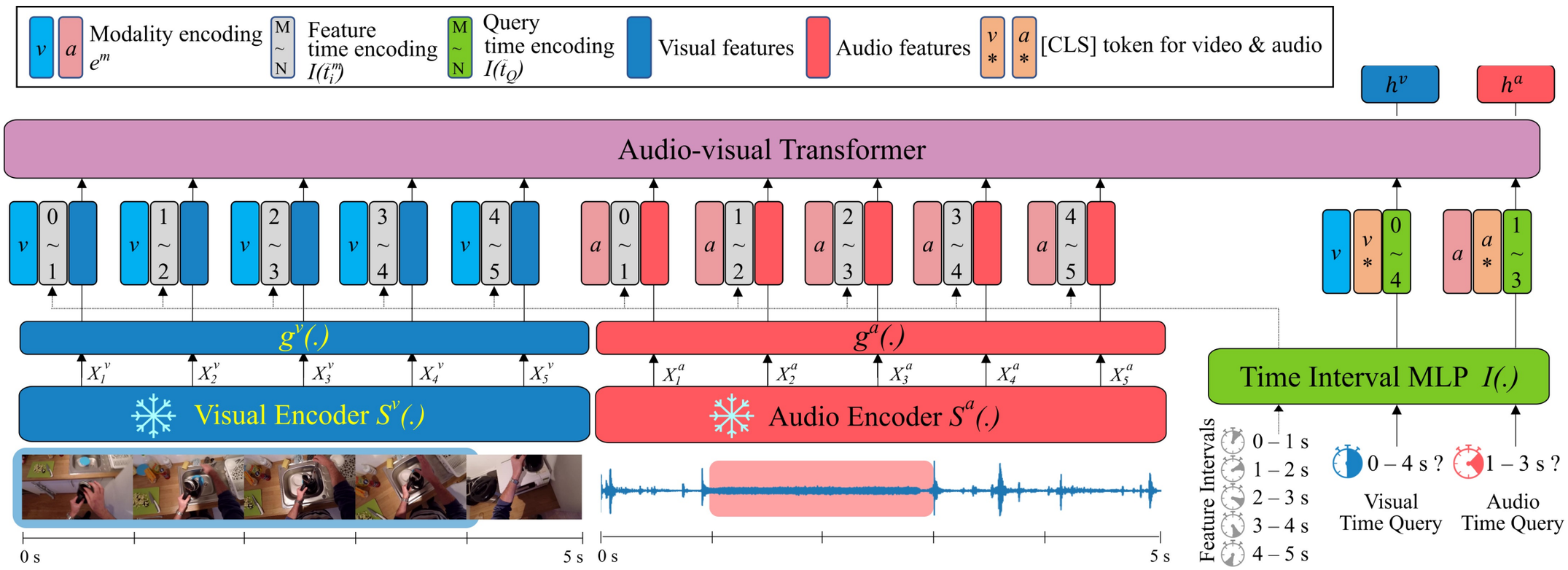
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



J Chalk*, J Huh*, E Kazakos, A Zisserman, D Damen (2024). TIM: A Time Interval Machine for Audio-Visual Action Recognition. CVPR

TIM: A Time-Interval Audio-Visual Machine

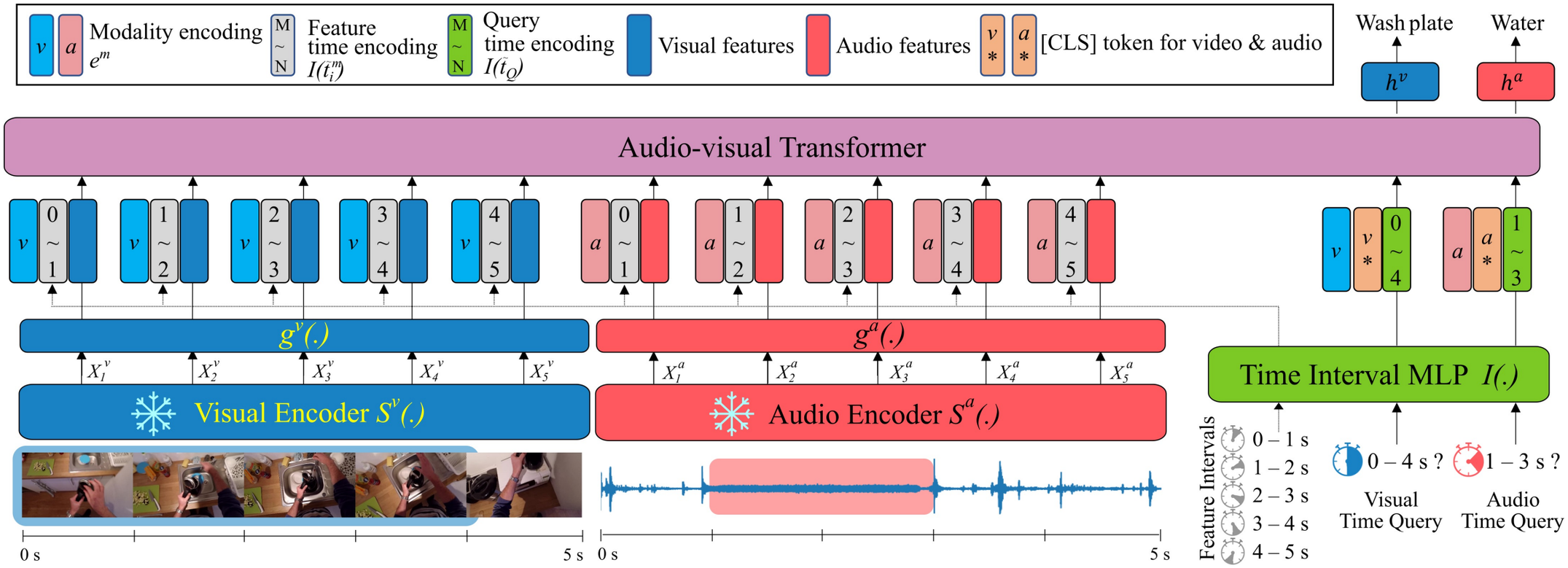
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



J Chalk*, J Huh*, E Kazakos, A Zisserman, D Damen (2024). TIM: A Time Interval Machine for Audio-Visual Action Recognition. CVPR

TIM: A Time-Interval Audio-Visual Machine

with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



J Chalk*, J Huh*, E Kazakos, A Zisserman, D Damen (2024). TIM: A Time Interval Machine for Audio-Visual Action Recognition. CVPR

TIM: A Time-Interval Audio-Visual Machine

with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman

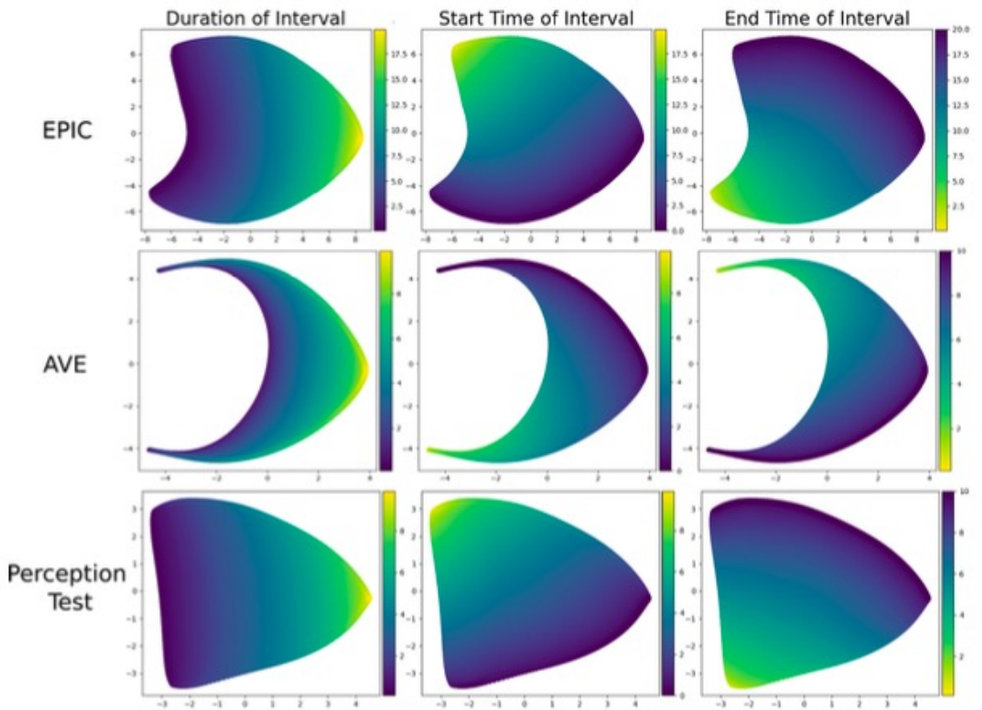
Model	*xp*	LLM	Verb	Noun	Action
Visual-only models					
MFormer-HR [37]	336p	✗	67.0	58.5	44.5
MoViNet-A6 [27]	320p	✗	72.2	57.3	47.7
MeMViT [55]	224p	✗	71.4	60.3	48.4
Omnivore [14]	224p	✗	69.5	61.7	49.9
MTV [59]	280p	✗	69.9	63.9	50.5
LaViLa (TSF-L) [63]	224p	✓	72.0	62.9	51.0
AVION (ViT-L) [62]	224p	✓	73.0	65.4	54.4
TIM (ours)	224p	✗	**76.2**	**66.4**	**56.4**
Audio-visual models					
TBN [24]	224p	✗	66.0	47.2	36.7
MBT [34]	224p	✗	64.8	58.0	43.4
MTCN [25]	336p	✗	70.7	62.1	49.6
M&M [57]	420p	✗	72.0	66.3	53.6
TIM (ours)	224p	✗	**77.5**	**67.4**	**57.9**

Perception Test Action				
Model	MLP (V)	MTCN [25](A+V)	**TIM (V)**	**TIM (A+V)**
Top-1 acc	43.7	51.2	56.1	**61.1**
Perception Test Sound				
Model	MLP (A)	MTCN [25](A+V)	**TIM (A)**	**TIM (A+V)**
Top-1 acc	50.6	52.9	54.8	**56.1**

Table 5. Comparisons to trained recognition baselines on the Perception Test validation split. We show both action and sound recognition and the benefit of including audio-visual in TIM for both challenges. **V** : visual and **A** : audio input features. MLP is the result by training an MLP classifier with the features directly.

TIM: A Time-Interval Audio-Visual Machine

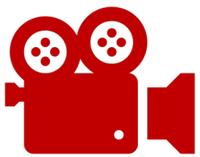
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



J Chalk*, J Huh*, E Kazakos, A Zisserman, D Damen (2024). TIM: A Time Interval Machine for Audio-Visual Action Recognition. CVPR

In today's tutorial

Motivation and Datasets in Egocentric Video Understanding



Video Understanding: Data and Tasks

Videos are Multimodal

Connected Videos of One's Life

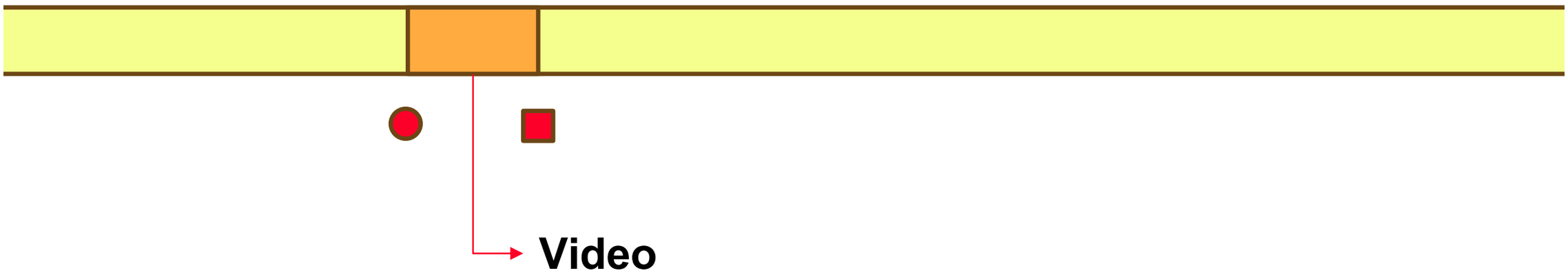
Video Understanding Out of the Frame

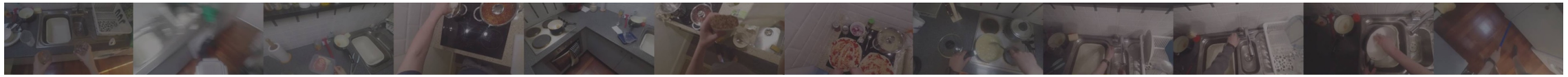
Teaser: The Wizard of Oz & Genie 3

Outlook into the Future of Egocentric Vision

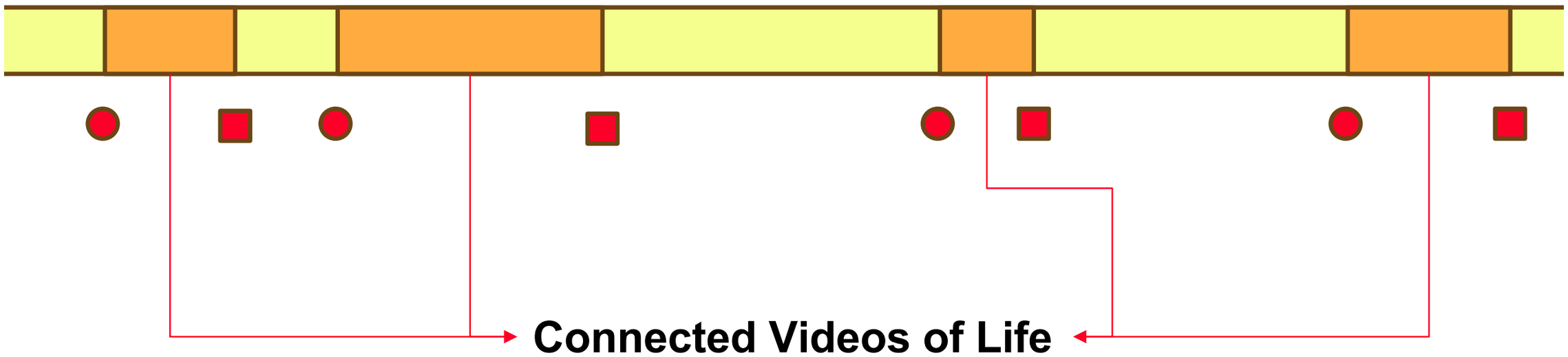
Conclusion

Current Video Understanding…



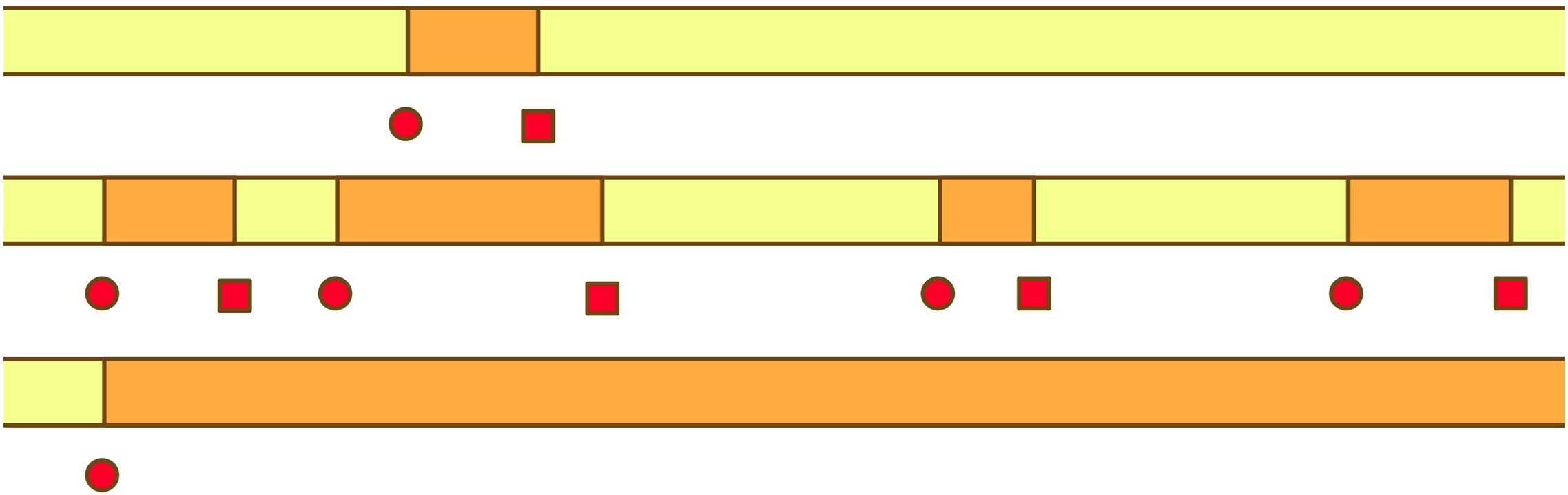


Upcoming Video Understanding…

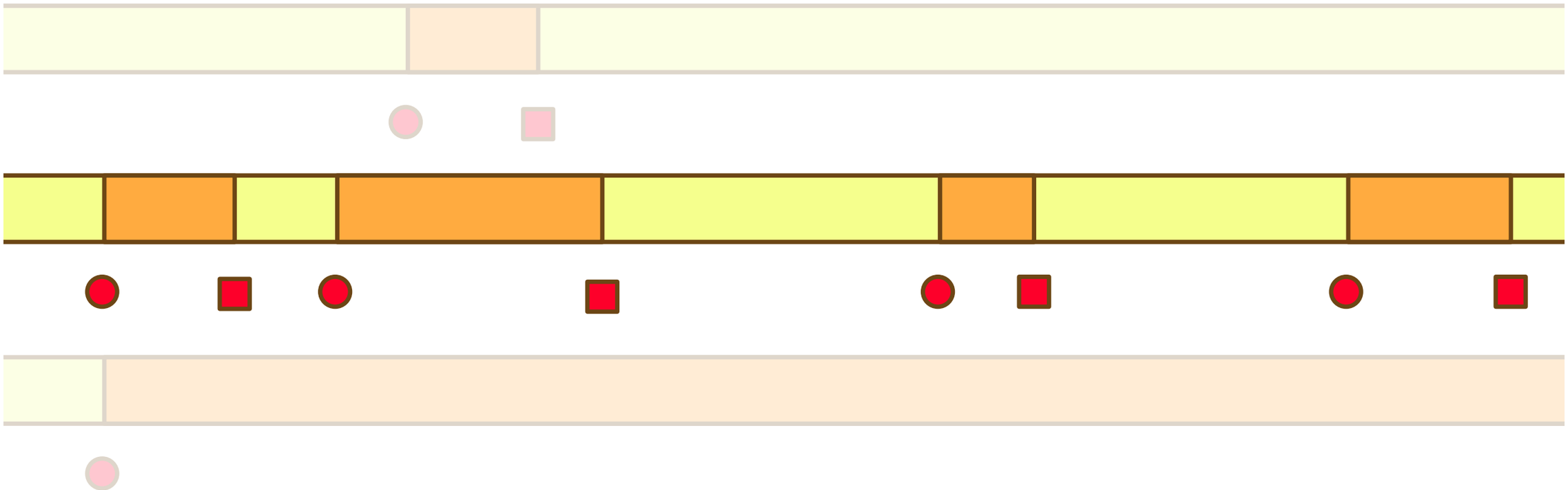


Eventually…

Egocentric Video Understanding

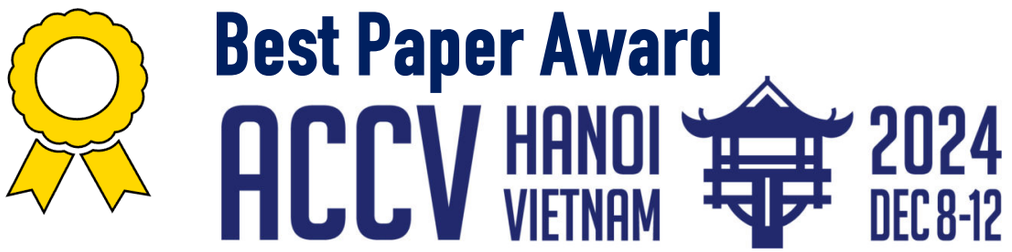


Egocentric Video Understanding



It’s Just Another Day: Unique Video Captioning by Discriminative Prompting

Toby Perrett, Tengda Han, Dima Damen, Andrew Zisserman



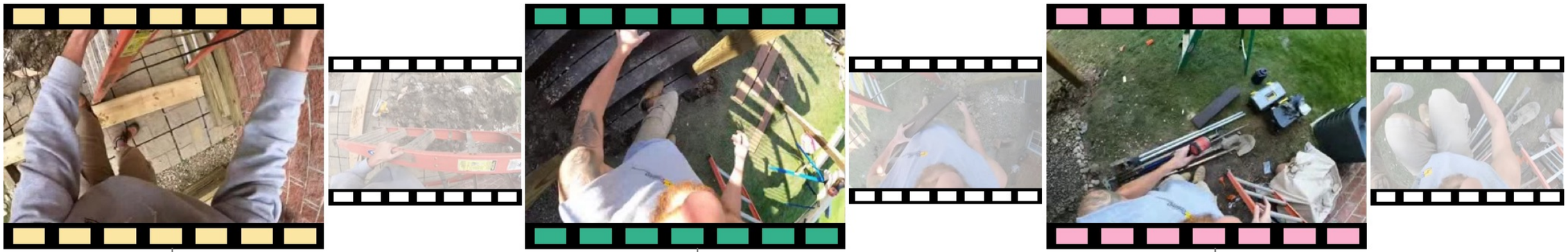
Unique Video Captioning

with: Toby Perrett
Tengda Han
Andrew Zisserman

Life is repetitive…

Unique Video Captioning

with: Toby Perrett
Tengda Han
Andrew Zisserman



- Current methods caption clips independently
- They generate the same caption for similar clips

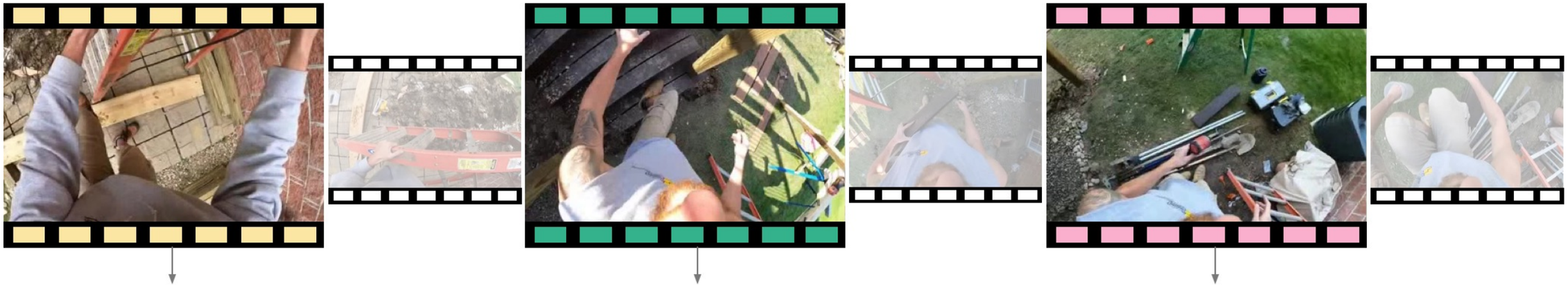
Unique Video Captioning

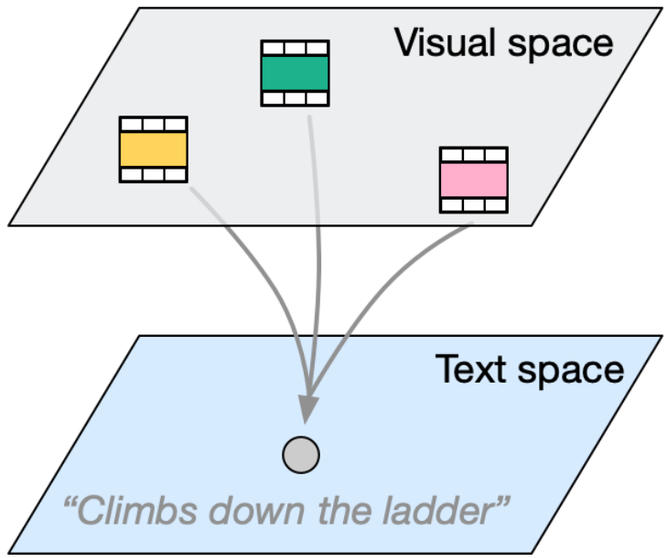
with: Toby Perrett
Tengda Han
Andrew Zisserman

Goal:
Generate a unique caption for every clip in a set

Unique Video Captioning

with: Toby Perrett
Tengda Han
Andrew Zisserman

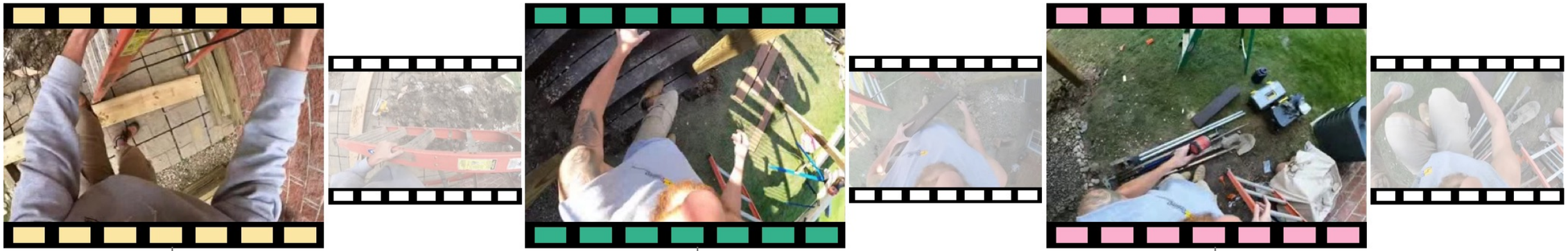




Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Unique Video Captioning

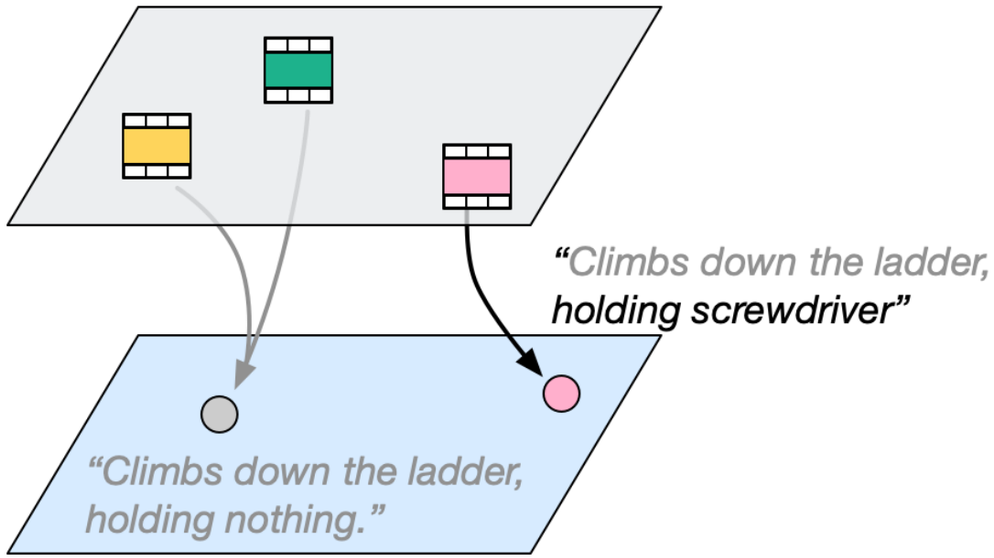
with: Toby Perrett
Tengda Han
Andrew Zisserman



"Climbs down the ladder"

"Climbs down the ladder"

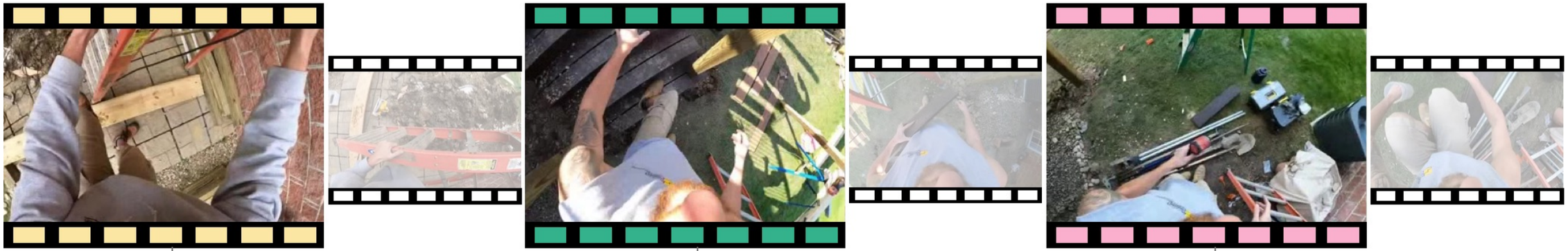
"Climbs down the ladder"



Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Unique Video Captioning

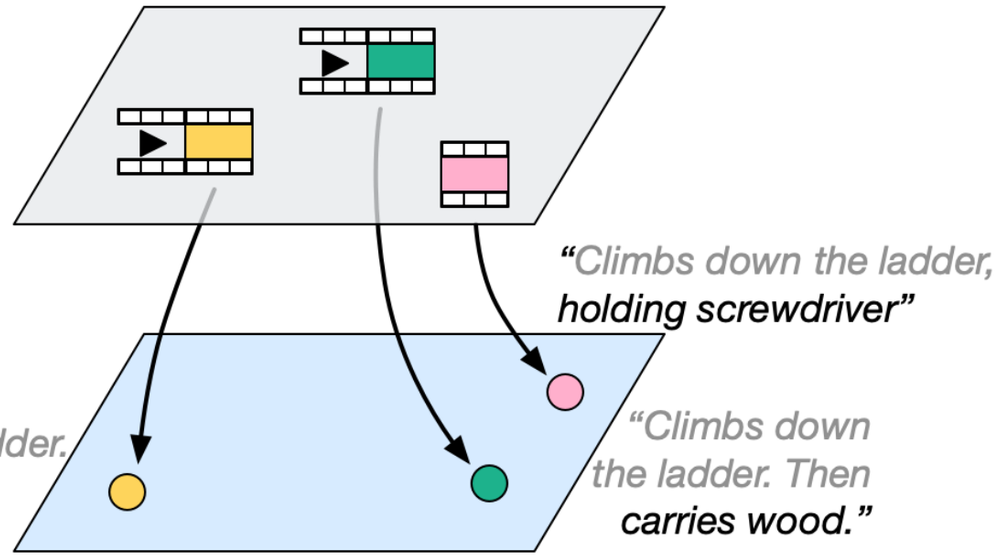
with: Toby Perrett
Tengda Han
Andrew Zisserman



"Climbs down the ladder"

"Climbs down the ladder"

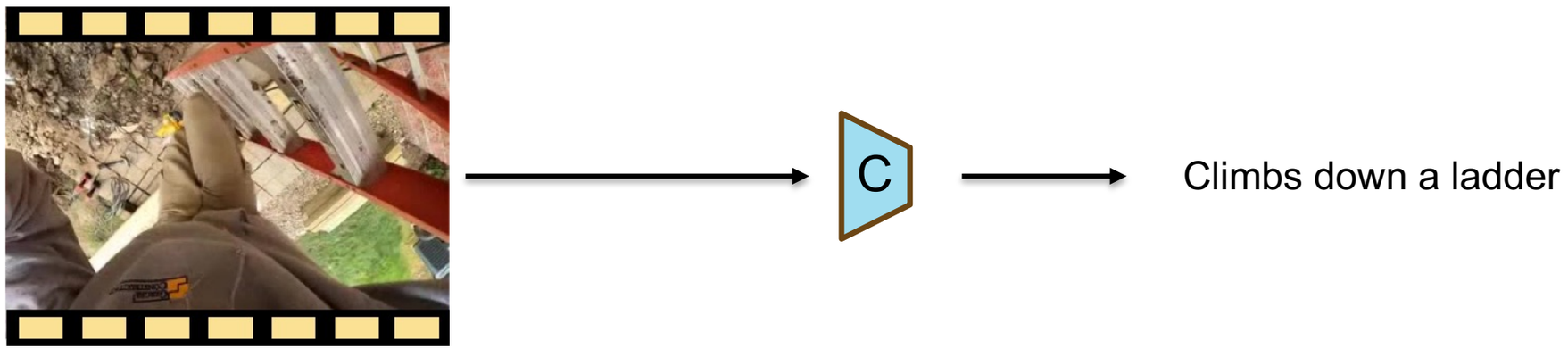
"Climbs down the ladder"



Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Captioning

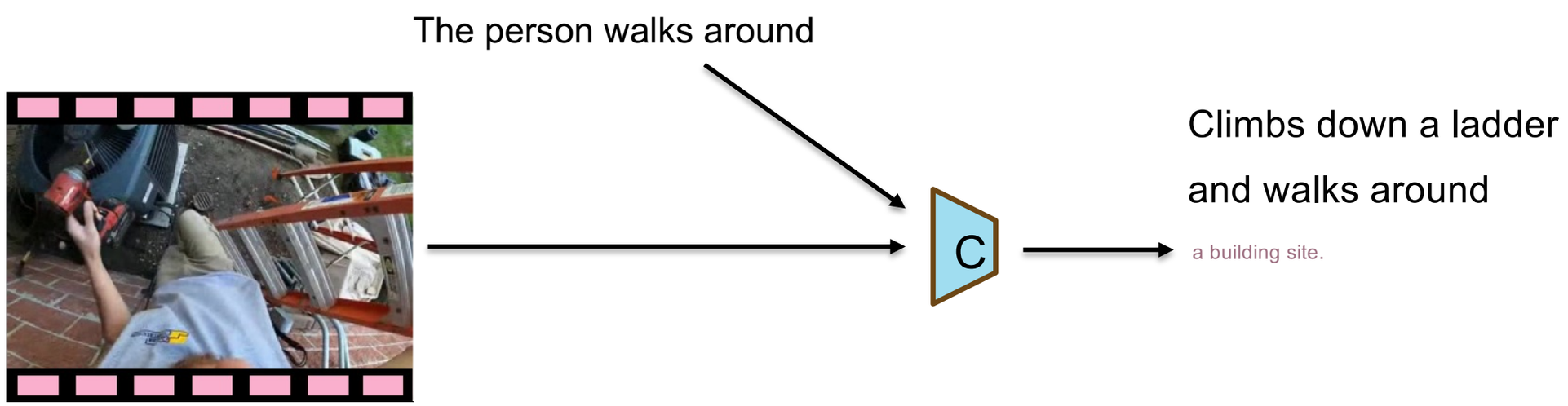
with: Toby Perrett
Tengda Han
Andrew Zisserman



Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Captioning

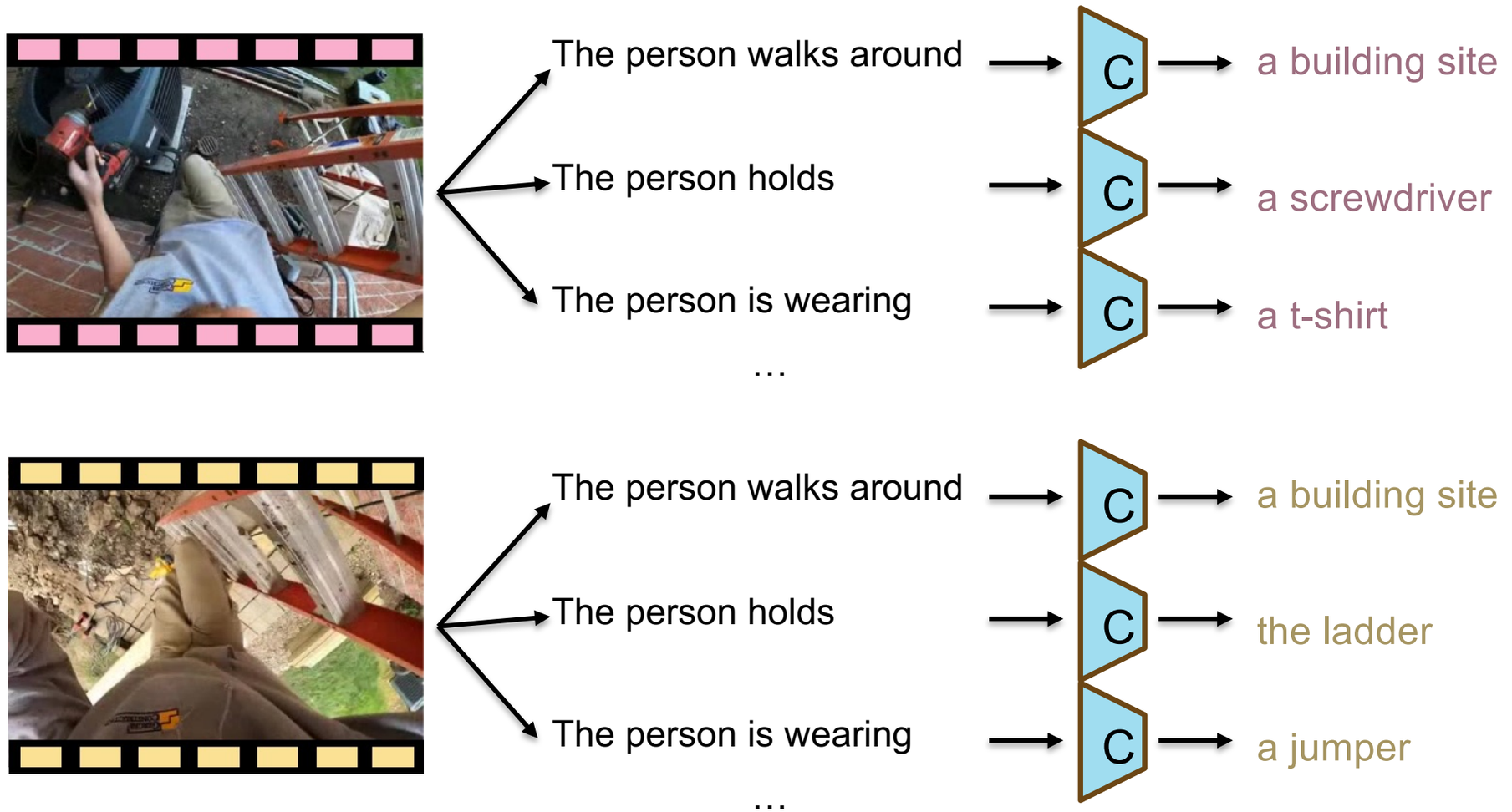
with: Toby Perrett
Tengda Han
Andrew Zisserman



Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Captioning by Discriminative Prompting

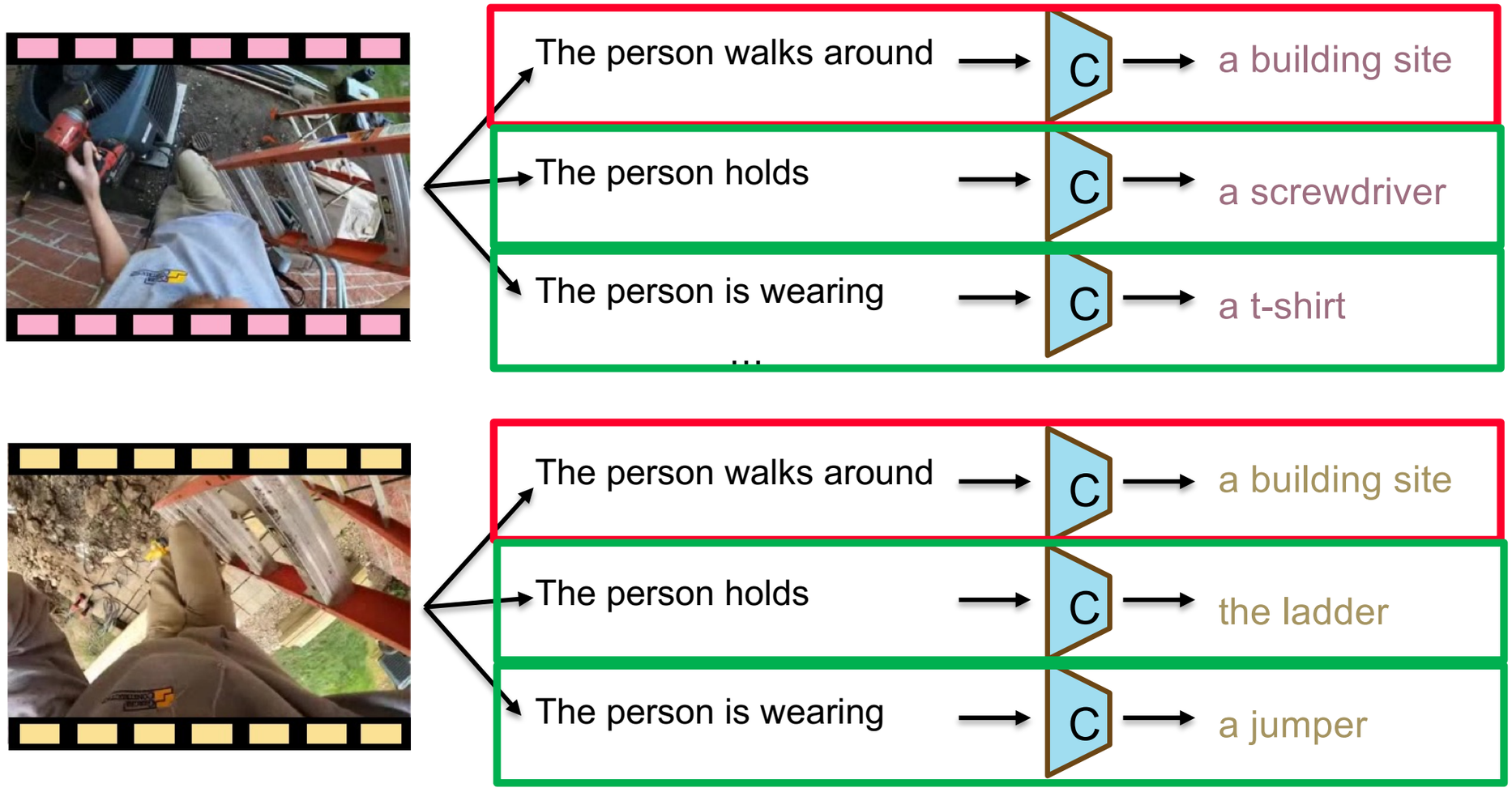
with: Toby Perrett
Tengda Han
Andrew Zisserman



Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Captioning by Discriminative Prompting

with: Toby Perrett
Tengda Han
Andrew Zisserman



Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Captioning by Discriminative Prompting

with: Toby Perrett
Tengda Han
Andrew Zisserman

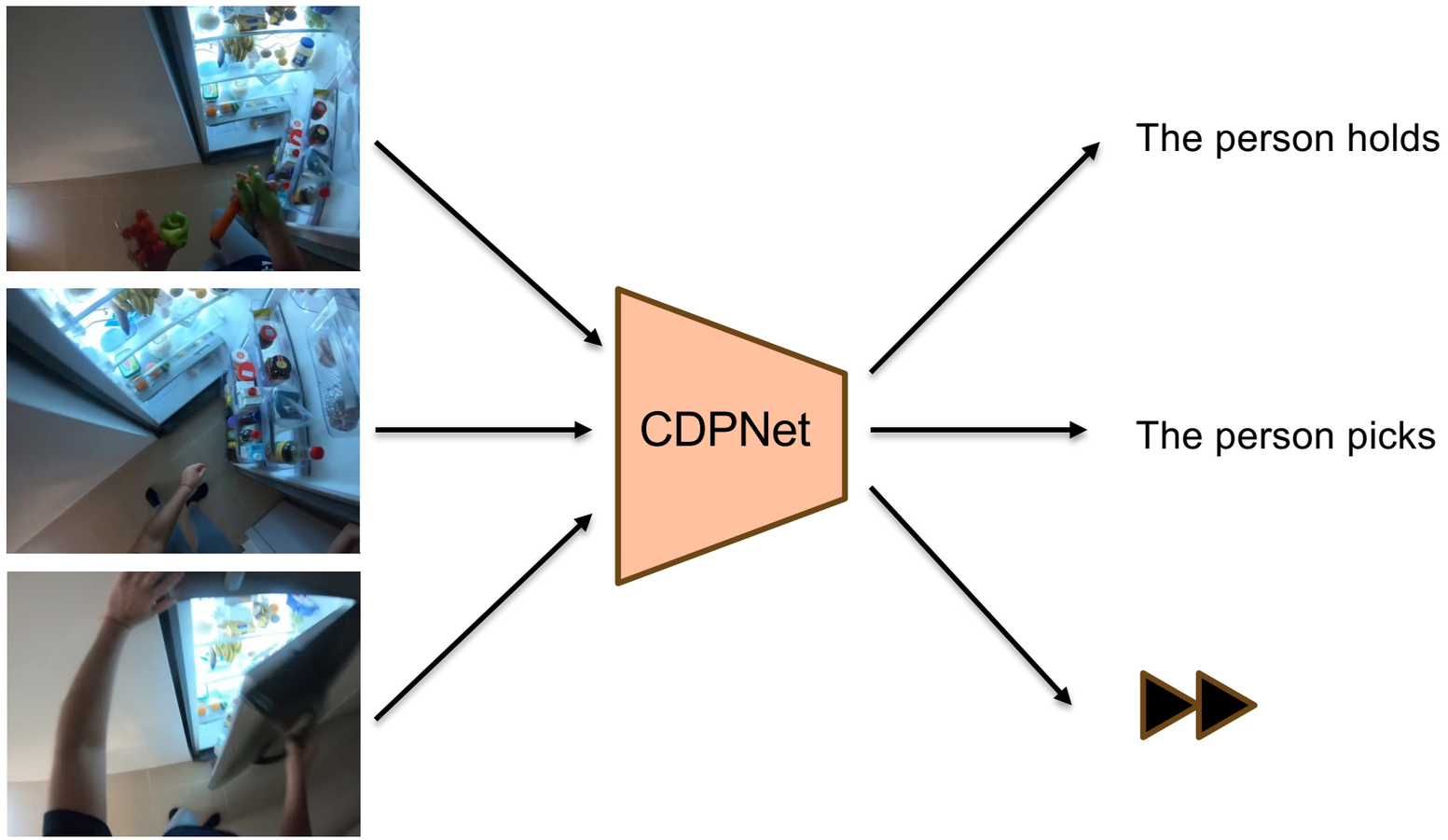
We propose to…
consider clips jointly
use a bank of discriminative prompts

But…
Expensive £££
What if there isn't a suitable prompt?

Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Captioning by Discriminative Prompting

with: Toby Perrett
Tengda Han
Andrew Zisserman

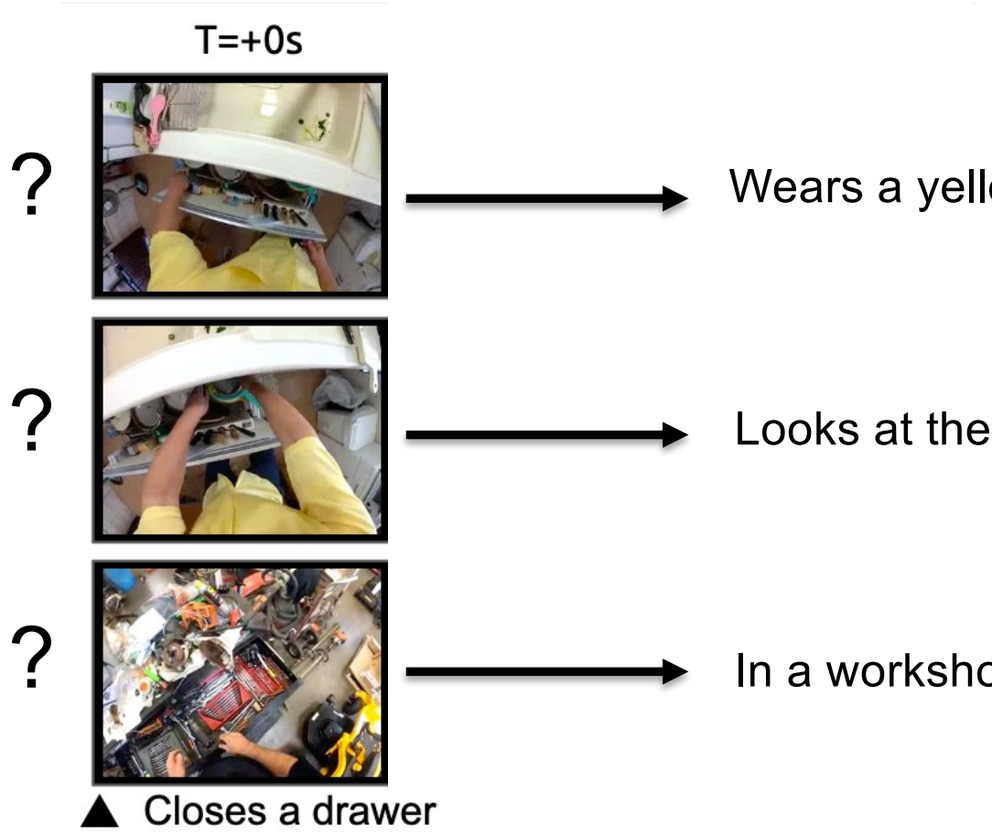


Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Benchmarks

with: Toby Perrett
Tengda Han
Andrew Zisserman

Task: Caption every clip, then evaluate retrieval.



Video-text R @ 1

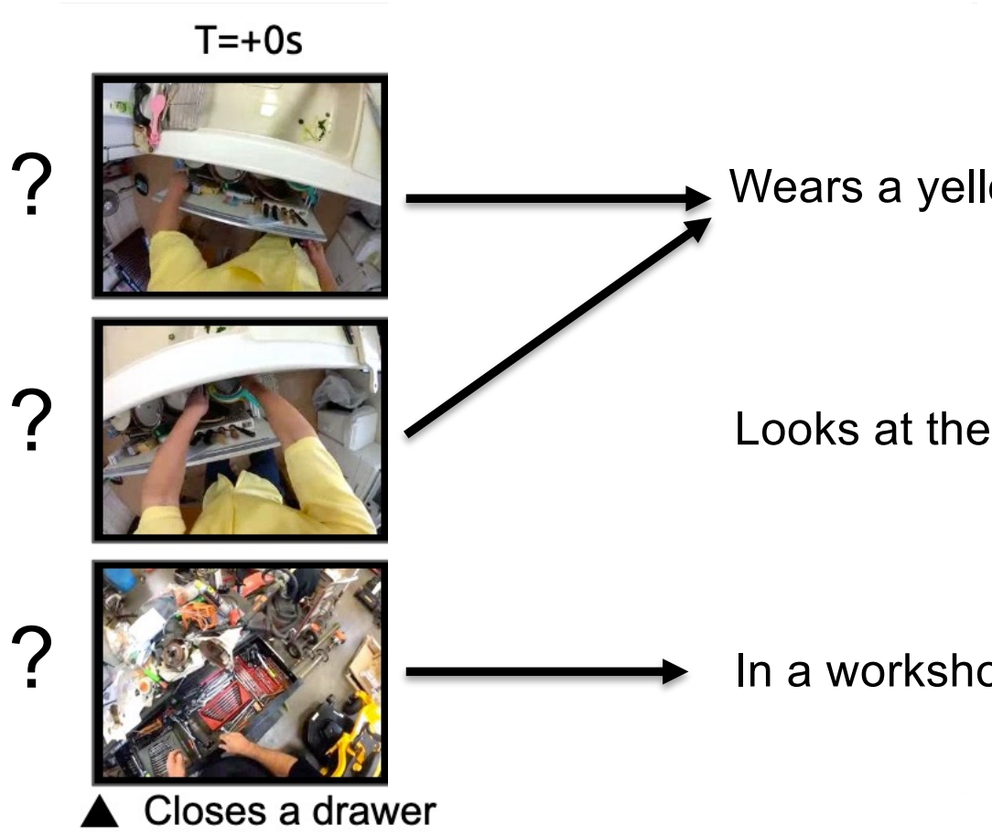
100%

Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Benchmarks

with: Toby Perrett
Tengda Han
Andrew Zisserman

Task: Caption every clip, then evaluate retrieval.



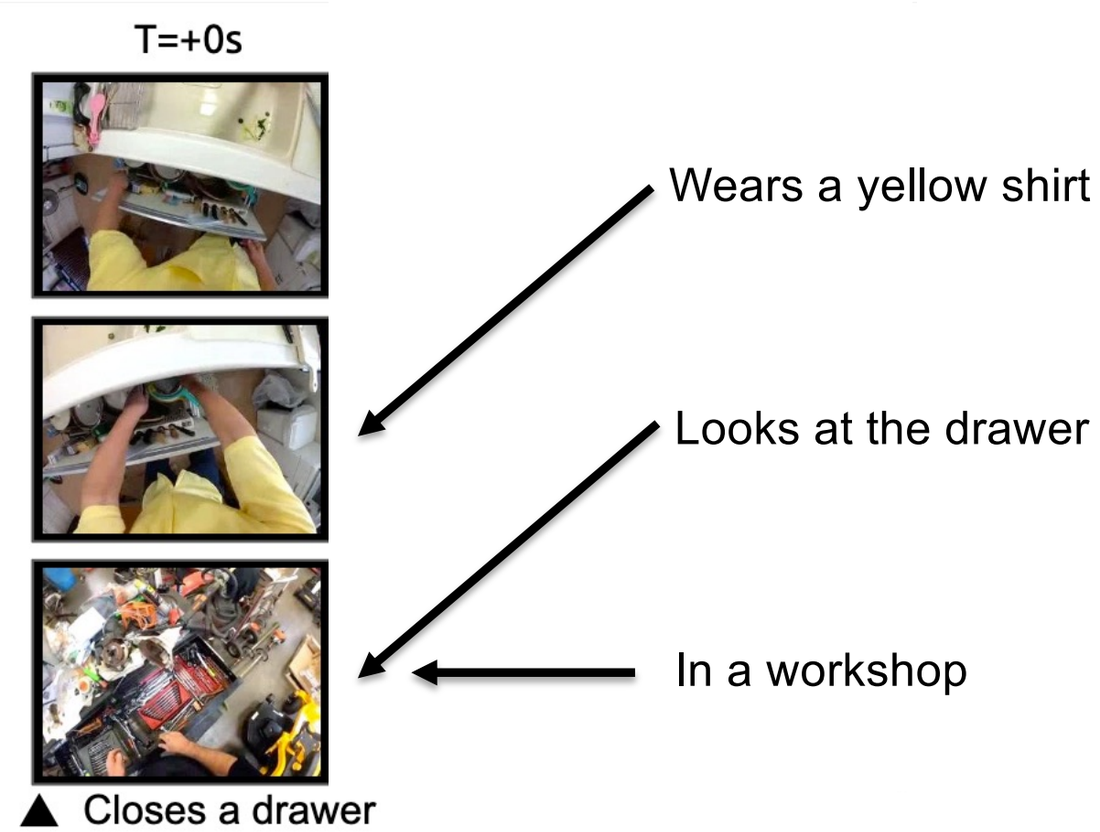
Video-text R @ 1

66%

Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Benchmarks

with: Toby Perrett
Tengda Han
Andrew Zisserman



Avg Recall @ 1
= (Video-text + Text-video R @ 1) / 2
= (33 + 66) / 2
= 50%

Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Unique Video Captioning

with: Toby Perrett
Tengda Han
Andrew Zisserman

Average recall @ 1

Egocentric	**+0s**
LaViLa	37
LaViLa + CDP	**45**

Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Unique Video Captioning

with: Toby Perrett
Tengda Han
Andrew Zisserman

Climbs the stairs

Unique Video Captioning

with: Toby Perrett
Tengda Han
Andrew Zisserman

Climbs the stairs

Climbs the stairs and

holds the phone

Climbs the stairs and

picks up the drill

Climbs the stairs and

holds a tape measure

Unique Video Captioning

with: Toby Perrett
Tengda Han
Andrew Zisserman

Looks around the shelves

Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Unique Video Captioning

with: Toby Perrett
Tengda Han
Andrew Zisserman

Looks around the shelves

Looks around the shelves and

the other man picks up a packet of biscuits from the shelf with his left hand

Looks around the shelves and

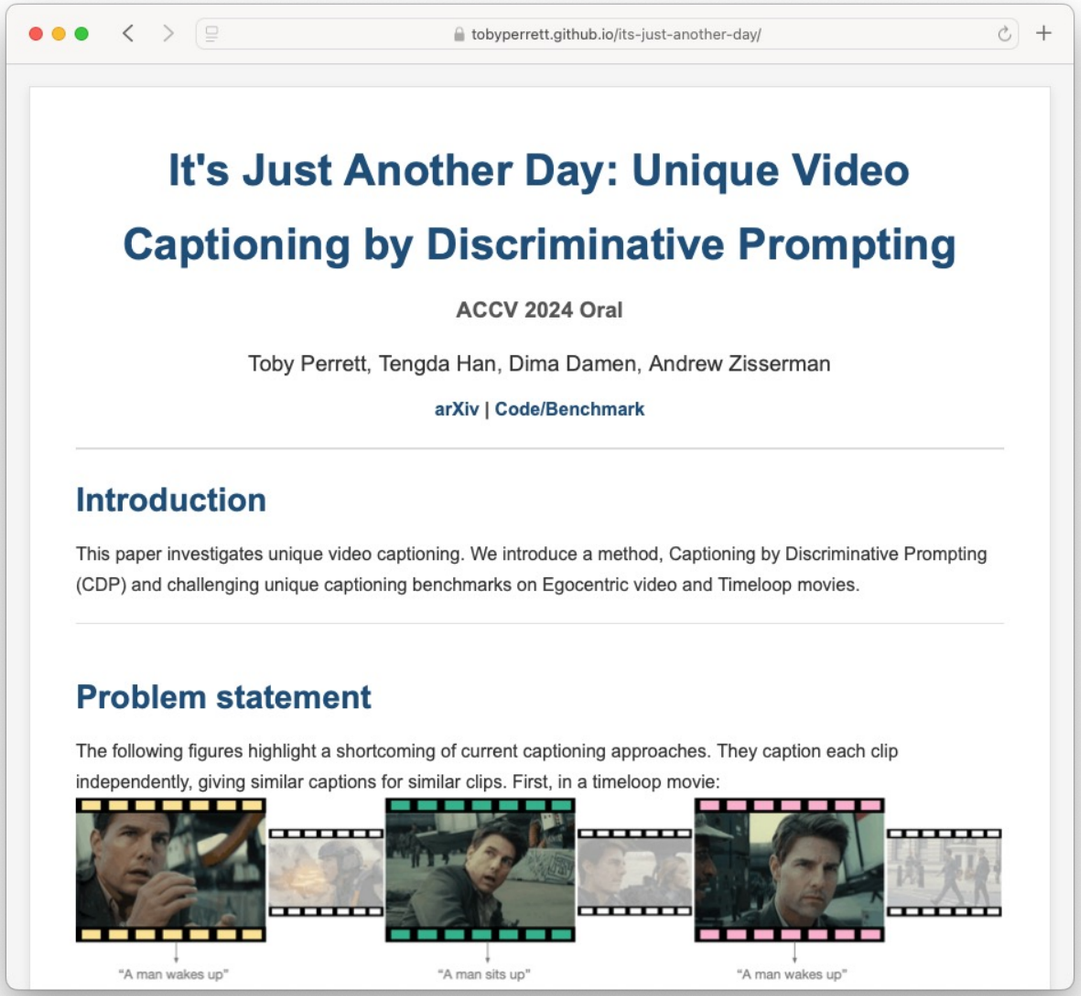
looks at the list

Looks around the shelves and then

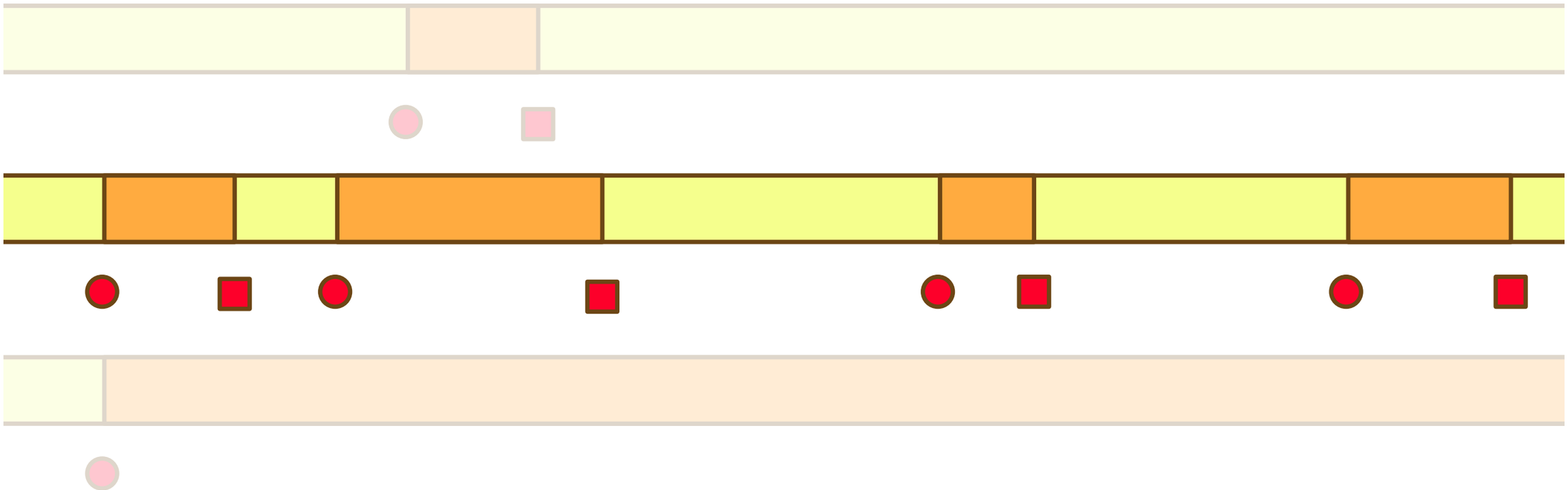
picks up a packet of cough rubs

Unique Video Captioning

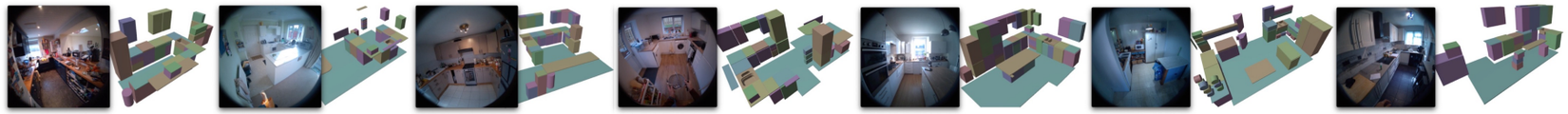
with: Toby Perrett
Tengda Han
Andrew Zisserman



Egocentric Video Understanding



CVPR Nashville JUNE 11-15, 2025



HD-EPIC

HD-EPIC: A Highly-Detailed Egocentric Video Dataset

University of BRISTOL

Universiteit Leiden

SINGAPORE MANAGEMENT UNIVERSITY

UNIVERSITY OF BATH

 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

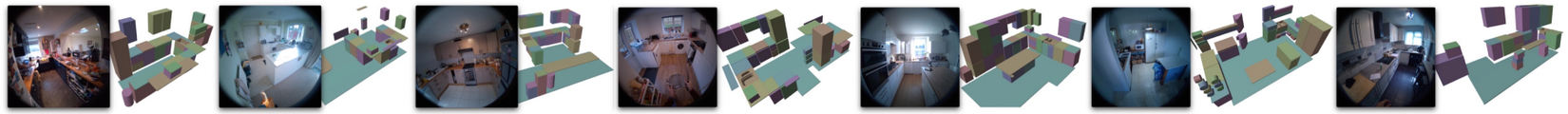
 Davide Moltisanti

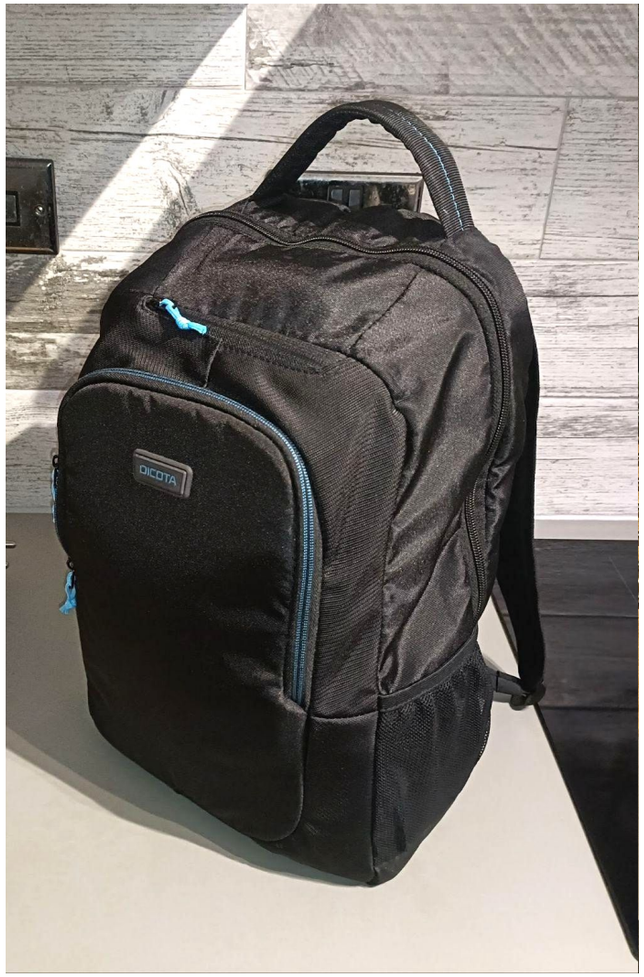
 Michael Wray

 Hazel Doughty

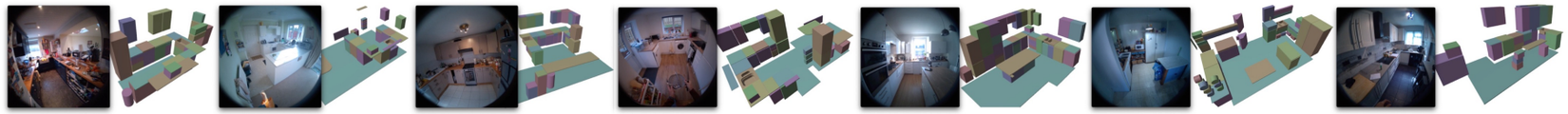
 Dima Damen

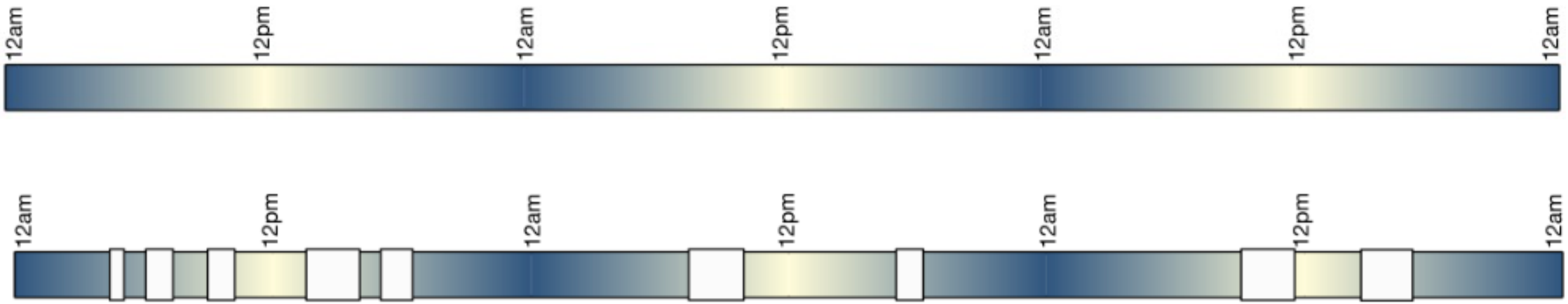
HD-EPIC



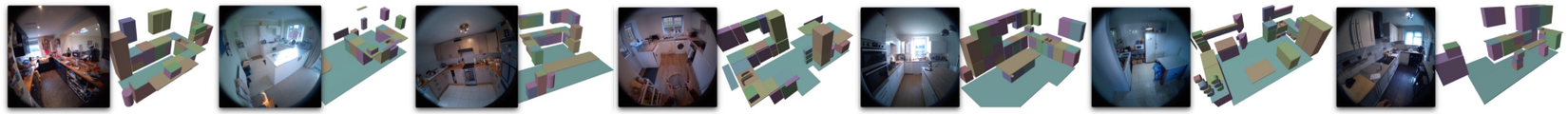


HD-EPIC

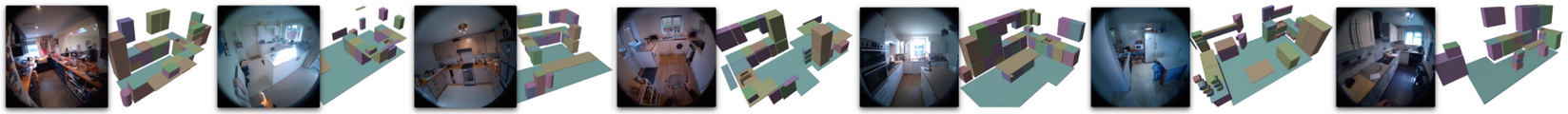


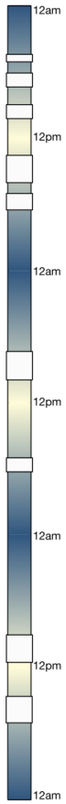

12am
12pm
12am
12pm
12am
12pm
12am
12am
12pm
12am
12pm
12am
12pm
12am

HD-EPIC

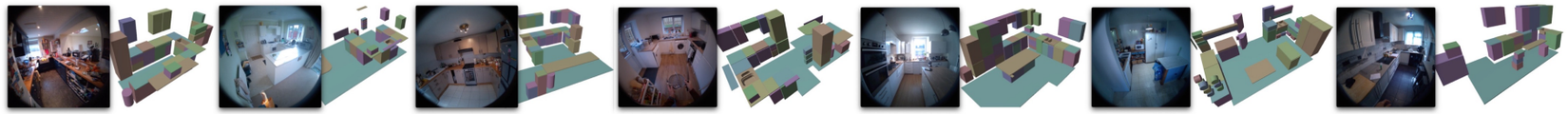


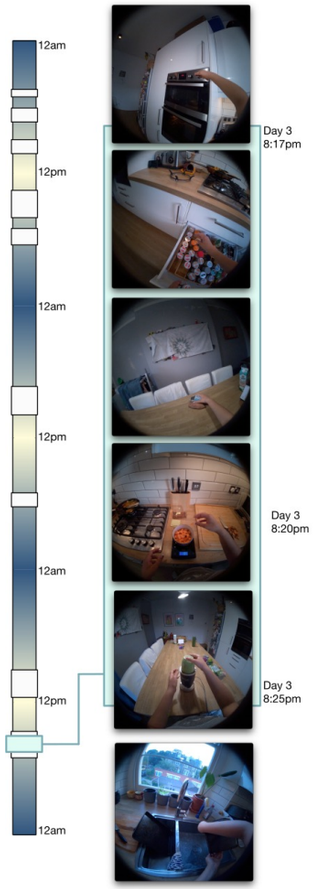
HD-EPIC



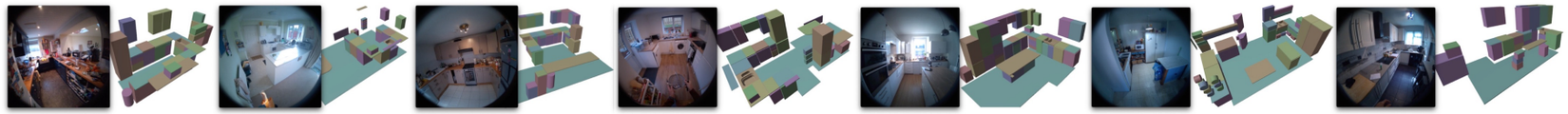


HD-EPIC




12am
12pm
12am
12pm
12am
12pm
12am
Day 3
8:17pm
Day 3
8:20pm
Day 3
8:25pm

HD-EPIC



Recipe: Southwestern Salad

1: Preheat the oven to 400F

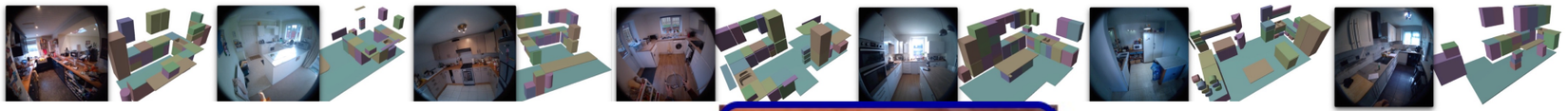
Day 3
1:17pm

2: Wash and peel the sweet potatoes and chop into bite-sized pieces. Put the sweet potatoes in a bowl and add the olive oil, cumin, and chili powder. Pour onto tray and roast for 10 mins.

3: Pulse all the dressing ingredients in a food processor until mostly smooth.

Recipe and nutrition

HD-EPIC



Cacio e Pepe (modified)

Ingredients:

~~200 g~~ → penne

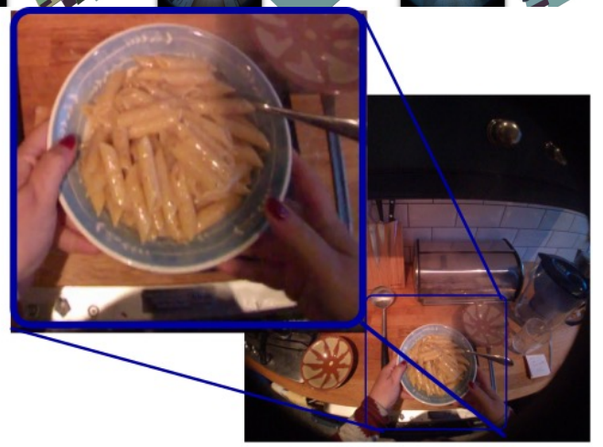
~~400g~~ of pasta of your choice
(we recommend bucatini)

~~2~~ 1 tablespoon of black peppercorn

30 g parmigiano

200g of freshly grated pecorino cheese

+25g of slightly salted butter



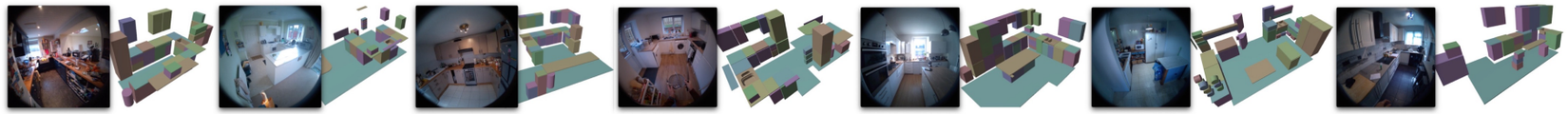
Steps:

1. Toast the peppercorns until fragrant in a dry frying pan over medium heat, about 2 minutes. Keep them moving to prevent them from burning. ~~Once toasted, roughly crush.~~ → step 2

2. Cook your choice of pasta in a large pot of generously salted boiling water ~~for around 4-6 minutes~~, or until al dente. → step 1

3. While the pasta cooks, add freshly grated cheese and crushed black peppercorns to a large serving bowl (on very low heat). Gradually add a cup of the boiling cooking water constantly mixing to obtain a silky, smooth sauce that's able to completely coat the pasta. → step 3

HD-EPIC

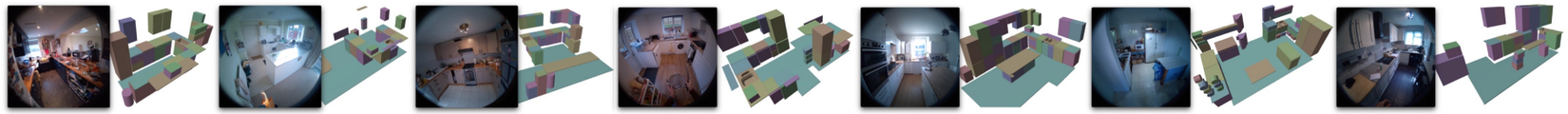


- The **prep** of a corresponding **step** is defined as

all essential actions the participant takes to get ready to execute a given step.

- For example, the **step** ‘chop tomato’:
  - **Prep:** retrieve tomato from storage, wash tomato, retrieve a knife and chopping board.
- the **step** ‘add chopped onions and stir’:
  - **Prep:** retrieve onion from storage, retrieve a knife and chopping board, **and chop the onions.**

HD-EPIC



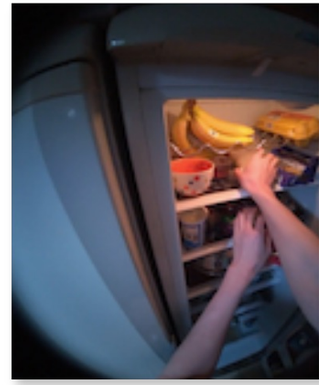
- Prep

- Step

pick up kettle from its base on the counter with my right hand

pick up packet of bacon

pour water from kettle into the pan with my right hand

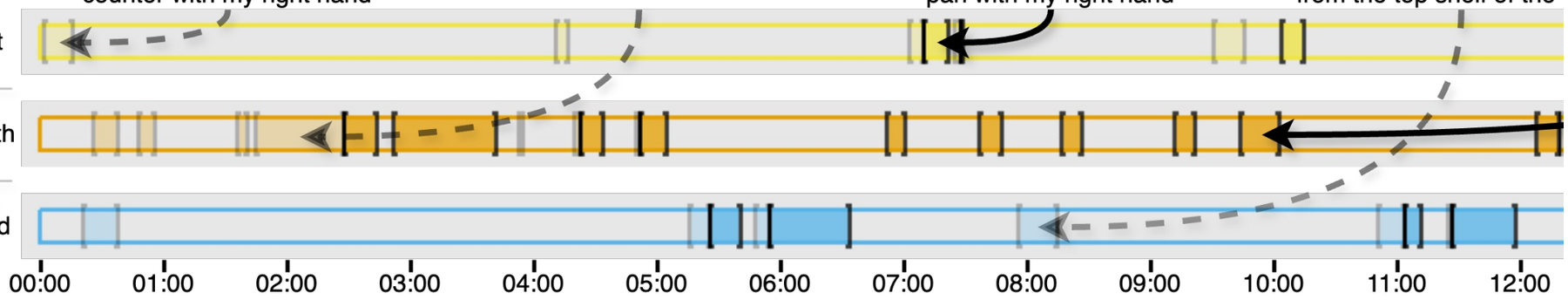


pick up block of cheese tha from the top shelf of the

Cook the pasta in a pan of boiling salted water according to the packet instructions.

Slice the bacon and place in a non-stick frying pan on a medium heat with half a tablespoon of olive oil and ...

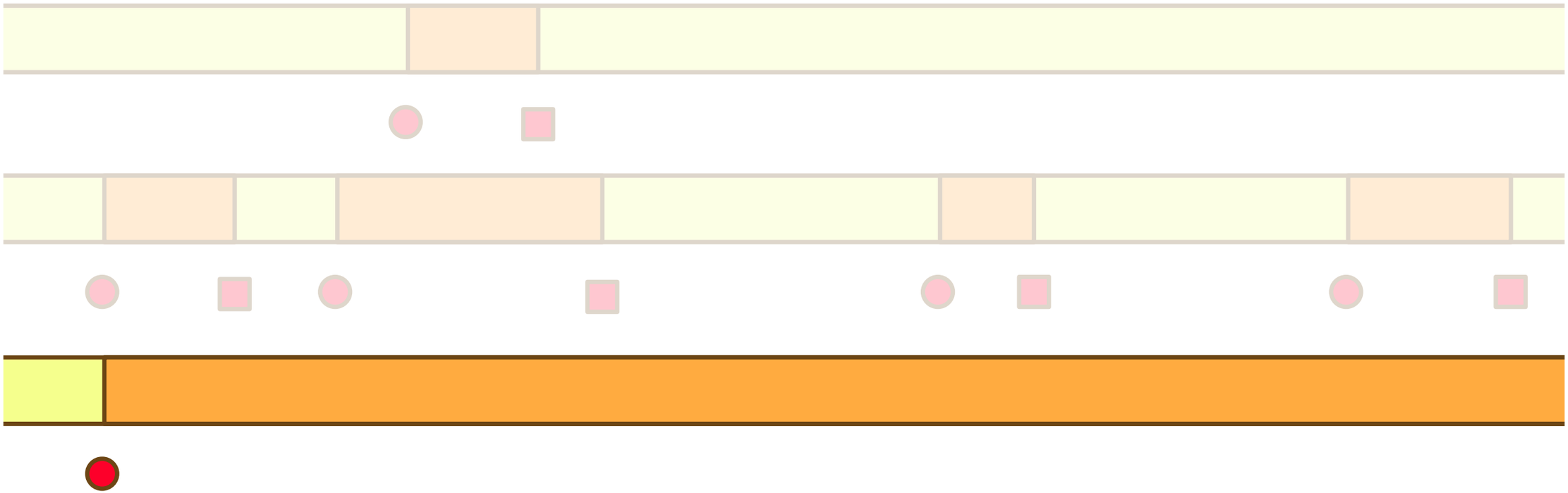
Meanwhile, beat the eggs in a bowl, then finely grate in the Parmesan and mix well.

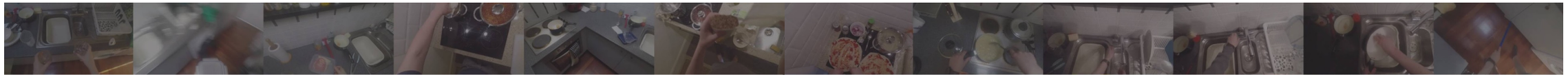


University of BRISTOL

Dima Damen
PAISS 2025

Egocentric Video Understanding





Eventually…

- No current model has the context required for this ...
- Impossible to store and process this influx of data …

But….

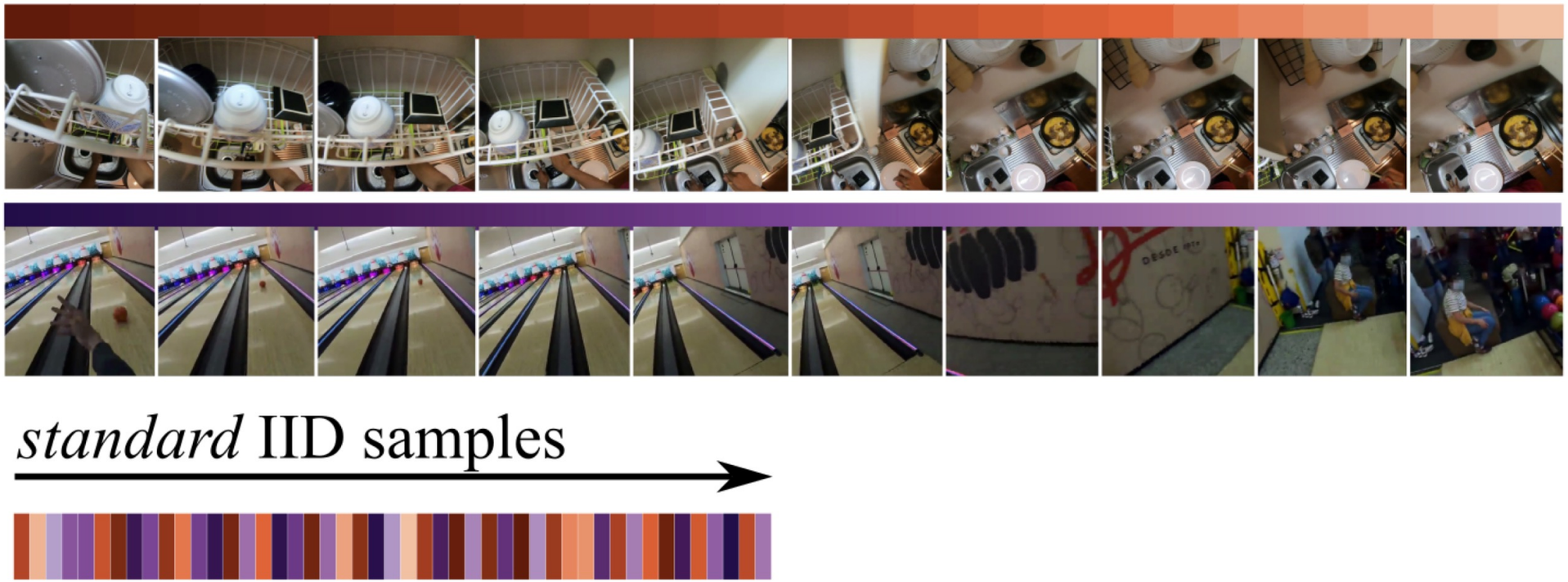
- Immense potential …

Learning from Streaming Video with Orthogonal Gradients

Tengda Han[◇], Dilara Gokay[◇], Joseph Heyward[◇], Chuhan Zhang[◇]
Daniel Zoran[◇], Viorica Pătrăucean[◇], João Carreira[◇], Dima Damen[◇†], Andrew Zisserman[◇‡]
[◇]Google DeepMind, [†]University of Bristol, [‡]University of Oxford

Han et al (2025). Learning from Streaming Video with Orthogonal Gradients. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

Learning from Streaming Videos with Orthogonal Gradients

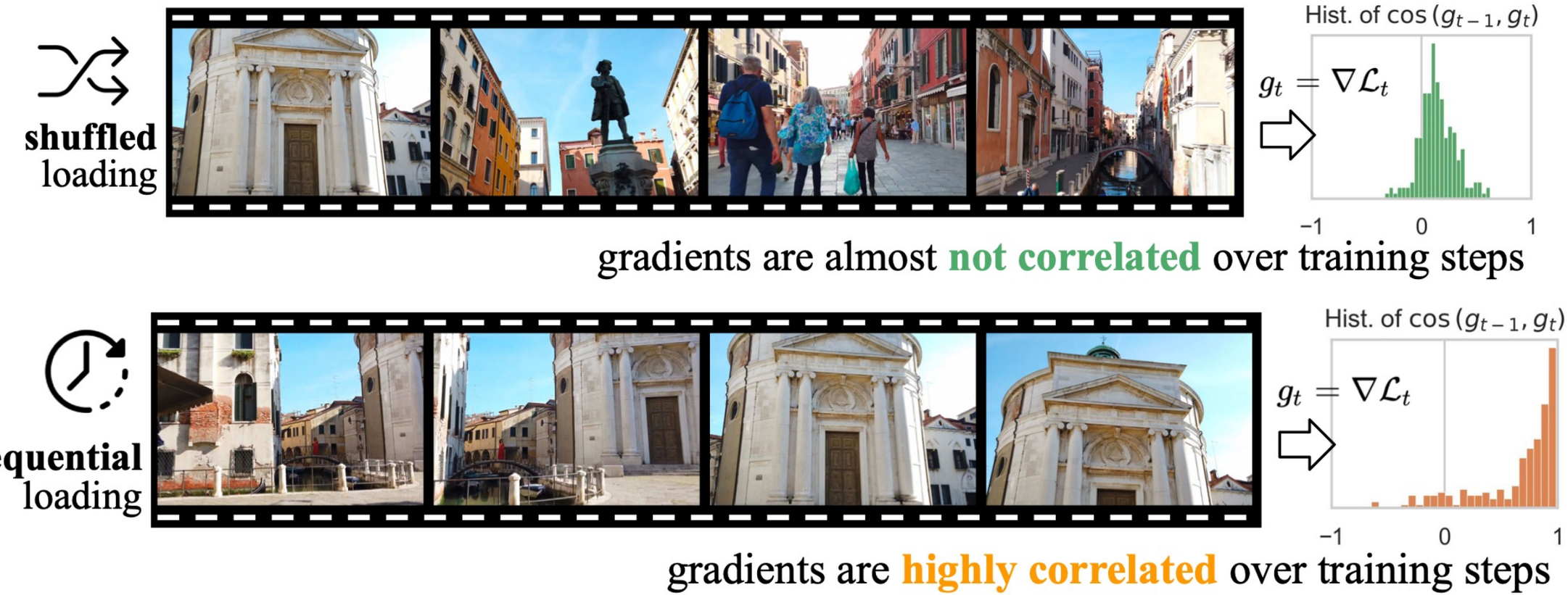


Carriera et al (2024). Learning from One Continuous Video Stream. CVPR

Learning from Streaming Videos with Orthogonal Gradients

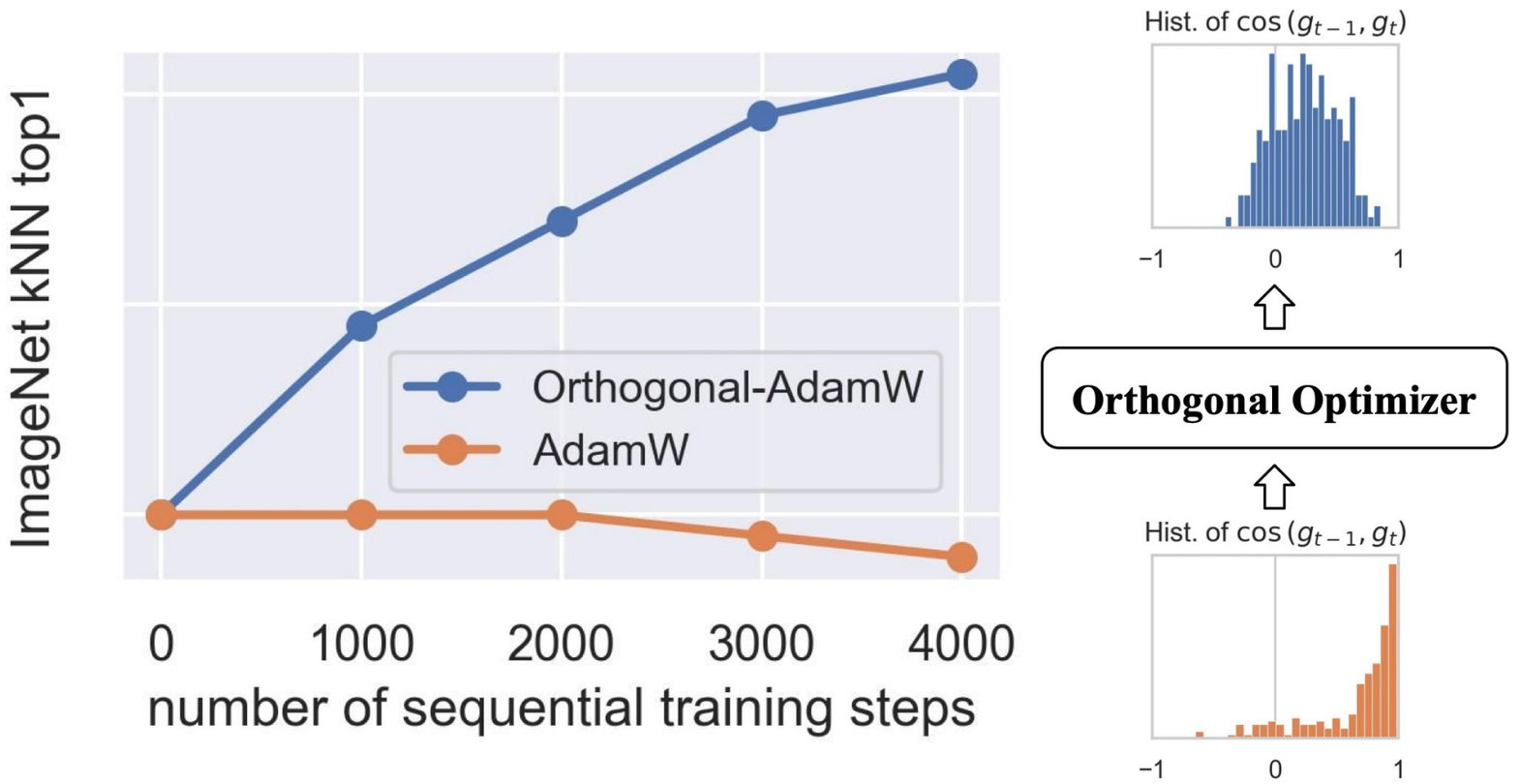
Han et al (2025). Learning from Streaming Video with Orthogonal Gradients. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

Learning from Streaming Videos with Orthogonal Gradients



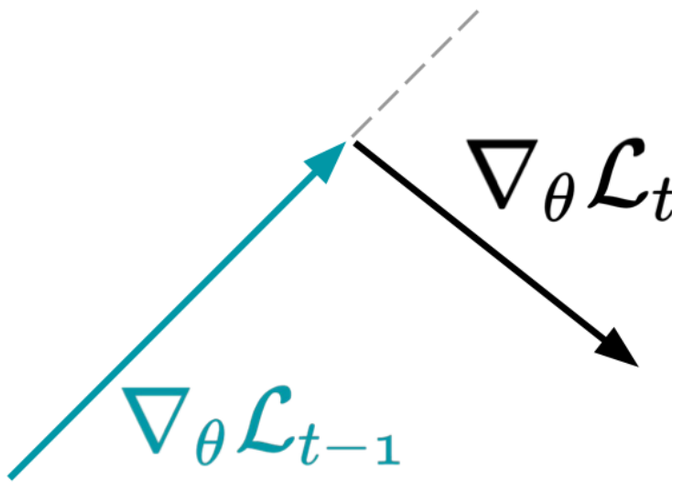
Han et al (2025). Learning from Streaming Video with Orthogonal Gradients. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

Learning from Streaming Videos with Orthogonal Gradients



Han et al (2025). Learning from Streaming Video with Orthogonal Gradients. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

Learning from Streaming Videos with Orthogonal Gradients



(a)

Han et al (2025). Learning from Streaming Video with Orthogonal Gradients. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

Learning from Streaming Videos with Orthogonal Gradients

Algorithm 2 AdamW

Require: Learning rate $\eta > 0$, weight decay coefficient $\lambda > 0$, decay rates $\beta_1, \beta_2 \in [0, 1)$, small constant $\epsilon > 0$, initial parameters θ_0, number of iterations T

Initialize first moment vector $m_0 = 0$, , and second moment vector $v_0 = 0$
for $t = 1$ to T **do**
  Sample a mini-batch of data $\mathcal{B}_t$ from the training set
  Compute the gradient: $g_t = \nabla_\theta \mathcal{L}(\theta_{t-1}; \mathcal{B}_t)$

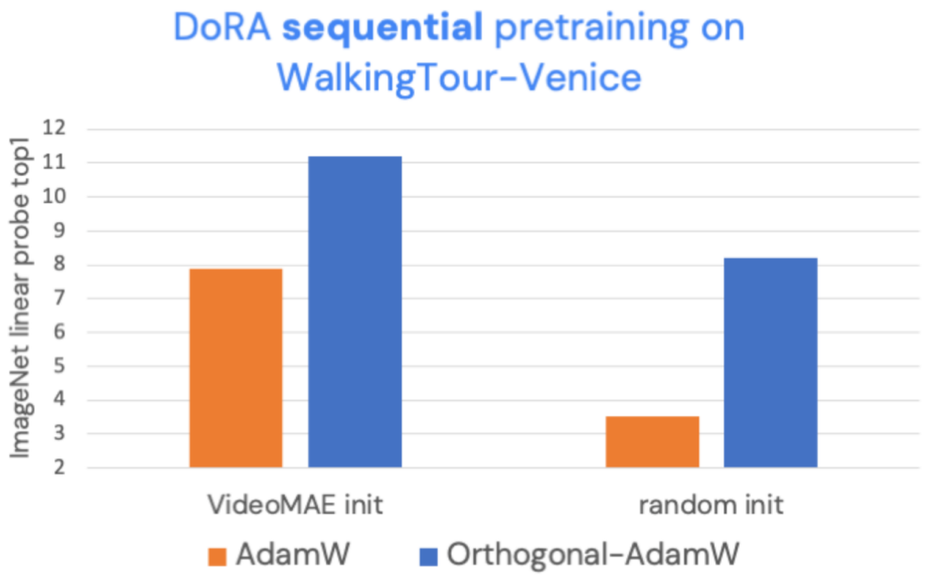
  Update biased first moment estimate: $m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$
  Update biased second moment estimate: $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$
  Compute bias-corrected first moment: $\hat{m}_t = \frac{m_t}{1-\beta_1^t}$
  Compute bias-corrected second moment: $\hat{v}_t = \frac{v_t}{1-\beta_2^t}$
  Apply weight decay: $\theta_{t-1} = \theta_{t-1} - \eta \lambda \theta_{t-1}$
  Update parameters: $\theta_t = \theta_{t-1} - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$
end for

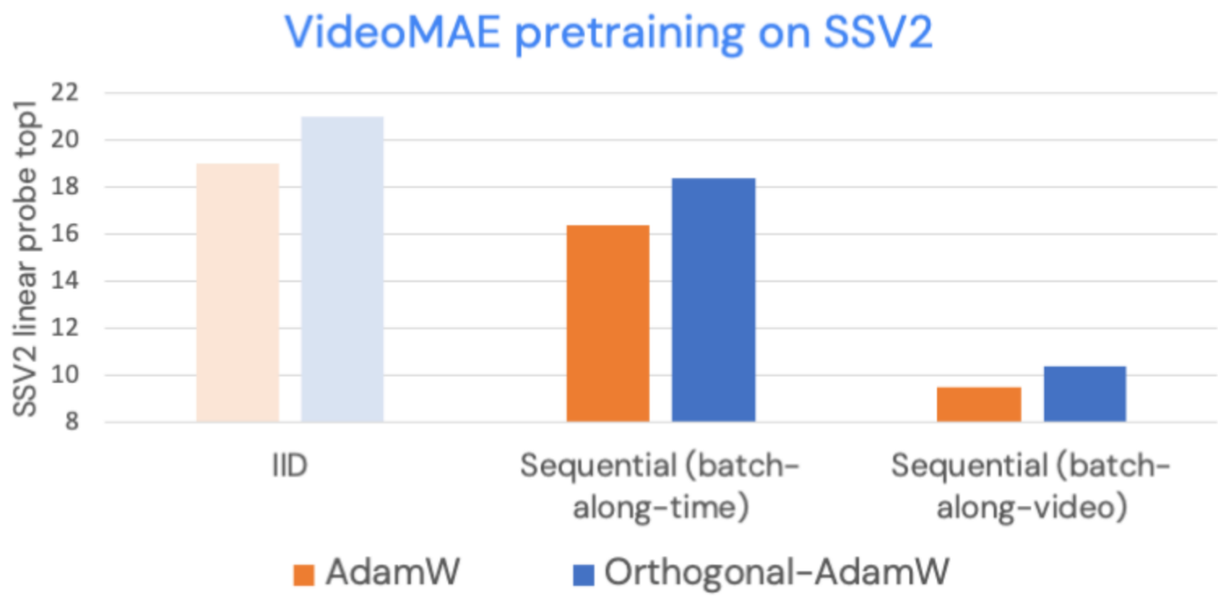
Han et al (2025). Learning from Streaming Video with Orthogonal Gradients. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

Learning from Streaming Videos with Orthogonal Gradients

Han et al (2025). Learning from Streaming Video with Orthogonal Gradients. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

Learning from Streaming Videos with Orthogonal Gradients



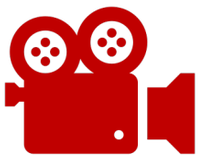


Han et al (2025). Learning from Streaming Video with Orthogonal Gradients. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

In today's tutorial

Motivation and Datasets in Egocentric Video Understanding

Video Understanding Out of the Frame



Video Understanding: Data and Tasks

Teaser: The Wizard of Oz & Genie 3

Videos are Multimodal

Outlook into the Future of Egocentric Vision

Connected Videos of One's Life

Conclusion

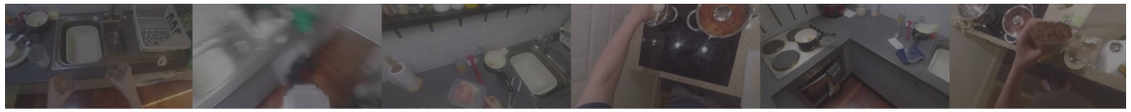
First-person Hyperlapse Videos

Johannes Kopf Michel F. Cohen Richard Szeliski
Microsoft Research

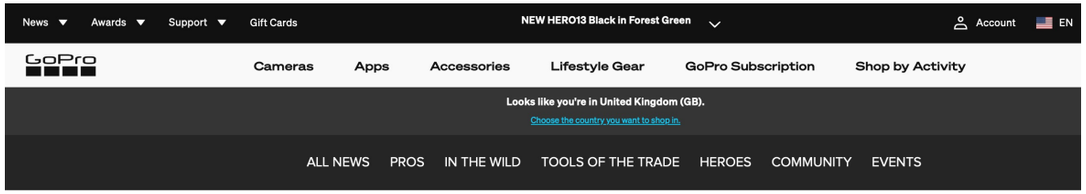
research.microsoft.com/hyperlapse

SIGGRAPH 2014

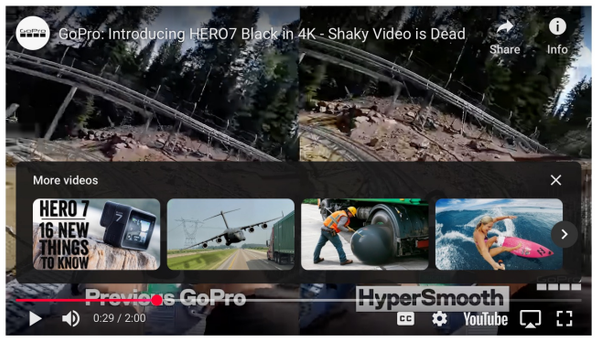

Rendered from single input frame



4 years later (2018)



SHAKY VIDEO IS DEAD: HERO7 BLACK IS HERE



Today, GoPro announced its new product lineup including the $399 flagship, HERO7 Black, which sets a new bar for video stabilization with its standout feature, HyperSmooth.

HyperSmooth is the best in-camera video stabilization ever featured in a camera. It makes it easy to capture professional-looking, gimbal-like stabilized video without the expense or hassle of a motorized gimbal. And HyperSmooth works underwater and in high-shock and wind situations where gimbals fail. HERO7 Black with HyperSmooth video stabilization – you've got to see it to believe it.

https://gopro.com/en/us/news/launch-2018-hero7-black

https://youtu.be/G9KDqfpCgws?si=t2yipPgTqSrAJLNo

Video Understanding Out of the Frame

While neighbouring frames have been used in on-board video stabilisation, approaches focused on video understanding within the frame

University of BRISTOL

Dima Damen
PAISS 2025

EPIC-KITCHENS

EPIC Fields

with: V Tschernezki*, A Darkhalil*, Z Zhu*,
D Fouhey, I Laina, D Larlus, A Vedaldi

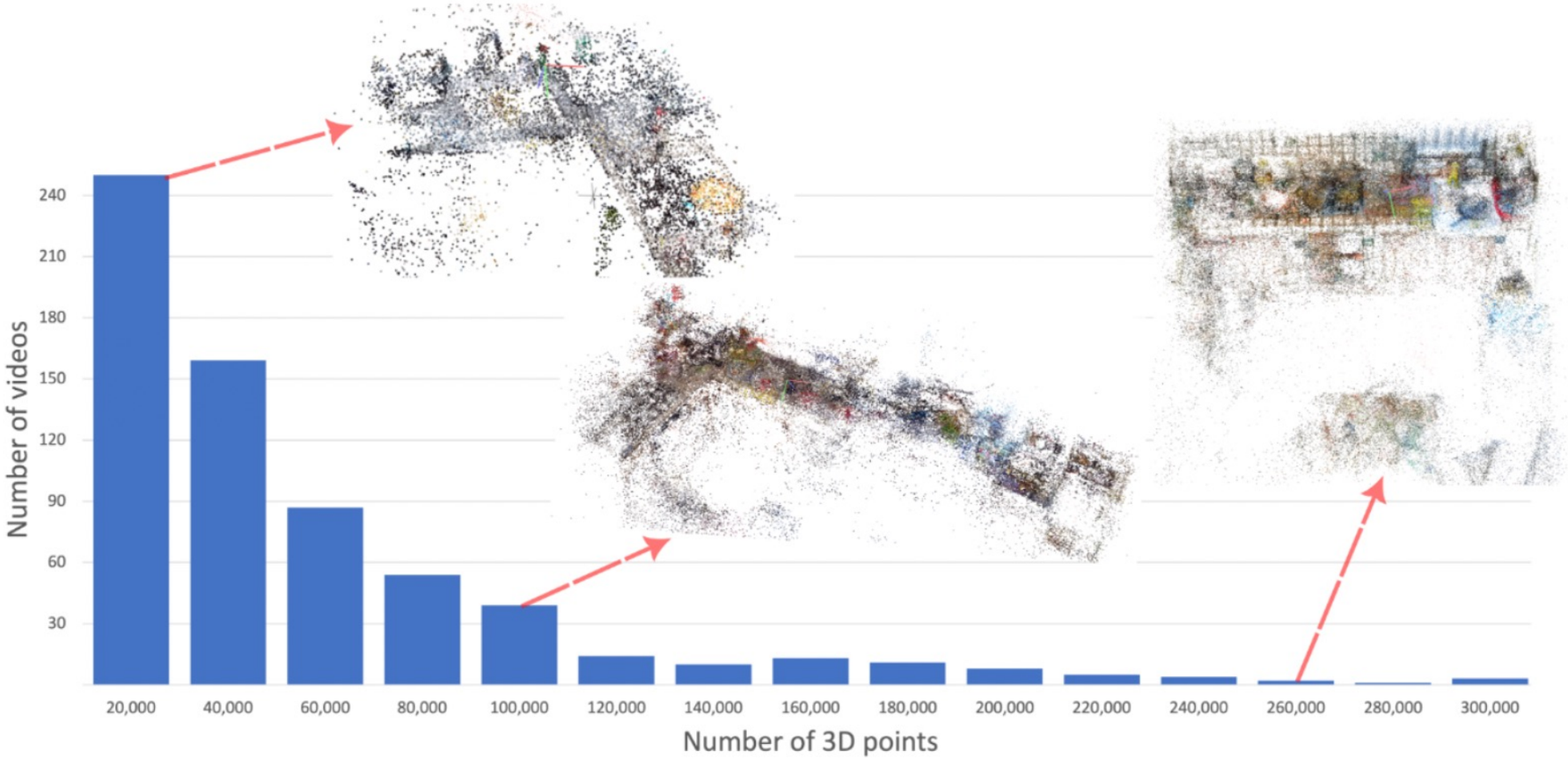


Figure 4: **Number of 3D points histogram.** The majority of our reconstructions generate less than 40,000 points that are enough to represent the kitchen. However, some reconstructions have more than 100,000, we include the point clouds for each points range showing the fine details covered by having more points

EPIC Fields

with: V Tschernezki*, A Darkhalil*, Z Zhu*, D Fouhey, I Laina, D Larlus, A Vedaldi

Table 1: Comparison of datasets commonly used in dynamic new-view synthesis.

Dataset	#Scenes	Seq. Length	Monocular	Semantics
Nerfies [37]	4	8–15 sec	-	-
D-NeRF [41]	8	1–3 sec	-	-
Plenoptic Video [22]	6	10-60 sec	-	-
NVIDIA Dynamic Scene Dataset [65]	12	1–5 sec	4 / 12	-
HyperNeRF [38]	16	8–15 sec	13 / 16	-
iPhone [13]	14	8–15 sec	7 / 14	-
SAFF [25]	8	1–5sec	-	✓
EPIC Fields (ours)	50	6–37 min (Avg 22)	50 / 50	✓

A Darkhalil et al (2022). EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. *NeurIPS*

Video Understanding Out of the Frame

What can we now do with these reconstructions:

- Point Tracking
- Object Tracking
- Gaze Estimation

EgoPoints: Advancing Point Tracking for Egocentric Videos

Ahmad Darkhalil[1] Rhodri Guerrier[1] Adam W. Harley[2] Dima Damen[1]

[1]University of Bristol [2]Stanford University

WACV 2025
TUCSON, ARIZONA • FEB 28 - MAR 4

Spatial Cognition from Egocentric Video:

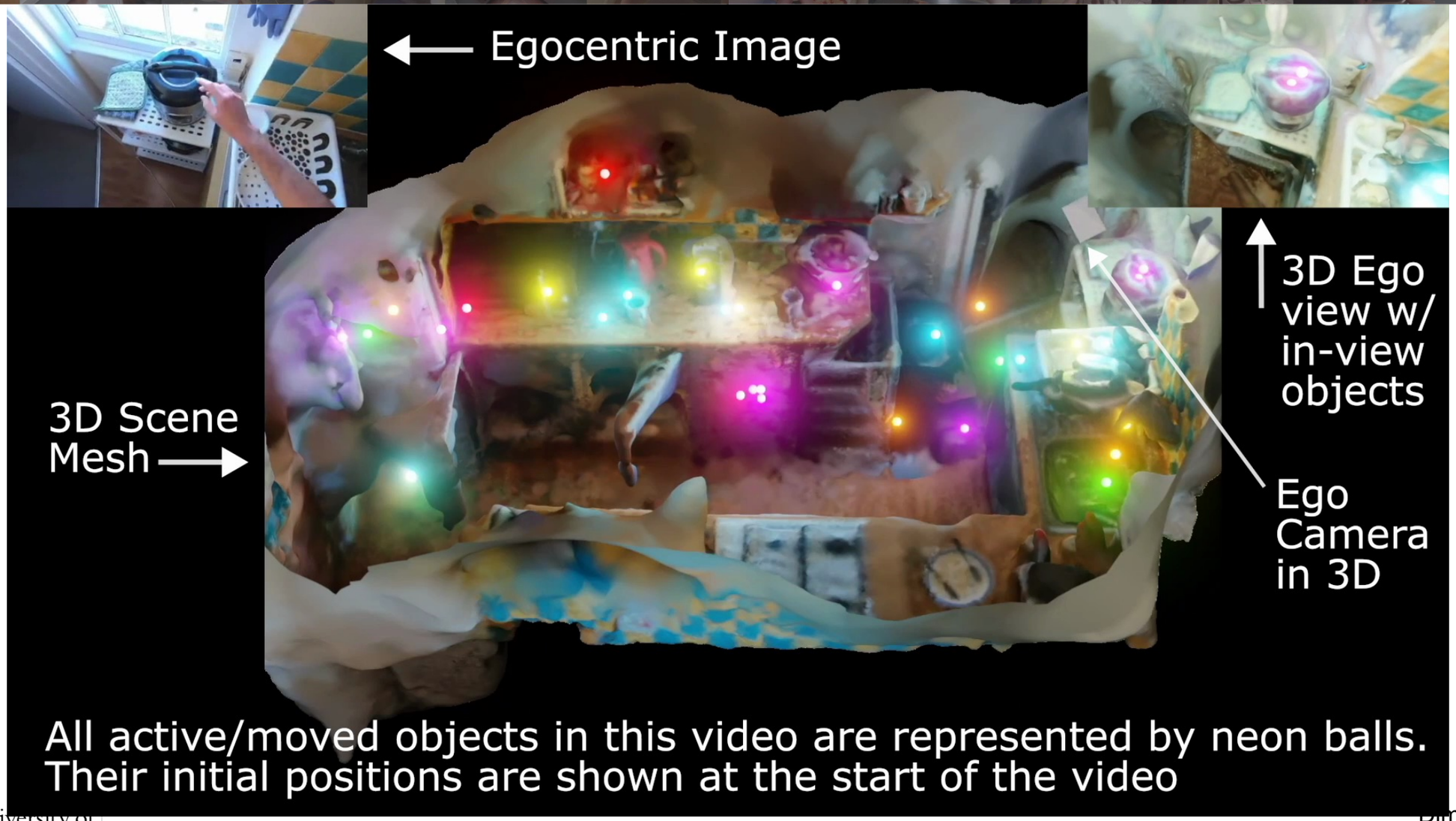
Out of Sight, Not Out of Mind

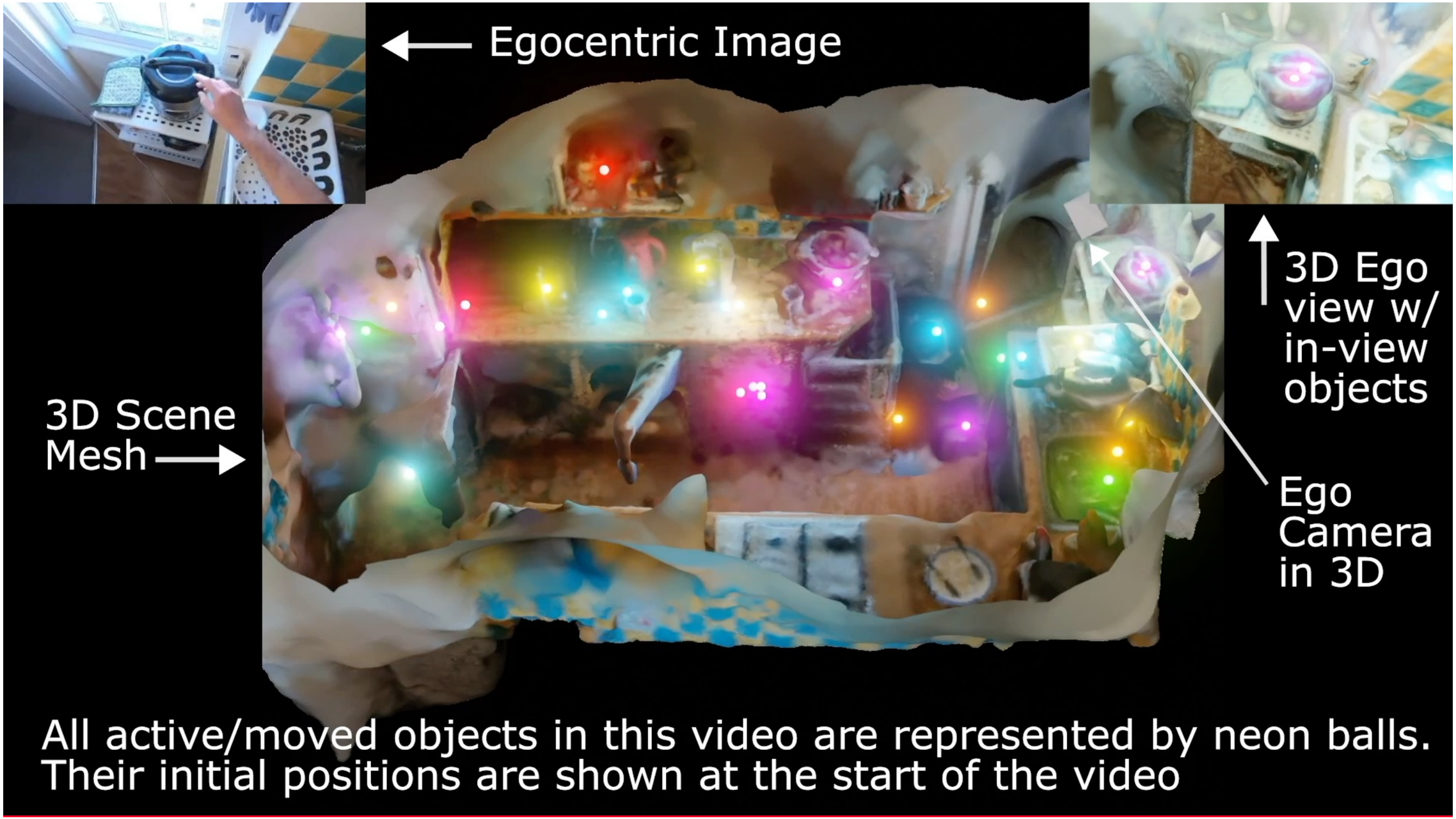
Chiara Plizzari Shubham Goel Toby Perrett Jacob Chalk
Angjoo Kanazawa Dima Damen

http://dimadamen.github.io/OSNOM

Out of Sight, not Out of Mind

with: Chiara Plizzari
Toby Perrett
Shubham Goel
Angjoo Kanazawa

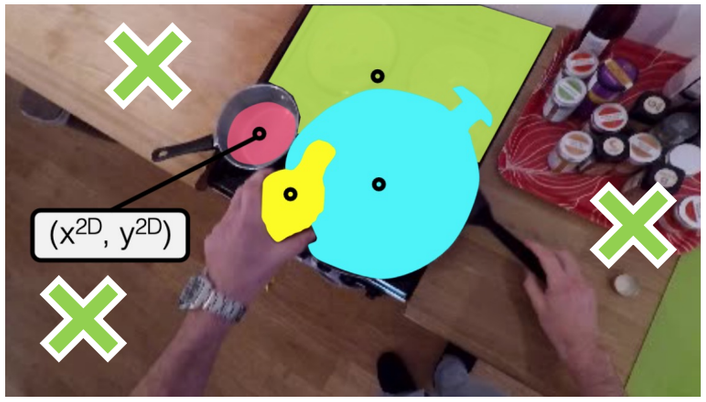


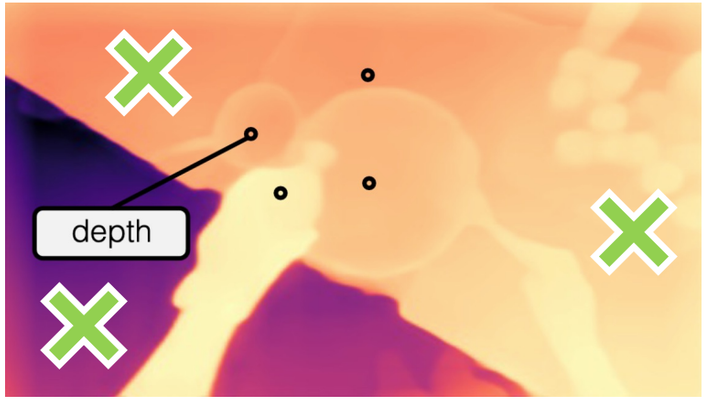

Egocentric Image
3D Ego view w/ in-view objects
3D Scene Mesh
Ego Camera in 3D
All active/moved objects in this video are represented by neon balls.
Their initial positions are shown at the start of the video

Out of Sight, not Out of Mind

with: Chiara Plizzari
Toby Perrett
Shubham Goel
Angjoo Kanazawa

Lift → Match → Keep





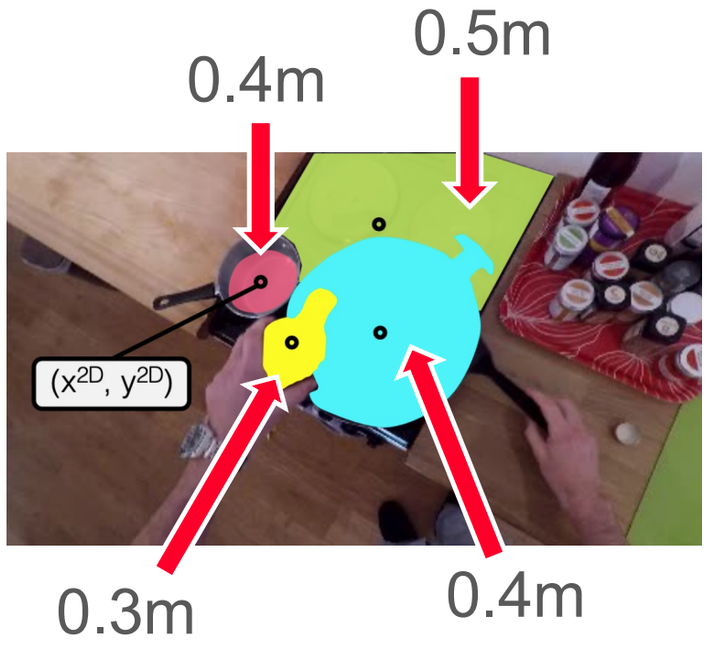
0.0 … 1.0

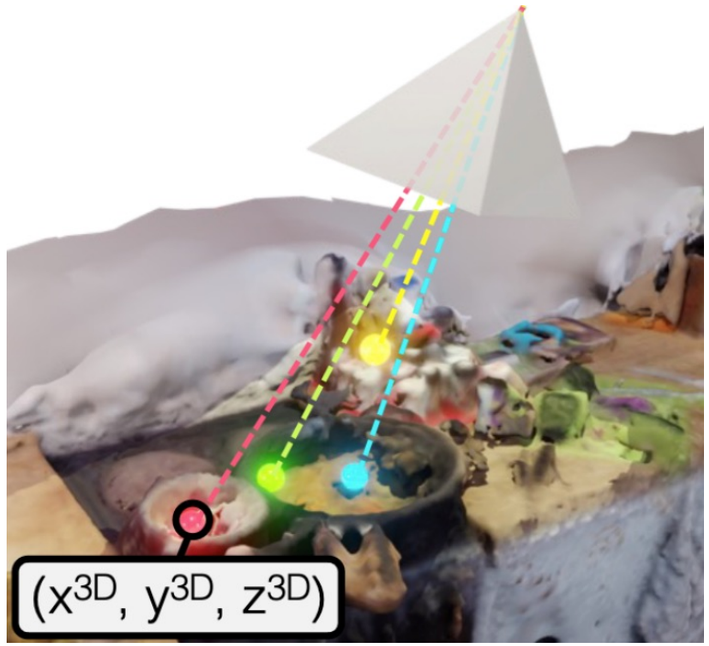
0.3m … 1.8m

Out of Sight, not Out of Mind

with: Chiara Plizzari
Toby Perrett
Shubham Goel
Angjoo Kanazawa

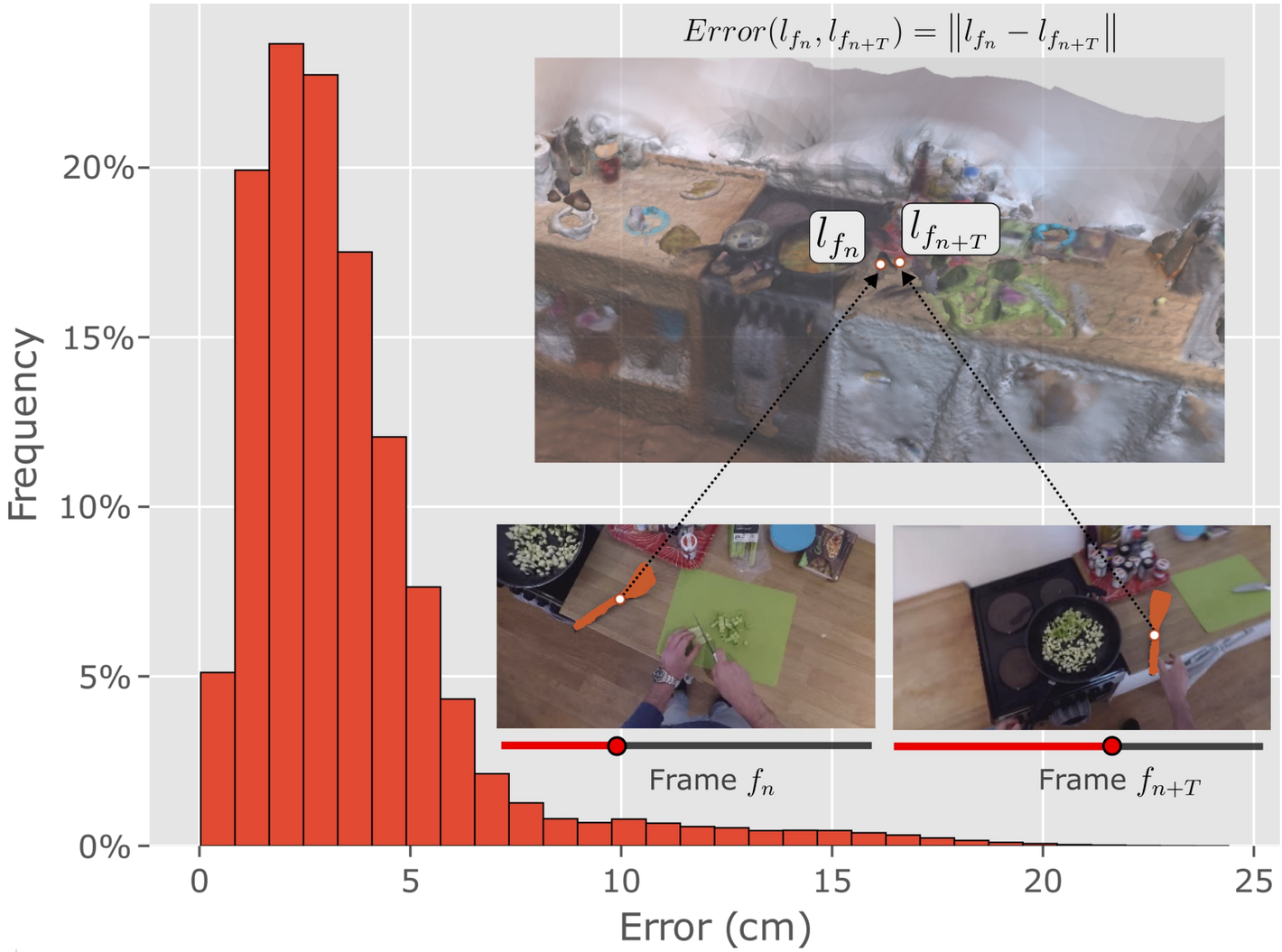
Lift → Match → Keep





Out of Sight, not Out of Mind

with: Chiara Plizzari
Toby Perrett
Shubham Goel
Angjoo Kanazawa



Out of Sight, not Out of Mind

with: Chiara Plizzari
Toby Perrett
Shubham Goel
Angjoo Kanazawa

Instead of tracking in 2D, we track in 3D, using combination of appearance and location distances

Out of Sight, not Out of Mind

with: Chiara Plizzari Shubham Goel
Toby Perrett Angjoo Kanazawa

After we Lift, Match and Keep (LMK), we can reason about an object's visibility and position

- In-View vs Out-of-View
- In-Sight vs Out-of-Sight (Occluded)
- Within-Reach vs Out-of-Reach (defining the camera wearer's near space)

Out of Sight, not Out of Mind

with: Chiara Plizzari Shubham Goel
Toby Perrett Angjoo Kanazawa

After we Lift, Match and Keep (LMK), we can reason about an object's visibility and position

- In-View vs Out-of-View
- In-Sight vs Out-of-Sight (Occluded)
- Within-Reach vs Out-of-Reach (defining the camera wearer's near space)

Out of Sight, not Out of Mind

with: Chiara Plizzari Shubham Goel
Toby Perrett Angjoo Kanazawa

After we Lift, Match and Keep (LMK), we can reason about an object's visibility and position

- In-View vs Out-of-View
- In-Sight vs Out-of-Sight (Occluded)
- Within-Reach vs Out-of-Reach (defining the camera wearer's near space)

Spatial Cognition from Egocentric Video:

Out of Sight, Not Out of Mind

Chiara Plizzari Shubham Toby Perrett Jacob Chalk

Angi Dima Damen

http://dimadamen.github.io/OSNOM

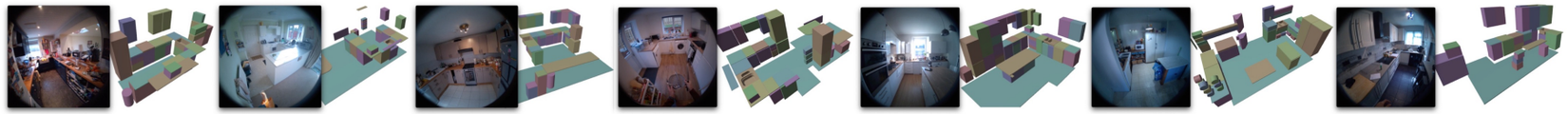
Ground-Truth??

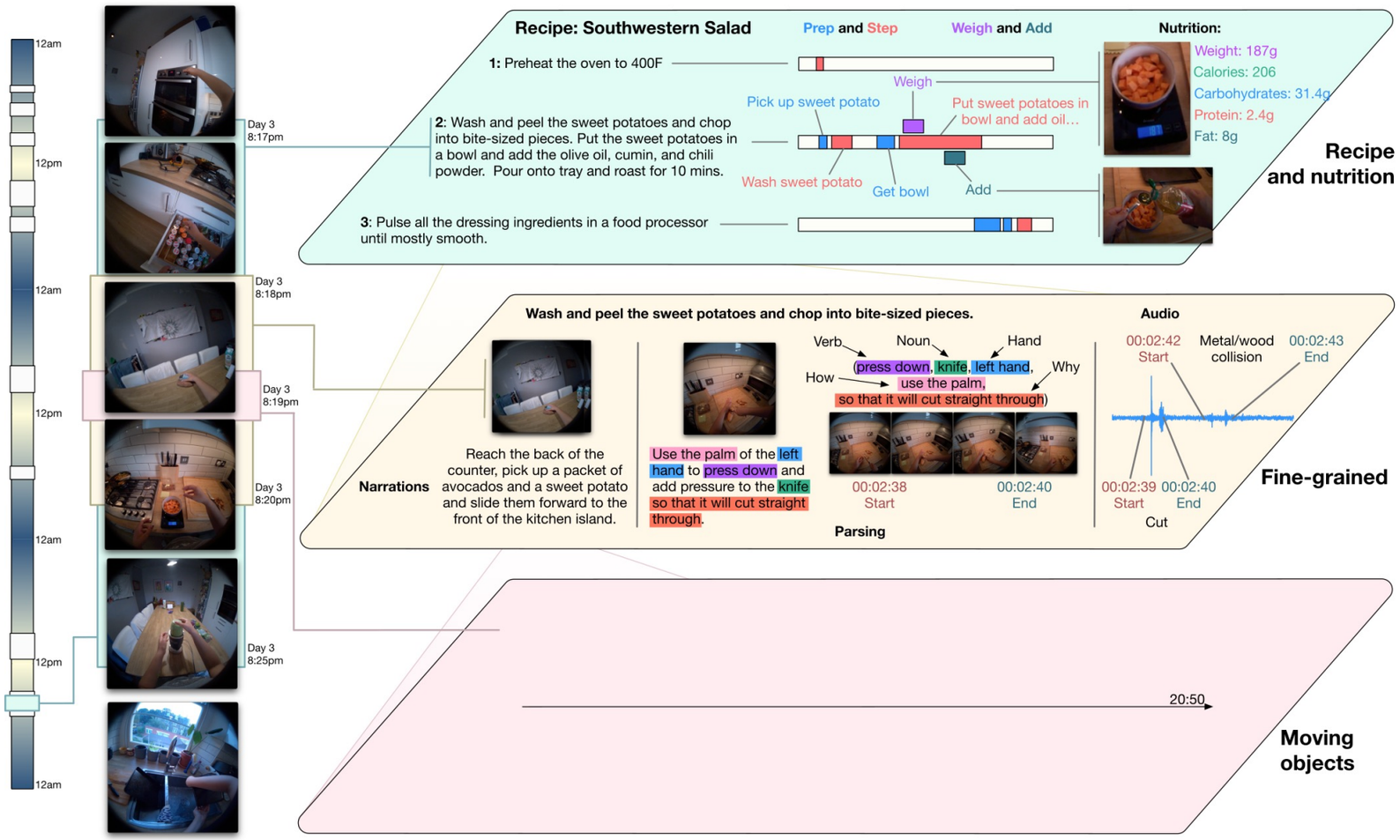
Politecnico di Torino 1859

Berkeley UNIVERSITY OF CALIFORNIA

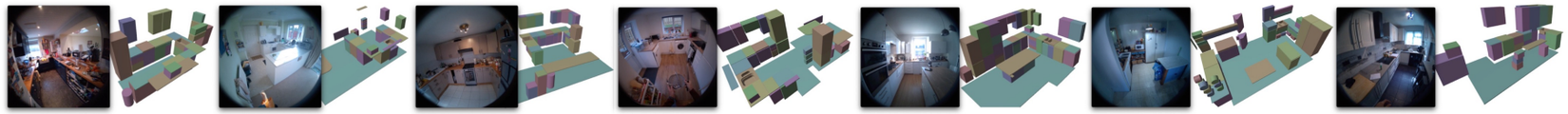
University of BRISTOL

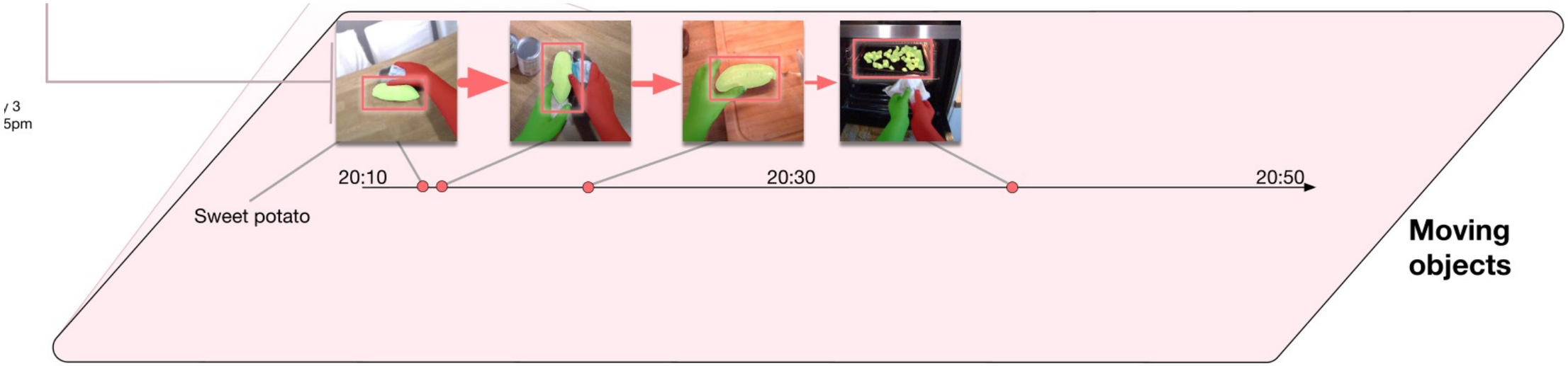
HD-EPIC



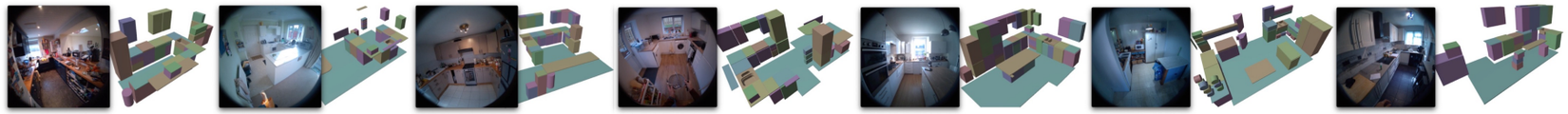

12am
12pm
12am
12pm
12am
12pm
12am
Day 3 8:17pm
Day 3 8:18pm
Day 3 8:19pm
Day 3 8:20pm
Day 3 8:25pm
Recipe: Southwestern Salad
Prep and Step
Weigh and Add
Nutrition:
1: Preheat the oven to 400F
2: Wash and peel the sweet potatoes and chop into bite-sized pieces. Put the sweet potatoes in a bowl and add the olive oil, cumin, and chili powder. Pour onto tray and roast for 10 mins.
3: Pulse all the dressing ingredients in a food processor until mostly smooth.
Weigh
Pick up sweet potato
Put sweet potatoes in bowl and add oil…
Wash sweet potato
Get bowl
Add
Weight: 187g
Calories: 206
Carbohydrates: 31.4g
Protein: 2.4g
Fat: 8g
Recipe and nutrition
Wash and peel the sweet potatoes and chop into bite-sized pieces.
Audio
Narrations
Reach the back of the counter, pick up a packet of avocados and a sweet potato and slide them forward to the front of the kitchen island.
Use the palm of the left hand to press down and add pressure to the knife so that it will cut straight through.
Verb
Noun
Hand
How
Why
(press down, knife, left hand, use the palm, so that it will cut straight through)
00:02:38 Start
00:02:40 End
Parsing
00:02:42 Start
Metal/wood collision
00:02:43 End
00:02:39 Start
00:02:40 End
Cut
Fine-grained
20:50
Moving objects

HD-EPIC

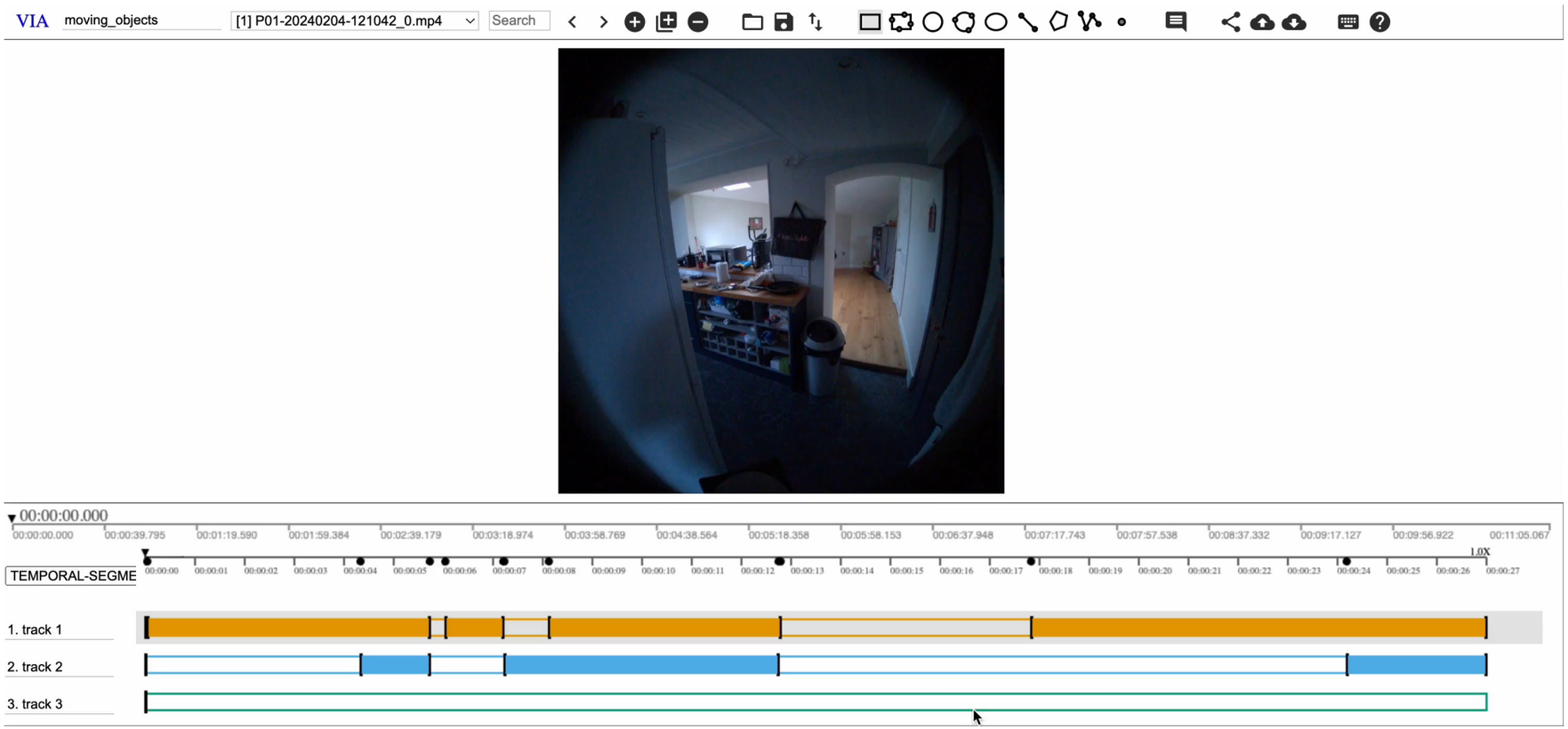




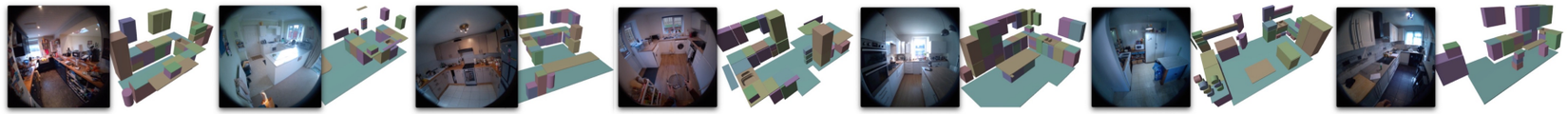
HD-EPIC



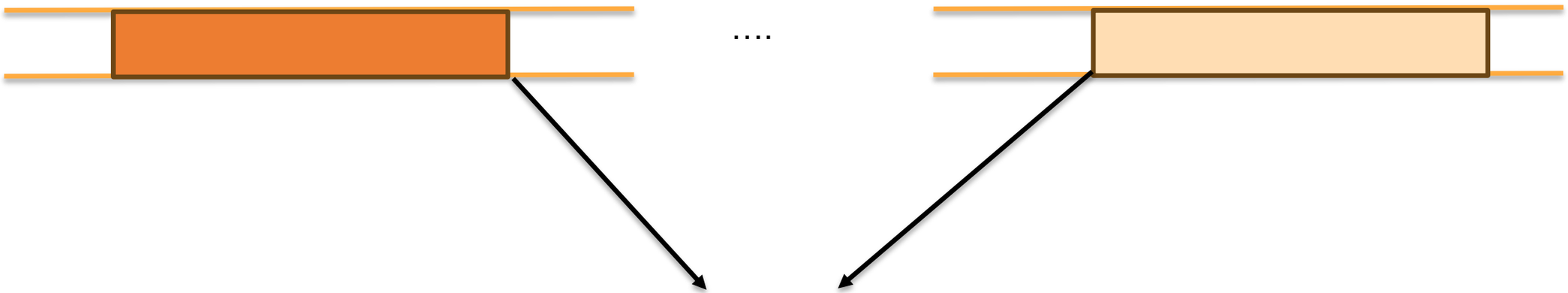
- How to minimize the annotations for tracking objects…



HD-EPIC

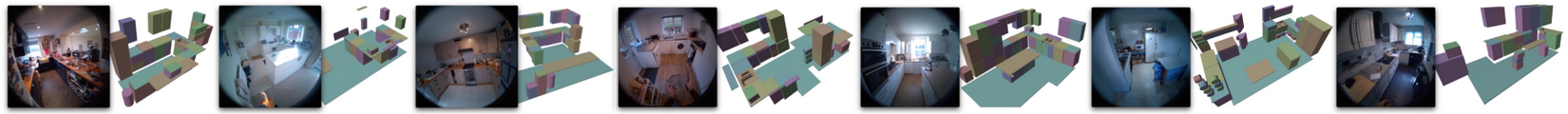


- How to minimize the annotations for tracking objects…



Using appearance & 3D location information to match
Manual confirmation in cases of confusion...

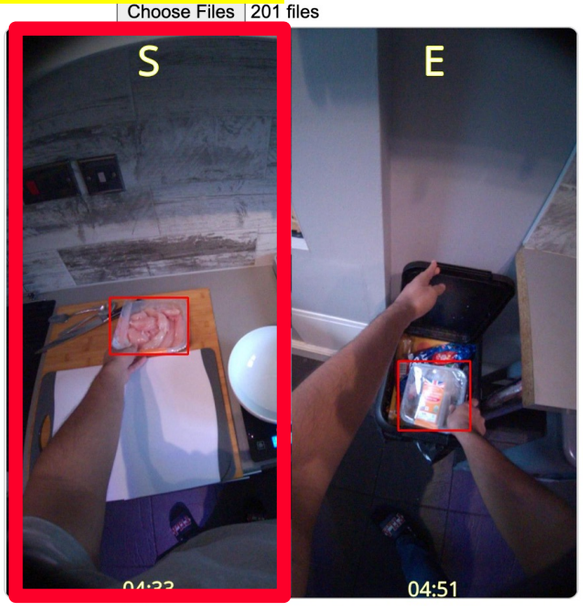
HD-EPIC



Current Track

Image

Choose Files | 201 files



42 / 199

← Previous | Next → | Undo

▼ rubbish bin | box of chicken | wooden chopping board

Enter Track Name (optional)

Create New Track

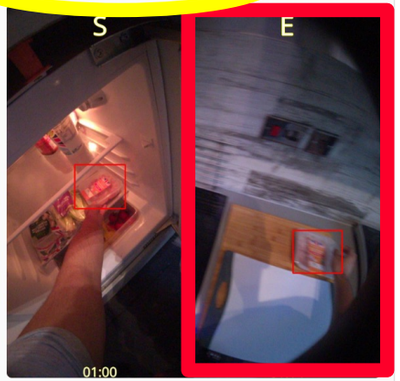
Inconsistent Query

Previous Tracks

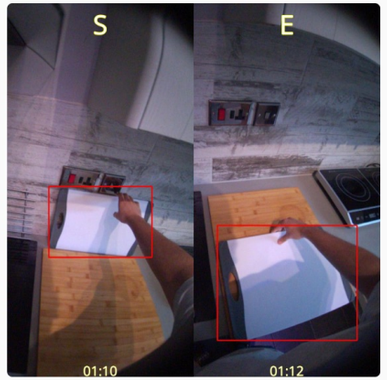
Sort by Distance

Save Tracks

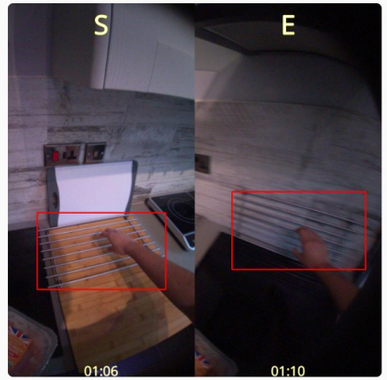
box of chicken (0.0m) Add



plastic chopping board (0.3m) Add

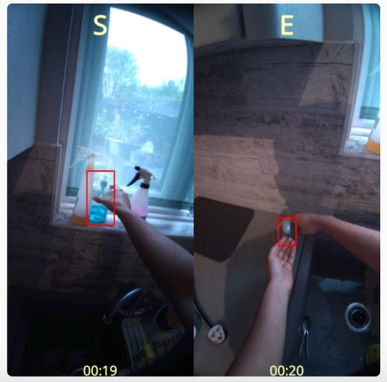


metal cooling rack (0.6m) Add



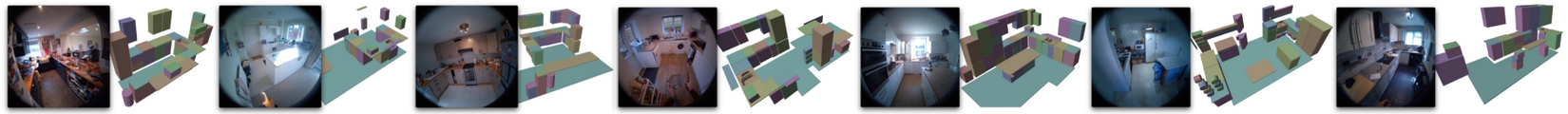
plastic measuring cup (1.0m) Add

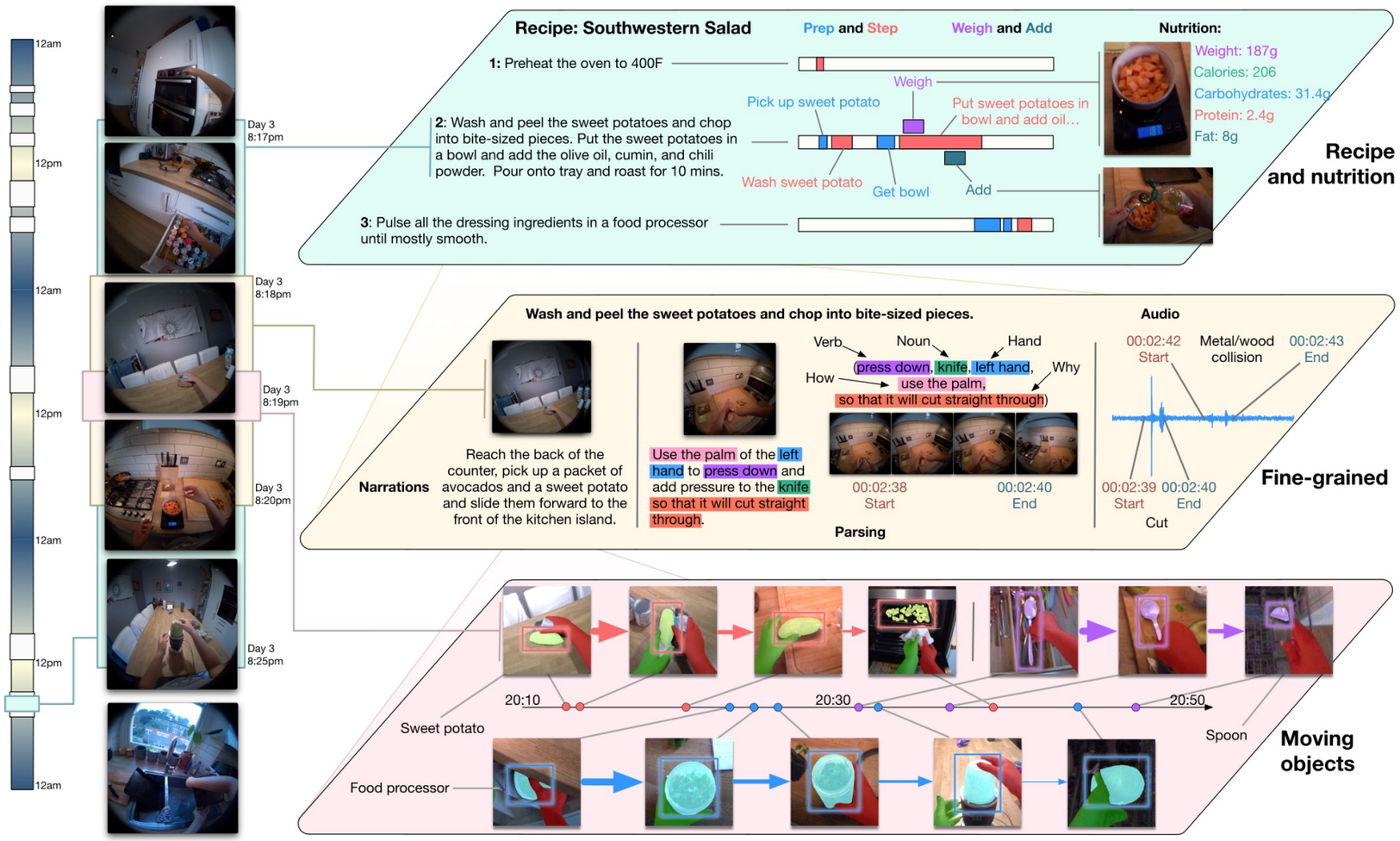
hand washing liquid (1.3m) Add



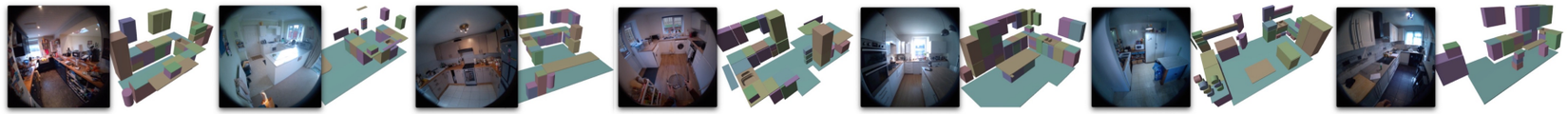
kitchen towel (1.5m) Add

HD-EPIC

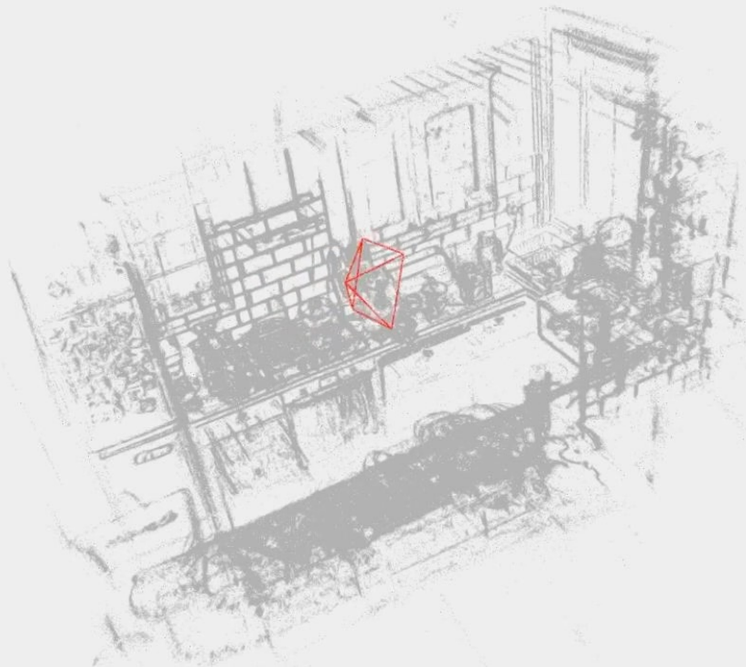



12am
12pm
12am
12pm
12am
12pm
12am
Day 3 8:17pm
Day 3 8:18pm
Day 3 8:19pm
Day 3 8:20pm
Day 3 8:25pm
Recipe: Southwestern Salad
Prep and Step
Weigh and Add
Nutrition:
1: Preheat the oven to 400F
2: Wash and peel the sweet potatoes and chop into bite-sized pieces. Put the sweet potatoes in a bowl and add the olive oil, cumin, and chili powder. Pour onto tray and roast for 10 mins.
3: Pulse all the dressing ingredients in a food processor until mostly smooth.
Weigh
Pick up sweet potato
Put sweet potatoes in bowl and add oil…
Wash sweet potato
Get bowl
Add
Weight: 187g
Calories: 206
Carbohydrates: 31.4g
Protein: 2.4g
Fat: 8g
Recipe and nutrition
Wash and peel the sweet potatoes and chop into bite-sized pieces.
Audio
Narrations
Reach the back of the counter, pick up a packet of avocados and a sweet potato and slide them forward to the front of the kitchen island.
Use the palm of the left hand to press down and add pressure to the knife so that it will cut straight through.
Verb
Noun
Hand
How
Why
(press down, knife, left hand, use the palm, so that it will cut straight through)
00:02:38 Start
00:02:40 End
Parsing
00:02:42 Start
Metal/wood collision
00:02:43 End
00:02:39 Start
00:02:40 End
Cut
Fine-grained
20:10
20:30
20:50
Sweet potato
Spoon
Food processor
Moving objects

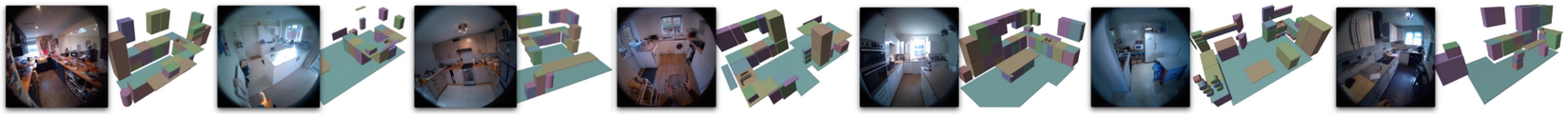
HD-EPIC

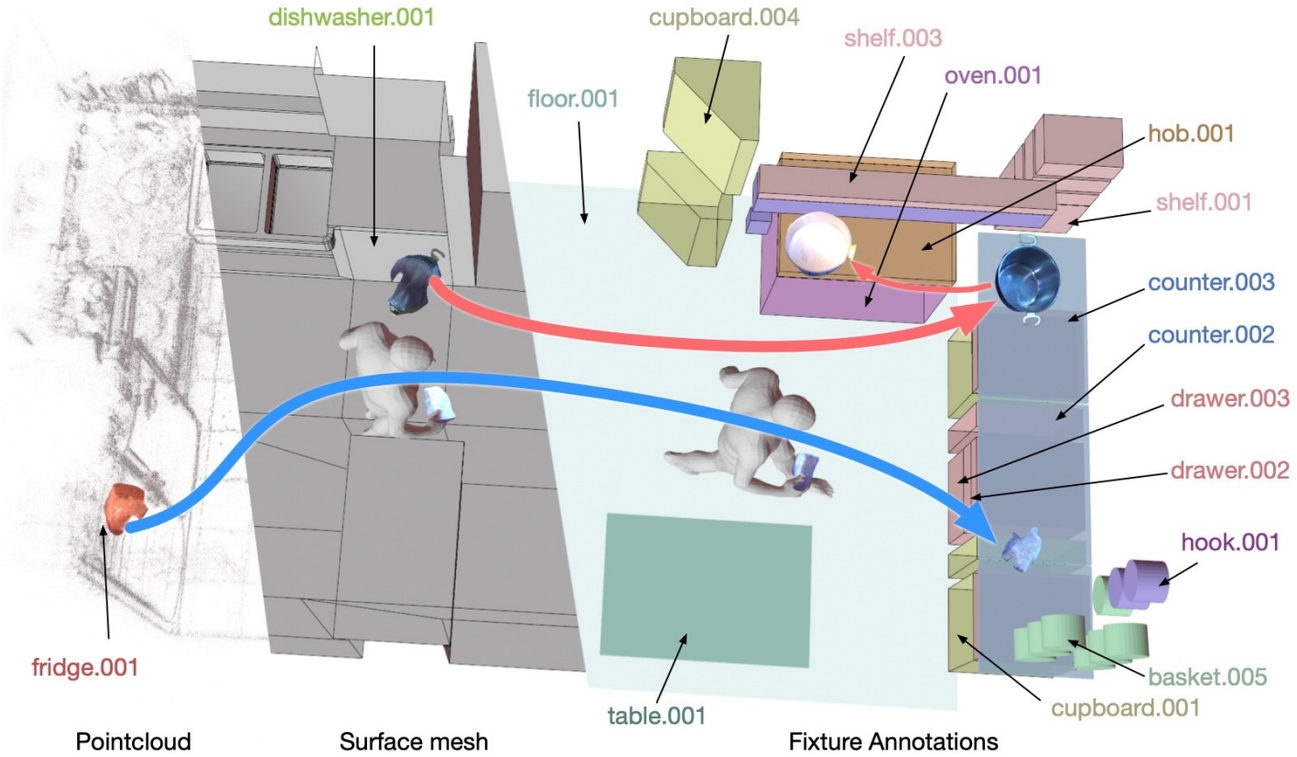


Digital Twin



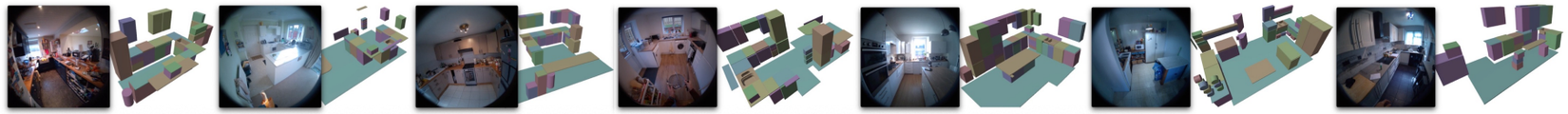
HD-EPIC





Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

HD-EPIC



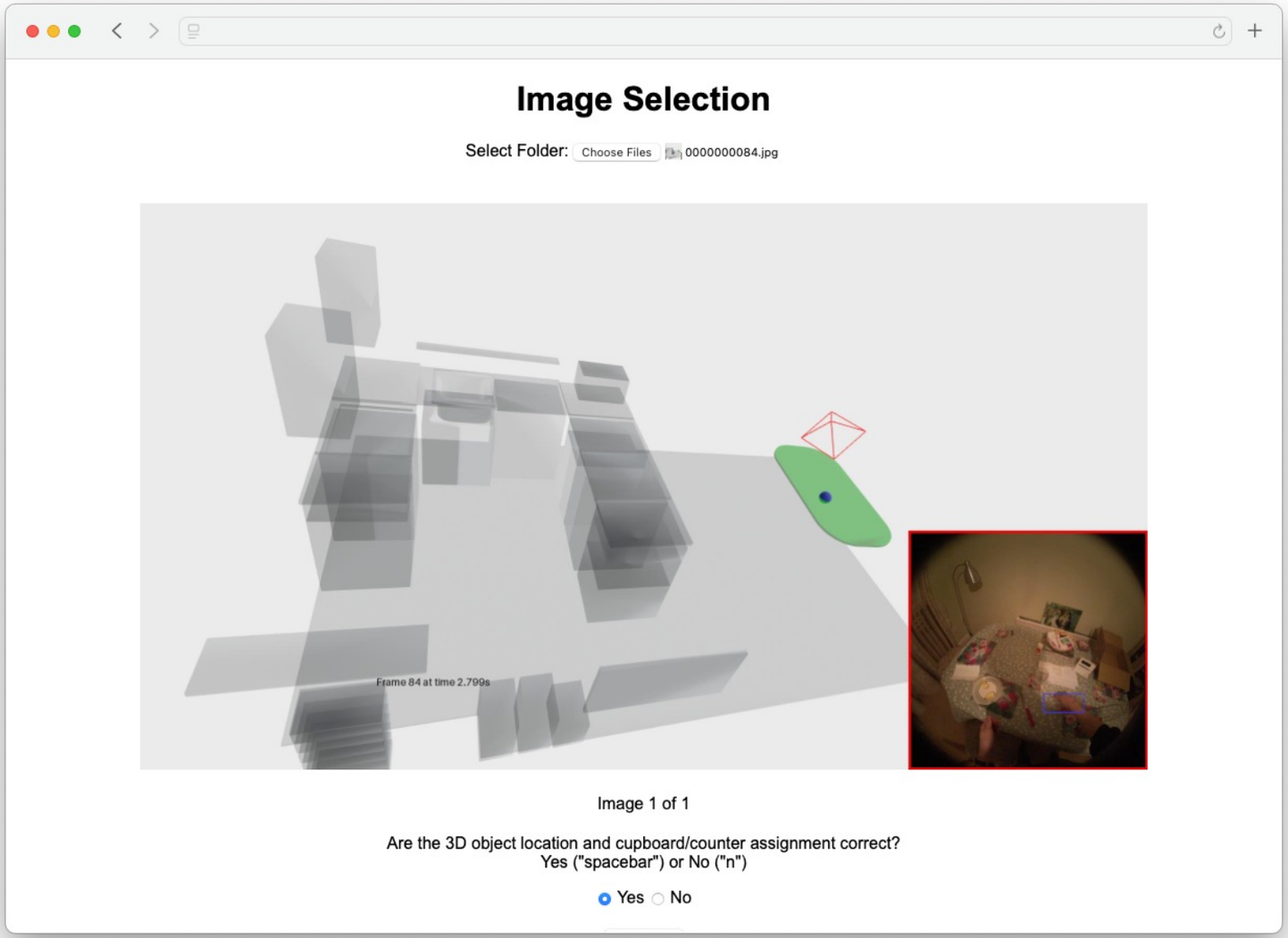
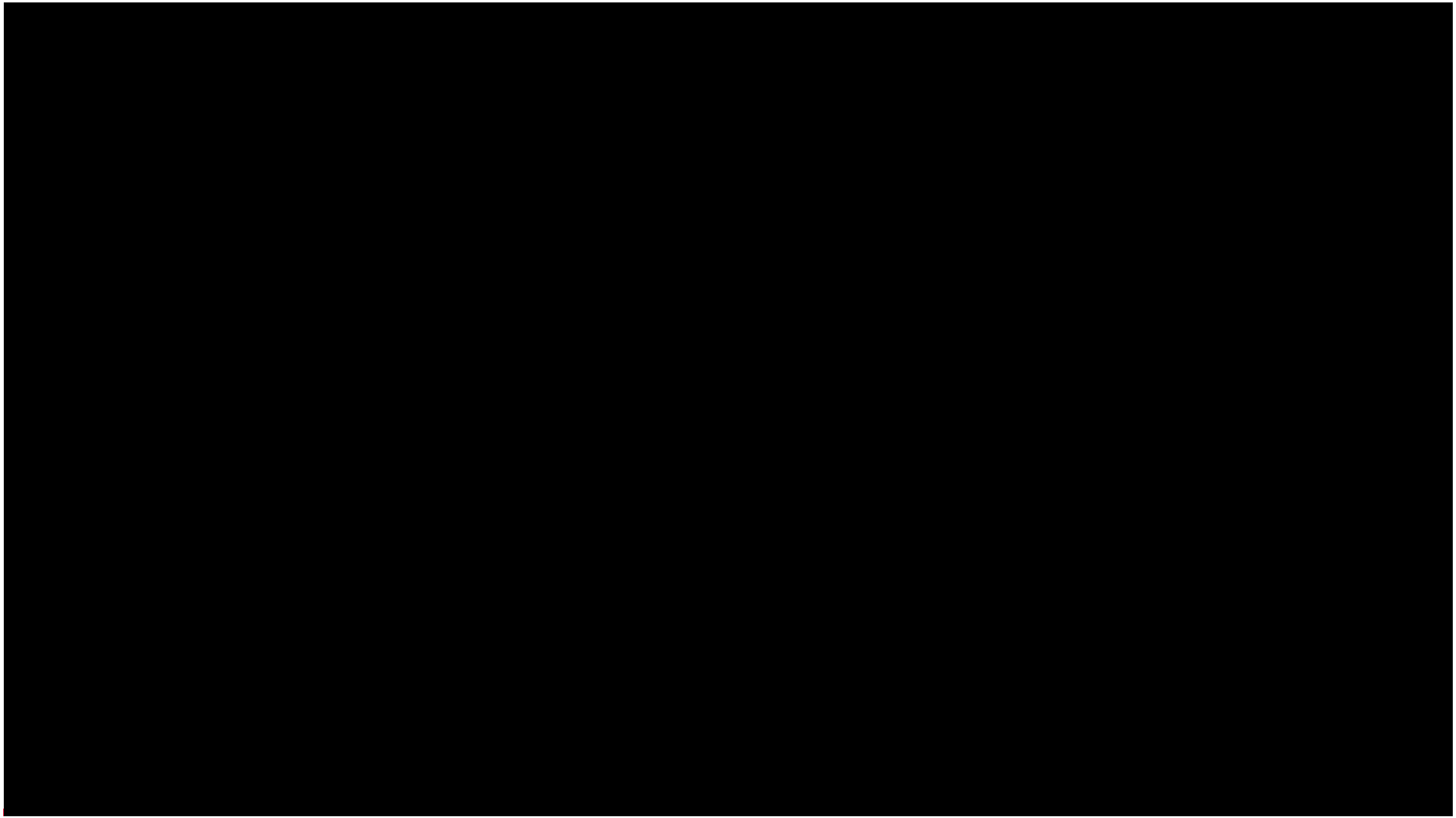
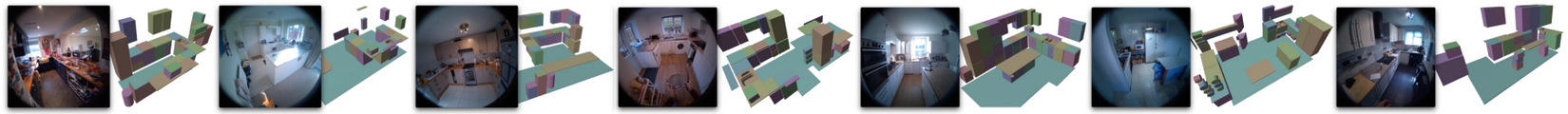


Image Selection
Select Folder:
Choose Files
0000000084.jpg
Frame 84 at time 2.799s
Image 1 of 1
Are the 3D object location and cupboard/counter assignment correct?
Yes ("spacebar") or No ("n")
Yes
No

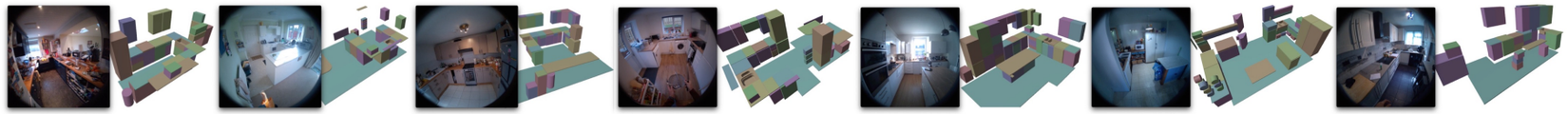




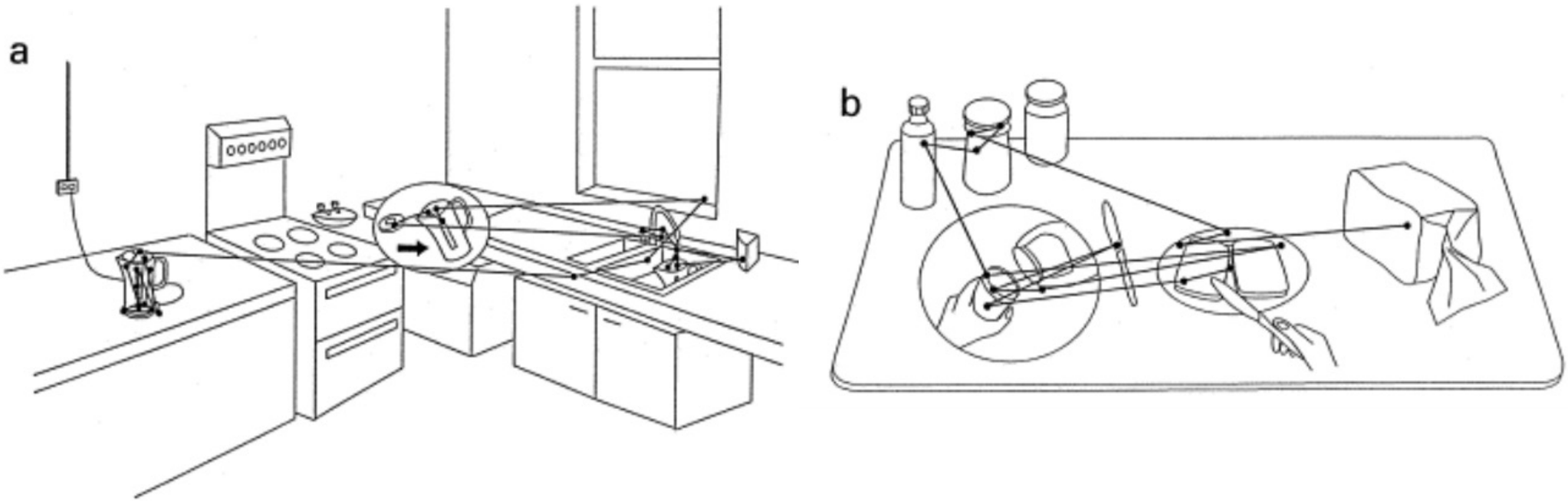
Video Understanding Out of the Frame

What can we now do with these reconstructions:

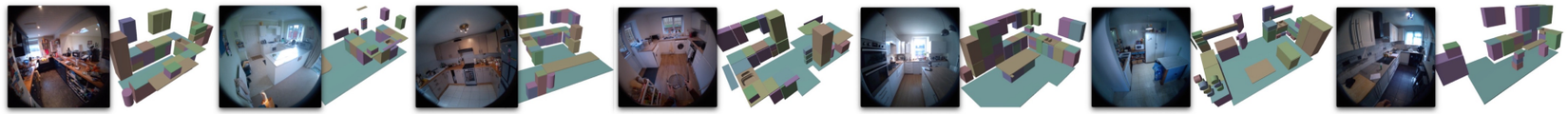
- Point Tracking
- Object Tracking
- Gaze Estimation



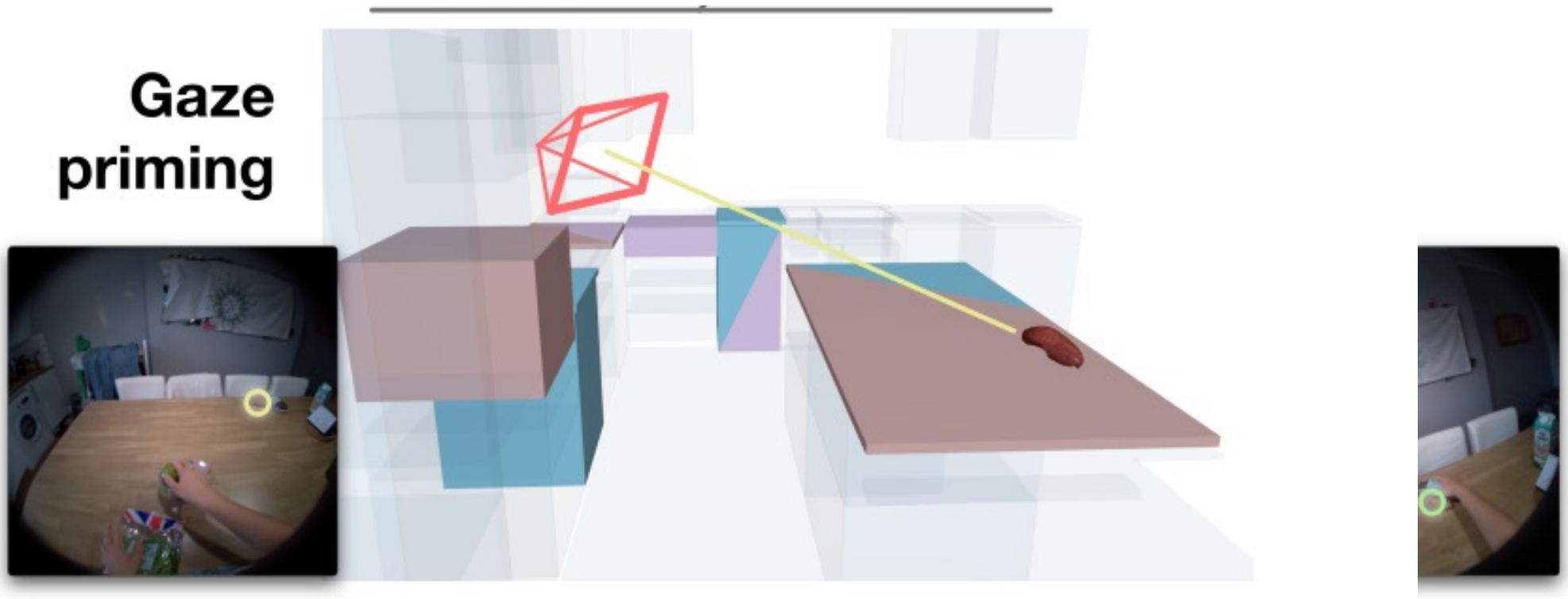
Gaze and Fixations



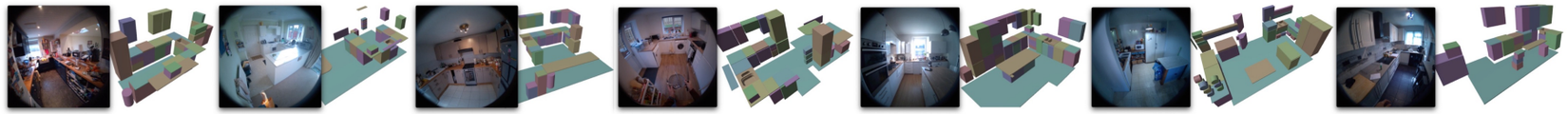
HD-EPIC



Gaze priming



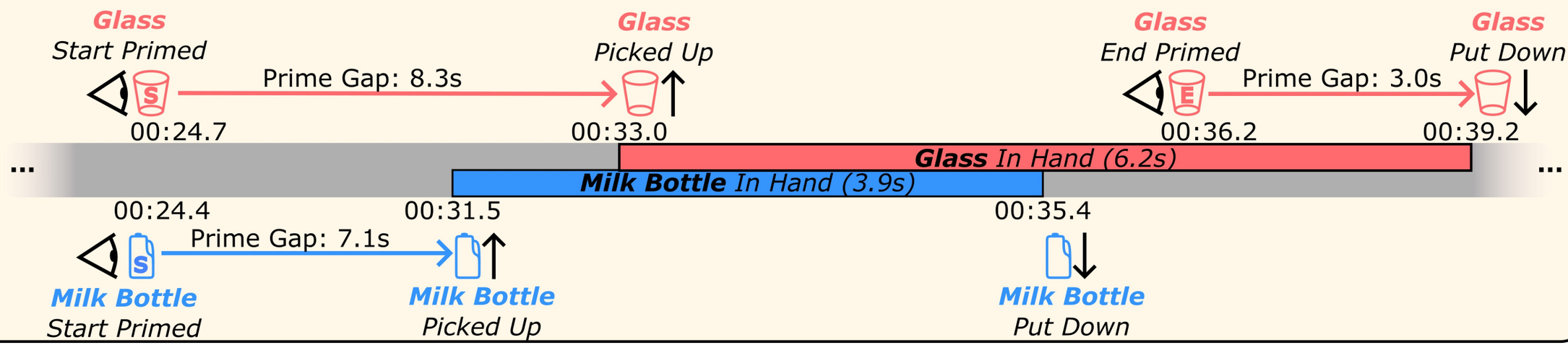
HD-EPIC



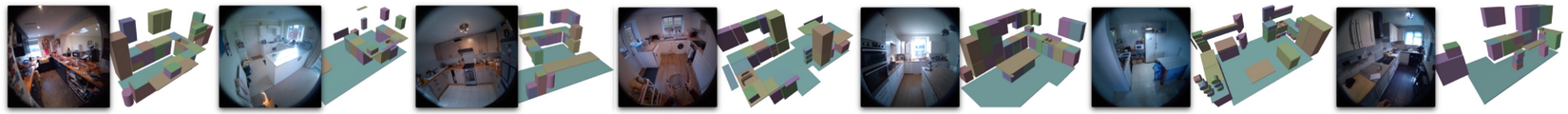
3D Scene

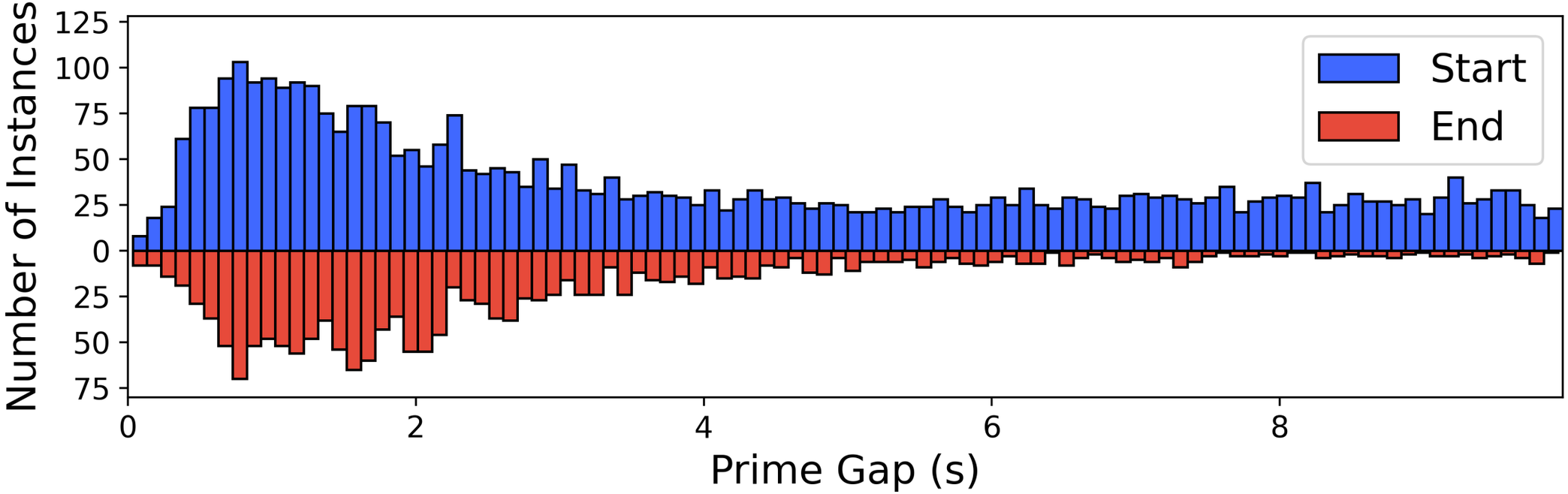
Frames w/ 2D Gaze

Object Movement



HD-EPIC



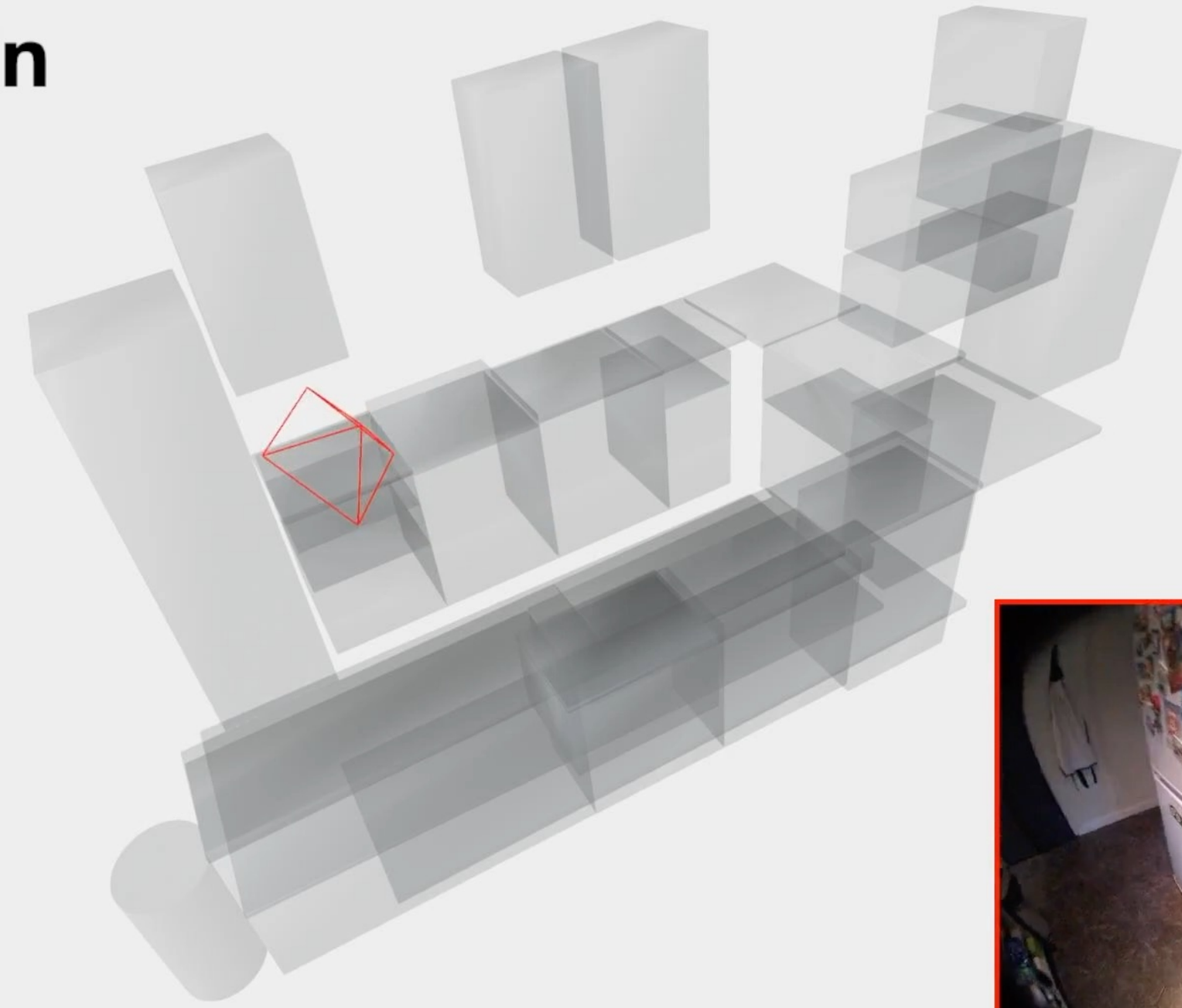


Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

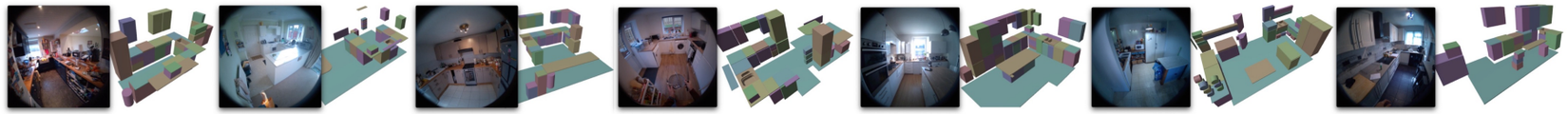
Digital Twin

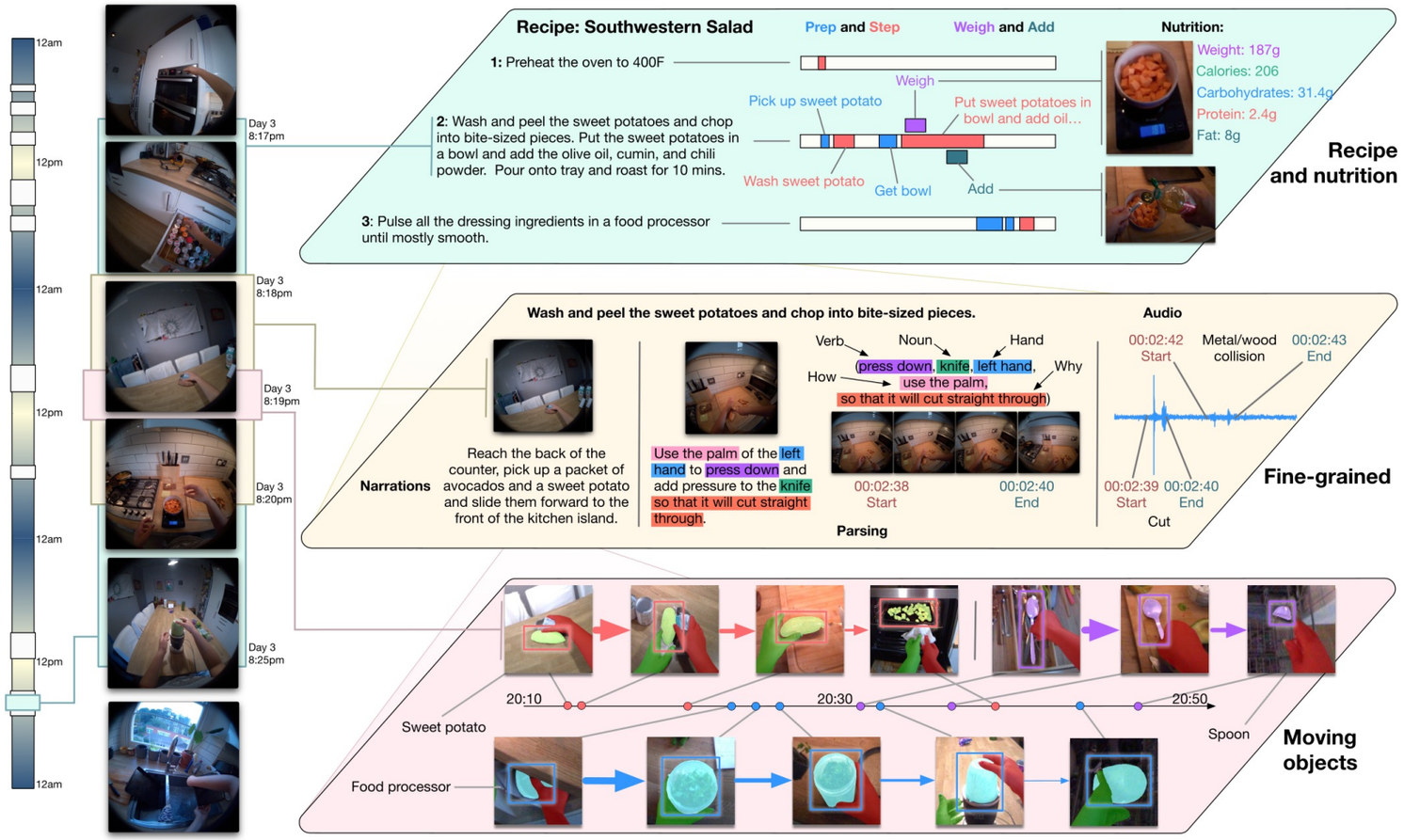
Fixtures

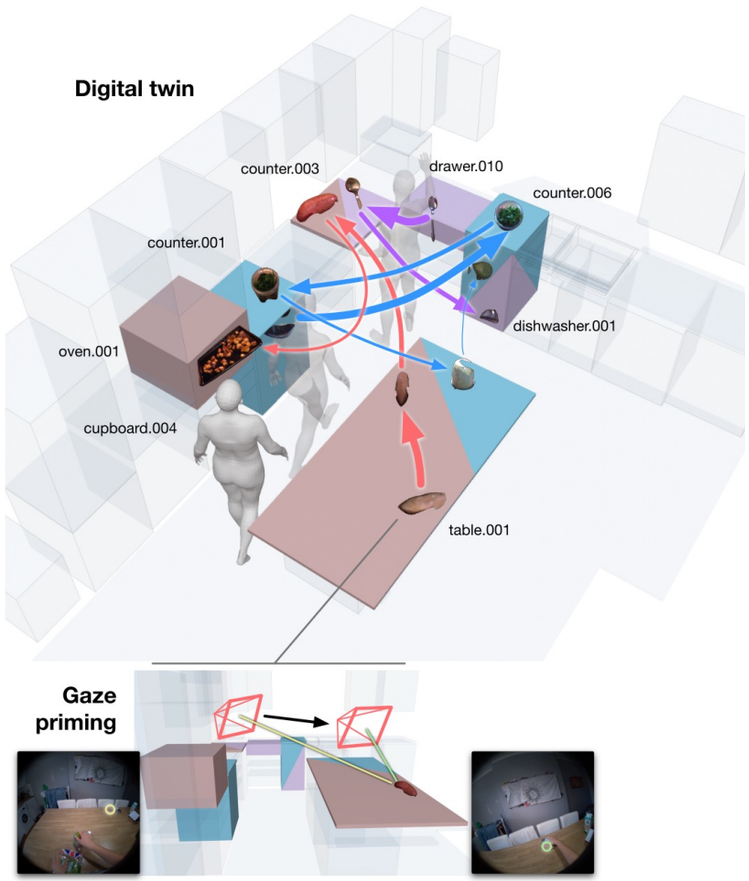
Open drawer



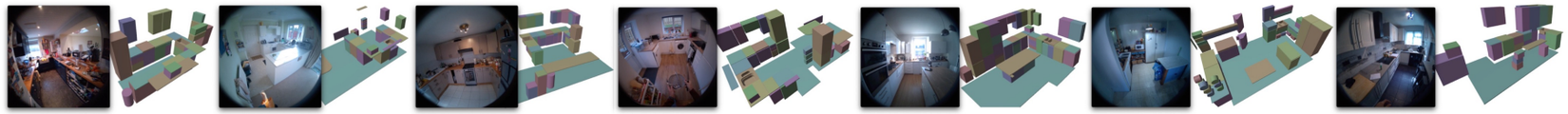
HD-EPIC




12am
12pm
12am
12pm
12am
12pm
12am
Day 3 8:17pm
Day 3 8:18pm
Day 3 8:19pm
Day 3 8:20pm
Day 3 8:25pm
Recipe: Southwestern Salad
Prep and Step
Weigh and Add
Nutrition:
1: Preheat the oven to 400F
2: Wash and peel the sweet potatoes and chop into bite-sized pieces. Put the sweet potatoes in a bowl and add the olive oil, cumin, and chili powder. Pour onto tray and roast for 10 mins.
3: Pulse all the dressing ingredients in a food processor until mostly smooth.
Weigh
Pick up sweet potato
Put sweet potatoes in bowl and add oil…
Wash sweet potato
Get bowl
Add
Weight: 187g
Calories: 206
Carbohydrates: 31.4g
Protein: 2.4g
Fat: 8g
Recipe and nutrition
Wash and peel the sweet potatoes and chop into bite-sized pieces.
Audio
Verb
Noun
Hand
How
Why
(press down, knife, left hand, use the palm, so that it will cut straight through)
00:02:42 Start
Metal/wood collision
00:02:43 End
Narrations
Reach the back of the counter, pick up a packet of avocados and a sweet potato and slide them forward to the front of the kitchen island.
Use the palm of the left hand to press down and add pressure to the knife so that it will cut straight through.
00:02:38 Start
00:02:40 End
Parsing
00:02:39 Start
00:02:40 End
Cut
Fine-grained
20:10
20:30
20:50
Sweet potato
Spoon
Food processor
Moving objects


Digital twin
counter.003
drawer.010
counter.006
counter.001
dishwasher.001
oven.001
cupboard.004
table.001
Gaze priming

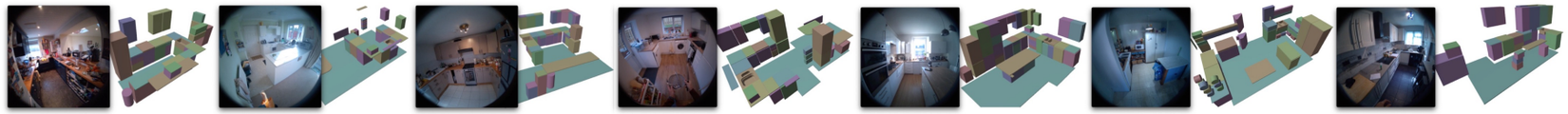
HD-EPIC

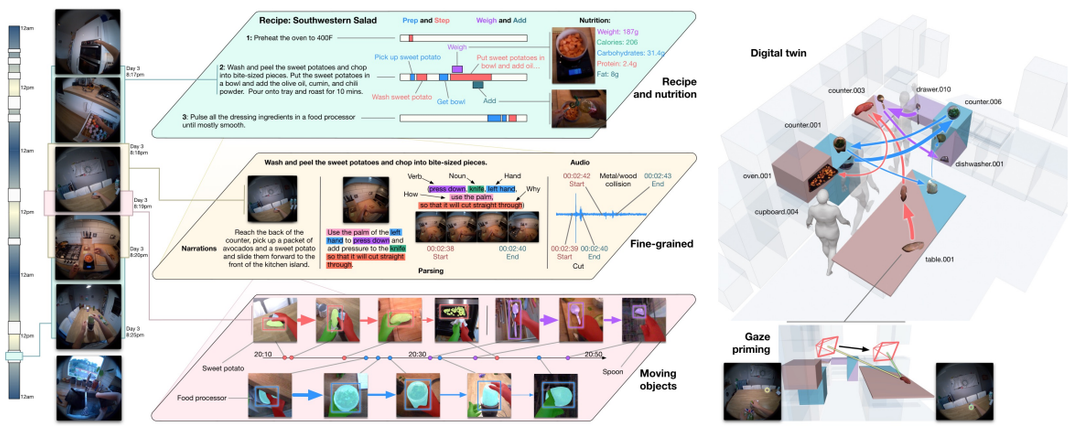


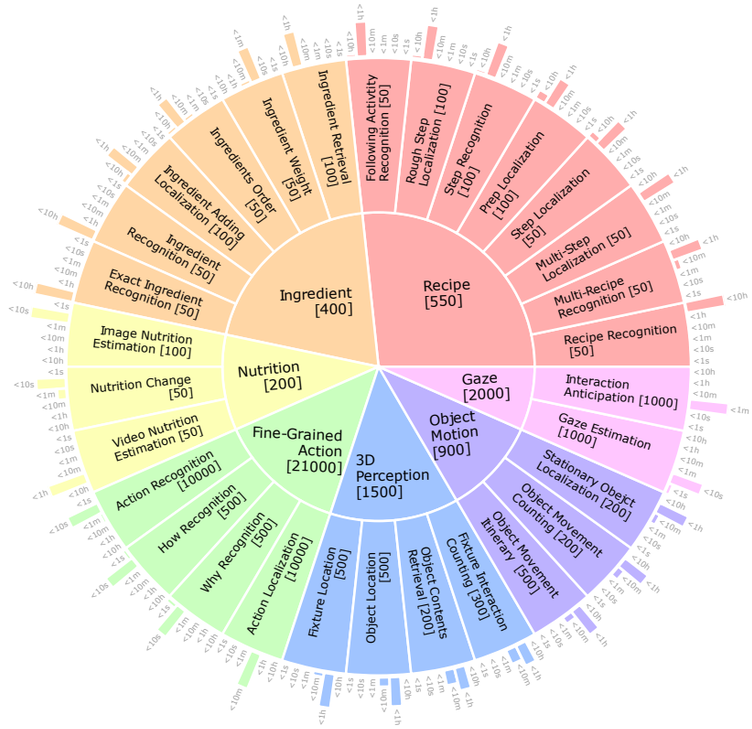
Annotation Type	Total annotations	Annotations/min
Narrations	59,454	24.0
Parsing (Verbs + Nouns + Hands + How + Why)	303,968	122.7
Recipes (Preps + Steps)	4,052	1.6
Sound	50,968	20.6
Action boundaries	59,454	24.0
Object Motion (Pick up + Put down + Fixtures + Bboxes + Masks)	153,480	62.0
Object Itinerary	4,881	2.0
Object Priming (Starts + Ends)	18,264	7.4
Total		263.2

Table A3. HD-EPIC annotations per minute

HD-EPIC



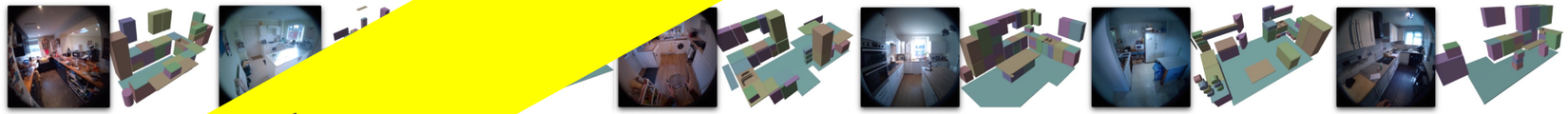




Sec 1: Highly-Detailed Dataset

Sec 2: HD-EPIC VQA Benchmark

HD-EPIC



Try it Yourself

Use Wise to Search through HD-EPIC

SCAN ME

https://meru.robots.ox.ac.uk/HD-EPIC/

Video Understanding Out of the Frame

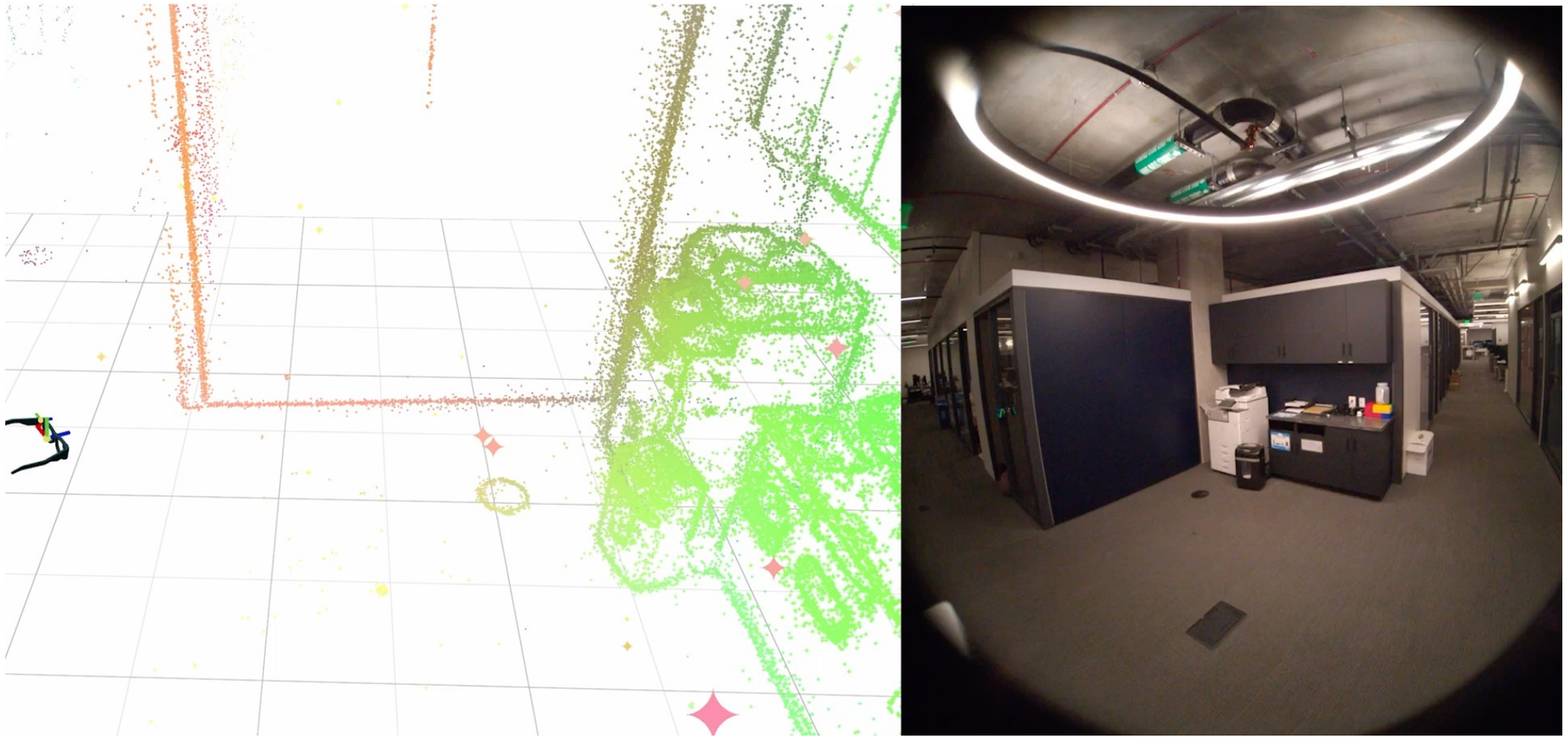
Body and Hand Motion Estimation “out of the frame”

University of BRISTOL

Dima Damen
PAISS 2025

EgoBody

EgoAllo uses egocentric ( ) SLAM poses and images



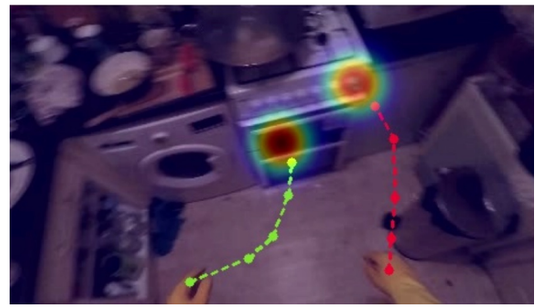
EgoHand Forecasting – Previous Works

with: Masashi Hatano
Zhifan Zhu
Hideo Saito

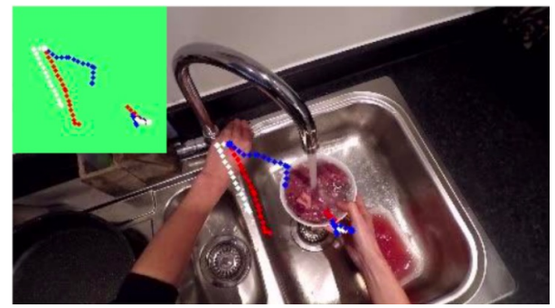
2D Hand Forecasting

Given an egocentric video,
forecast 2D hand positions of both hands

→ Limited in 2D image plane



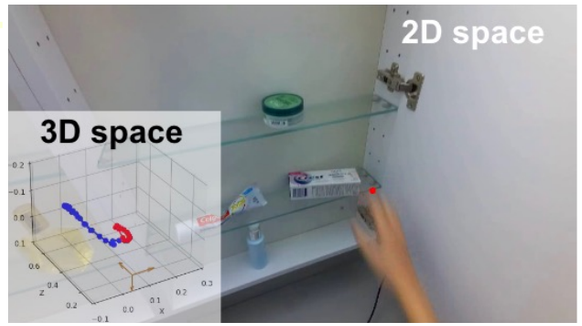
OCT [CVPR'22]



Diff-IP2D [IROS'25]

3D Hand Forecasting

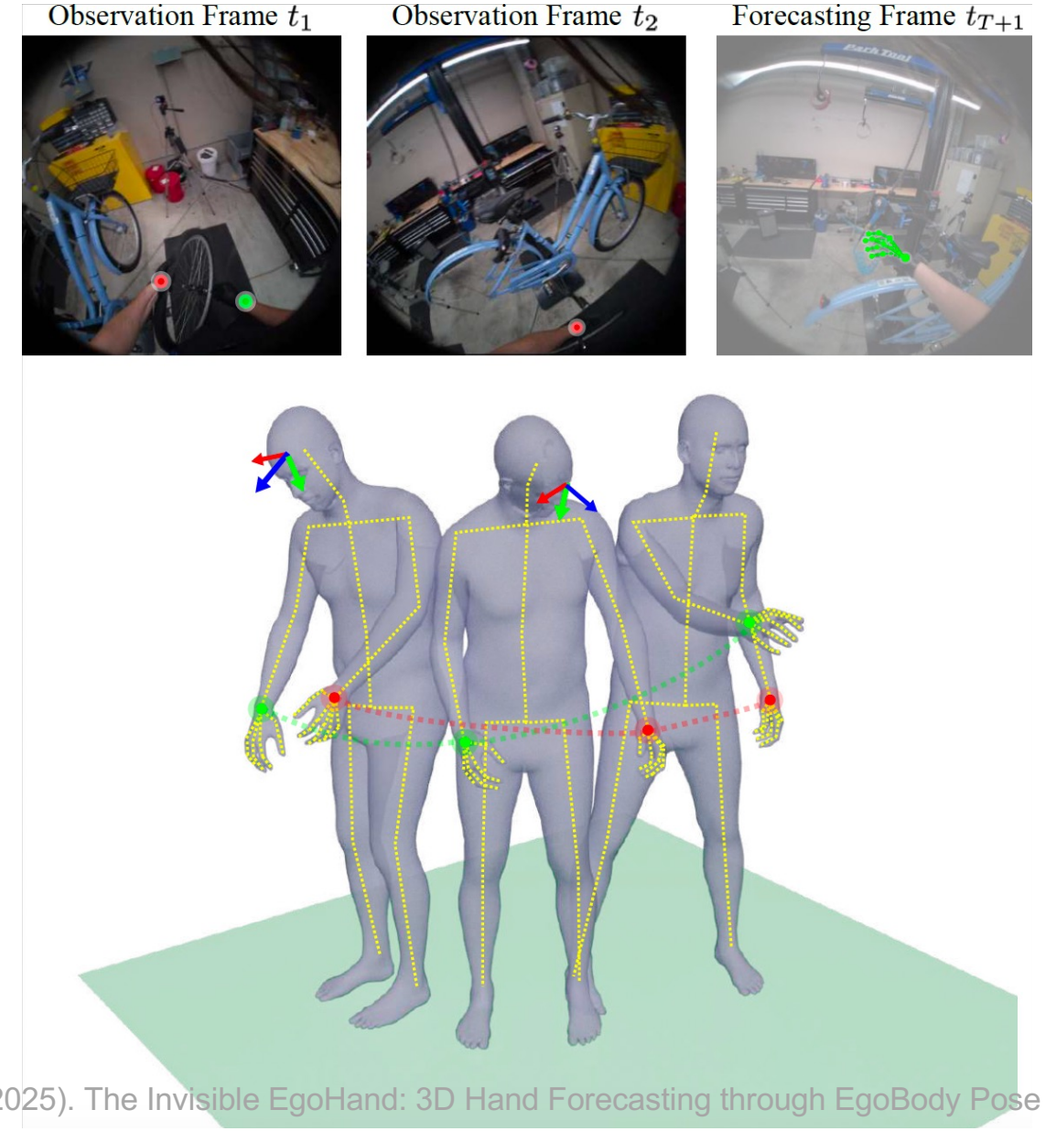
Given an egocentric video & 3D hand trajectory,
forecast 3D hand positions of one hand



USST [ICCV'23]

The Invisible EgoHand

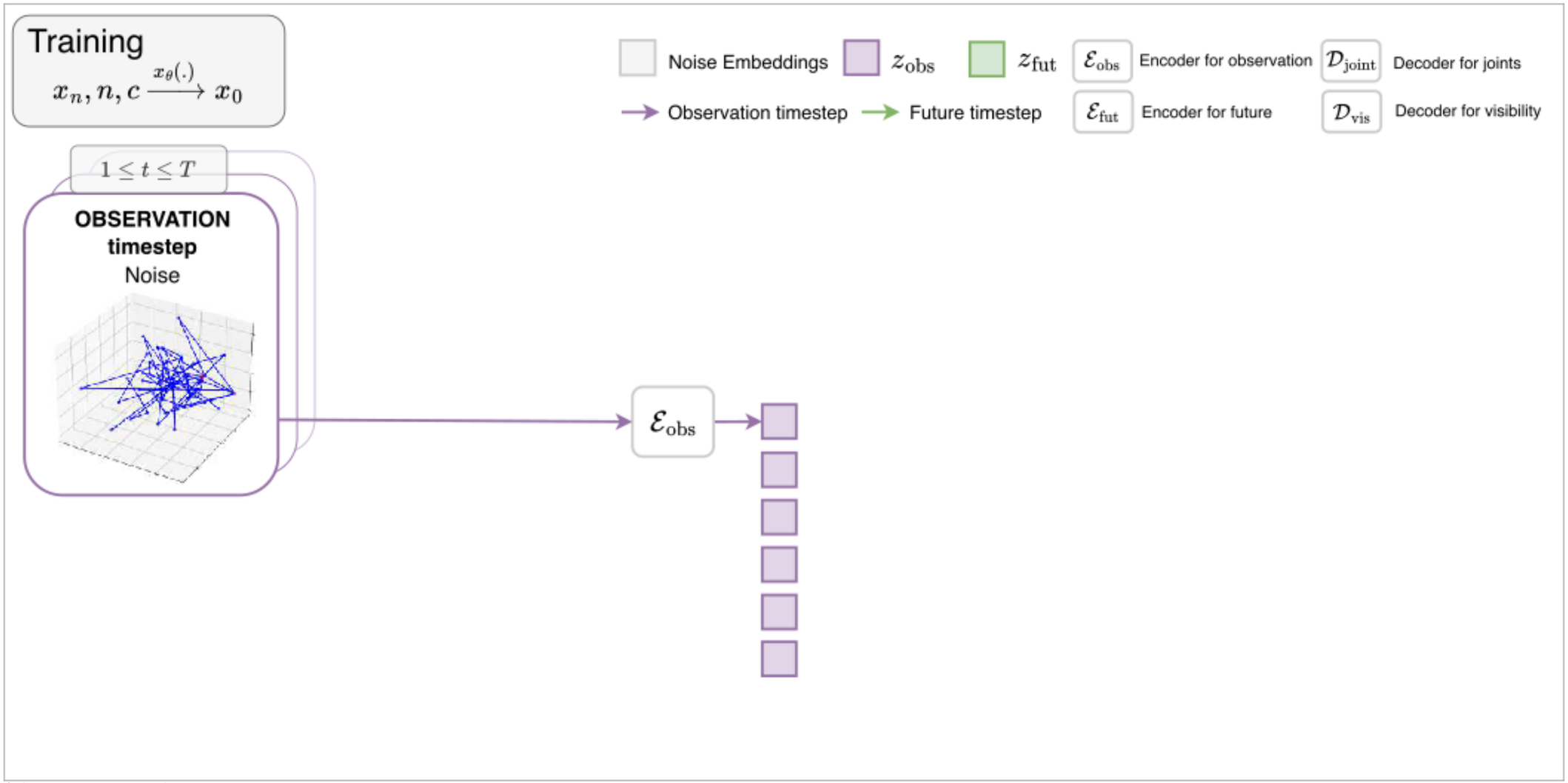
with: Masashi Hatano
Zhifan Zhu
Hideo Saito



Hatano et al (2025). The Invisible EgoHand: 3D Hand Forecasting through EgoBody Pose Estimation. ArXiv

The Invisible EgoHand

with: Masashi Hatano
Zhifan Zhu
Hideo Saito



Hatano et al (2025). The Invisible EgoHand: 3D Hand Forecasting through EgoBody Pose Estimation. ArXiv

The Invisible EgoHand

with: Masashi Hatano
Zhifan Zhu
Hideo Saito

Method	Hand Trajectory Forecasting		Hand Pose Forecasting	
	All		All	
	ADE	FDE	MPJPE	MPJPE-F
Static	0.335	0.405	0.166	0.179
CVM [61]	0.346	0.467	0.166	0.183
EgoEgoForecast	0.295	0.352	0.166	0.177
USST [3]	0.562	0.581	-	-
Ours	**0.261**	**0.324**	**0.115**	**0.143**

The Invisible EgoHand

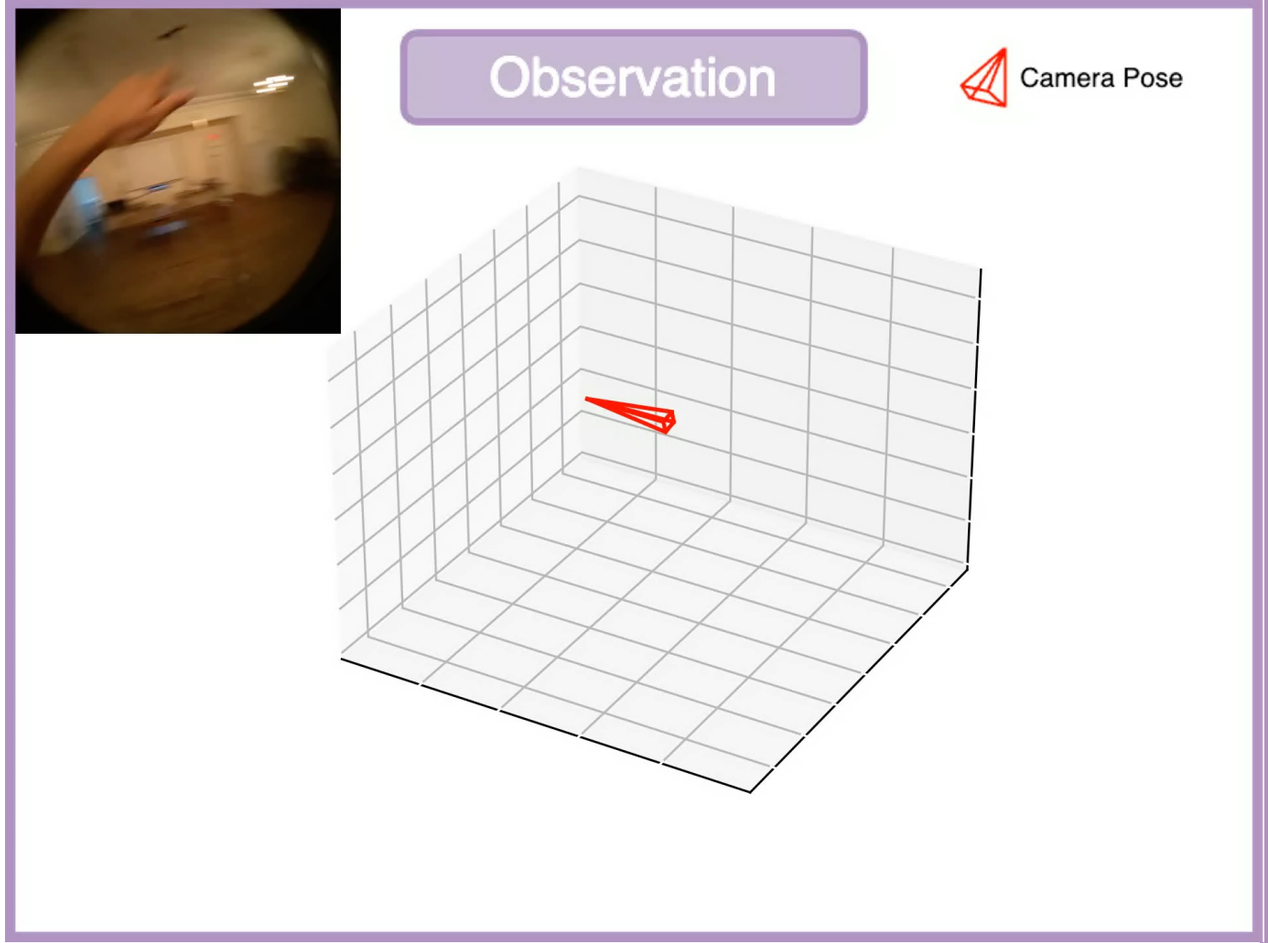
with: Masashi Hatano
Zhifan Zhu
Hideo Saito

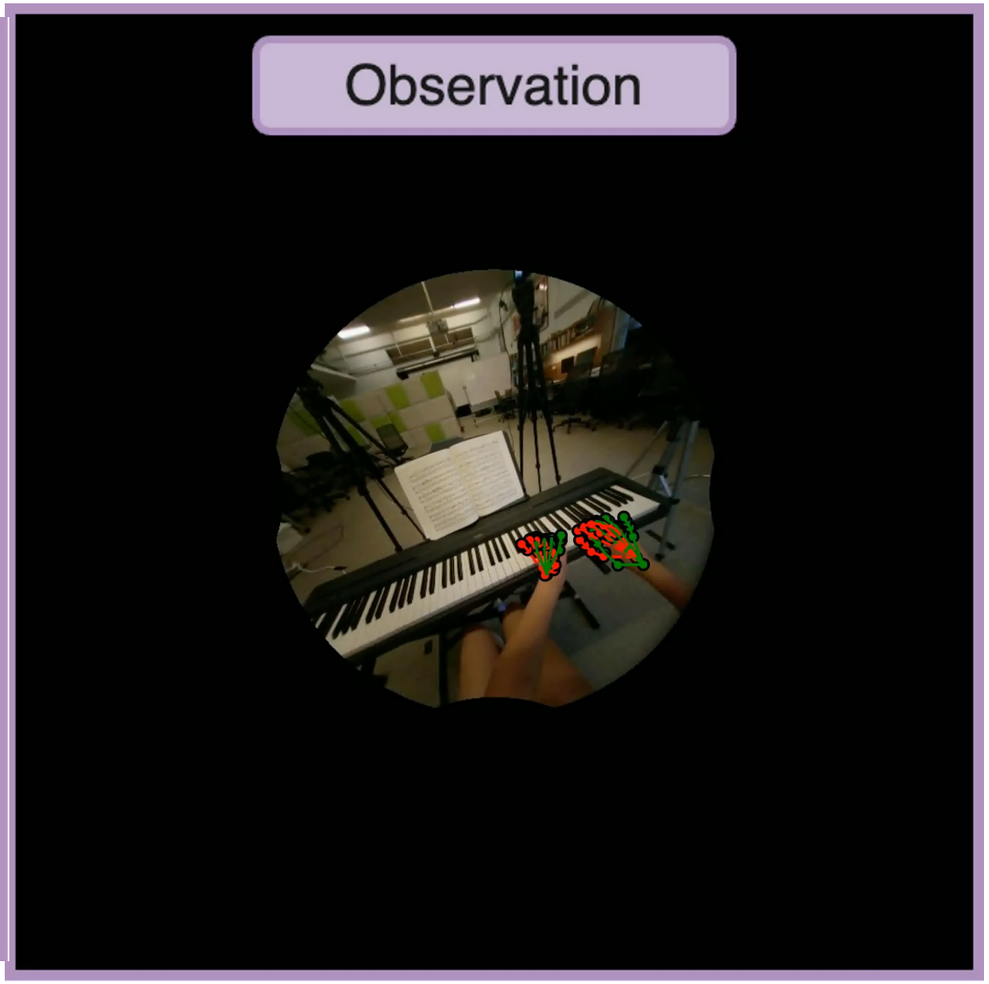
Method	Hand Trajectory Forecasting			Hand Pose Forecasting		
	In-view	Out-of-view	All	In-view	Out-of-view	All
EgoEgoForecast	0.171	0.385	0.295	0.162	0.299	0.166
Ours w/o. 2D joint	0.151	0.377	0.282	0.139	0.269	0.142
Ours w/o. image	**0.116**	0.367	**0.261**	0.117	**0.234**	0.120
Ours w/o. $\mathcal{L}_{\text{reproj}}$	0.132	0.368	0.269	0.125	0.250	0.128
Ours w/o. $\mathcal{L}_{\text{vis}}$	0.127	0.377	0.272	0.121	0.240	0.124
Ours w/o. $\mathcal{L}_{\text{body}}$	0.129	0.385	0.277	0.120	0.258	0.123
Ours w/o. $\mathcal{L}_{\text{obs}}$	0.149	0.390	0.289	0.139	0.250	0.142
Ours	**0.116**	**0.366**	**0.261**	**0.112**	0.240	**0.115**

- Without visible 2D joints, significant performance drops can be seen
- 2D reprojection loss serves as effective regularization
- Visibility loss & Body joints loss contribute for out-of-view scenario

The Invisible EgoHand

with: Masashi Hatano
Zhifan Zhu
Hideo Saito





Hatano et al (2025). The Invisible EgoHand: 3D Hand Forecasting through EgoBody Pose Estimation. ArXiv

Video Understanding Out of the Frame

From First-Point View to Second- and Third-

University of BRISTOL

Dima Damen
PAISS 2025

Ego-Exo4D

with: Kristen Grauman
+102 authors

University of BRISTOL

K Grauman et al (2024). Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives.. *CVPR*

Ego-Exo4D

with: Kristen Grauman
+102 authors

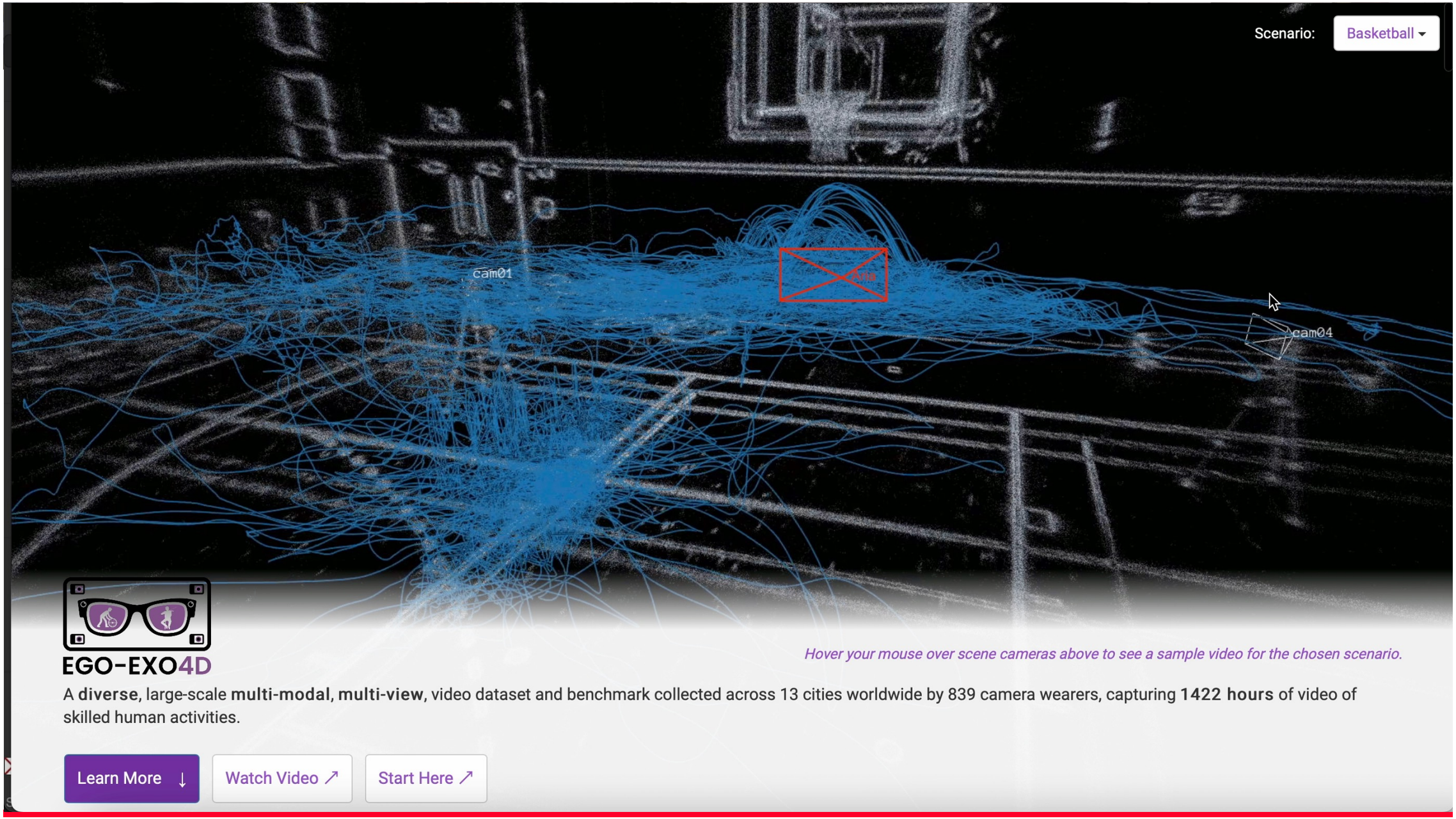
Ego-Exo Relation

K Grauman et al (2024). Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives.. *CVPR*

Ego-Exo4D

with: Kristen Grauman
+102 authors

K Grauman et al (2024). Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives.. *CVPR*



EGO-EXO4D

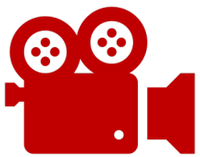
Hover your mouse over scene cameras above to see a sample video for the chosen scenario.

A **diverse**, large-scale **multi-modal**, **multi-view**, video dataset and benchmark collected across 13 cities worldwide by 839 camera wearers, capturing **1422 hours** of video of skilled human activities.

Learn More ↓ | Watch Video ↗ | Start Here ↗

In today's tutorial

Motivation and Datasets in Egocentric Video Understanding



Video Understanding: Data and Tasks

Videos are Multimodal

Connected Videos of One's Life

Video Understanding Out of the Frame

Teaser: The Wizard of Oz & Genie 3

Outlook into the Future of Egocentric Vision

Conclusion

The Wizard of Oz at the Sphere

Sold out tickets – August 2025

https://behindthecurtain.withgoogle.com

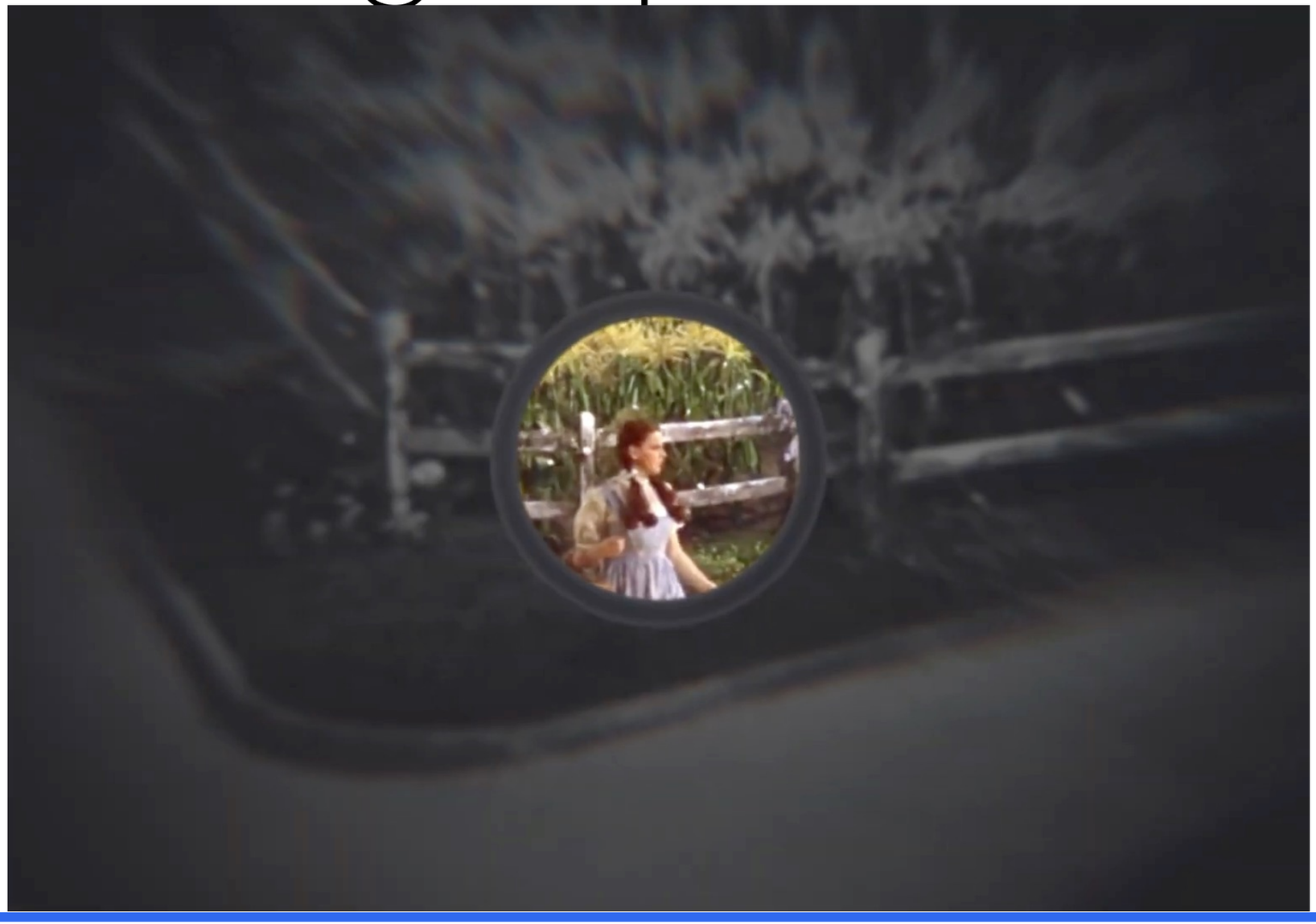
Dima Damen
PAISS 2025

The Wizard of Oz @ The Sphere

- The Movie (1939)
- Technocolour pioneer
- Iconic characters

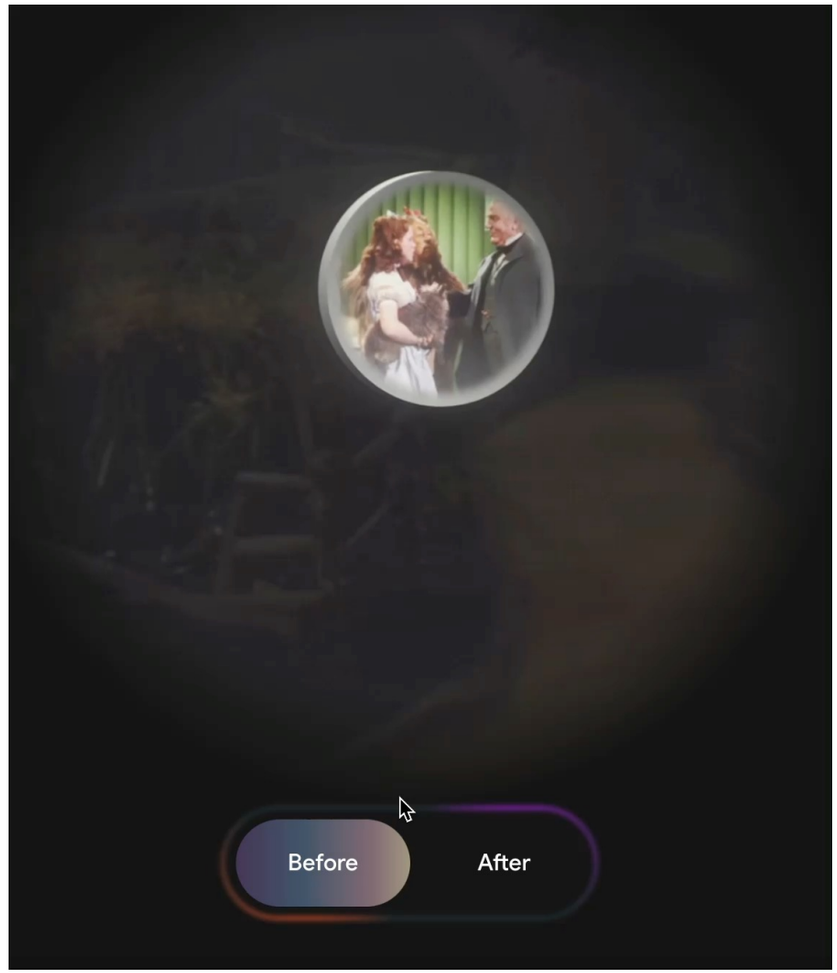
The Wizard of Oz @ The Sphere

The Wizard of Oz @ The Sphere



The Wizard of Oz @ The Sphere

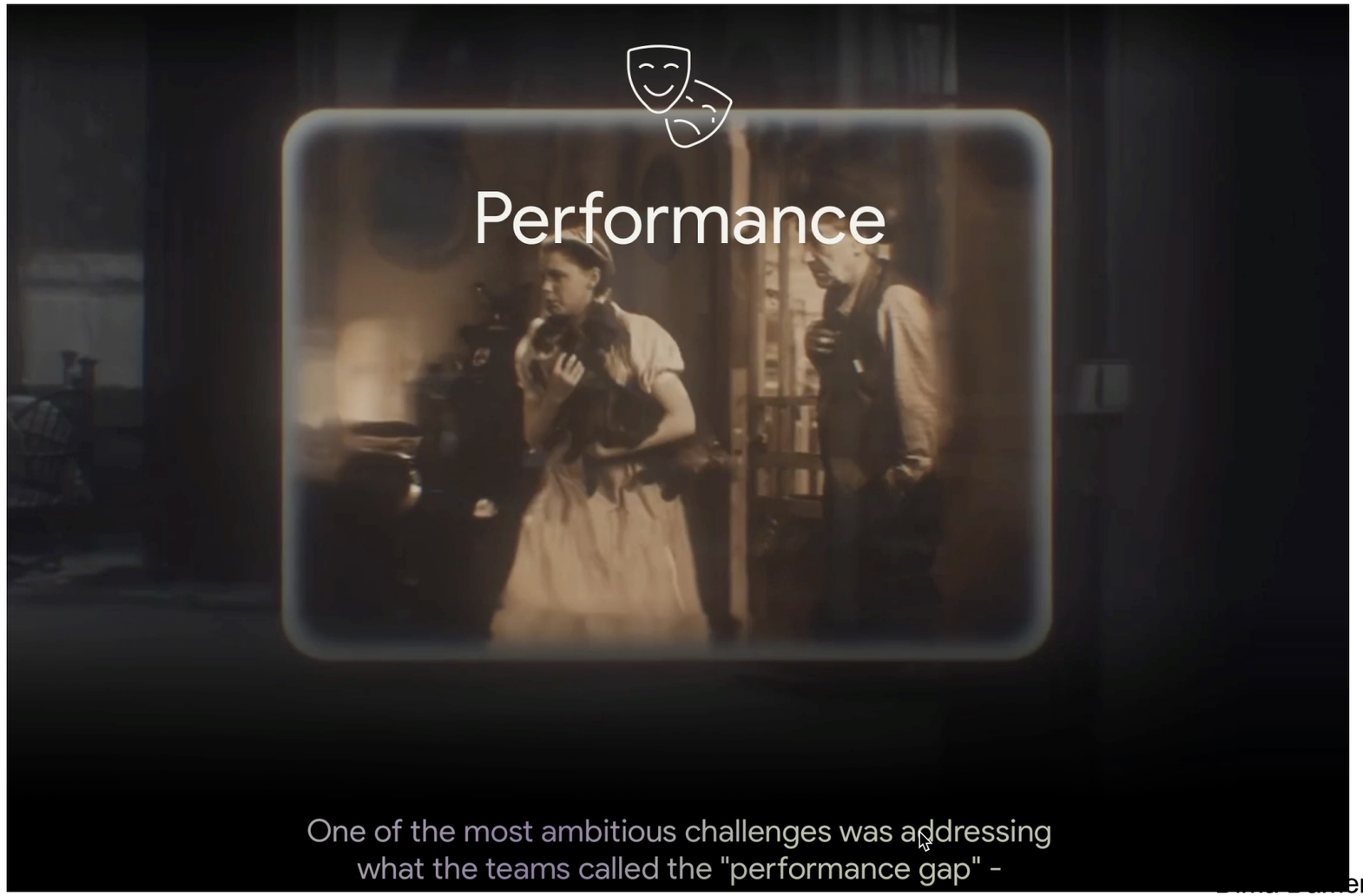
- Super-resolution,
- Outpainting…



https://behindthecurtain.withgoogle.com

The Wizard of Oz @ The Sphere

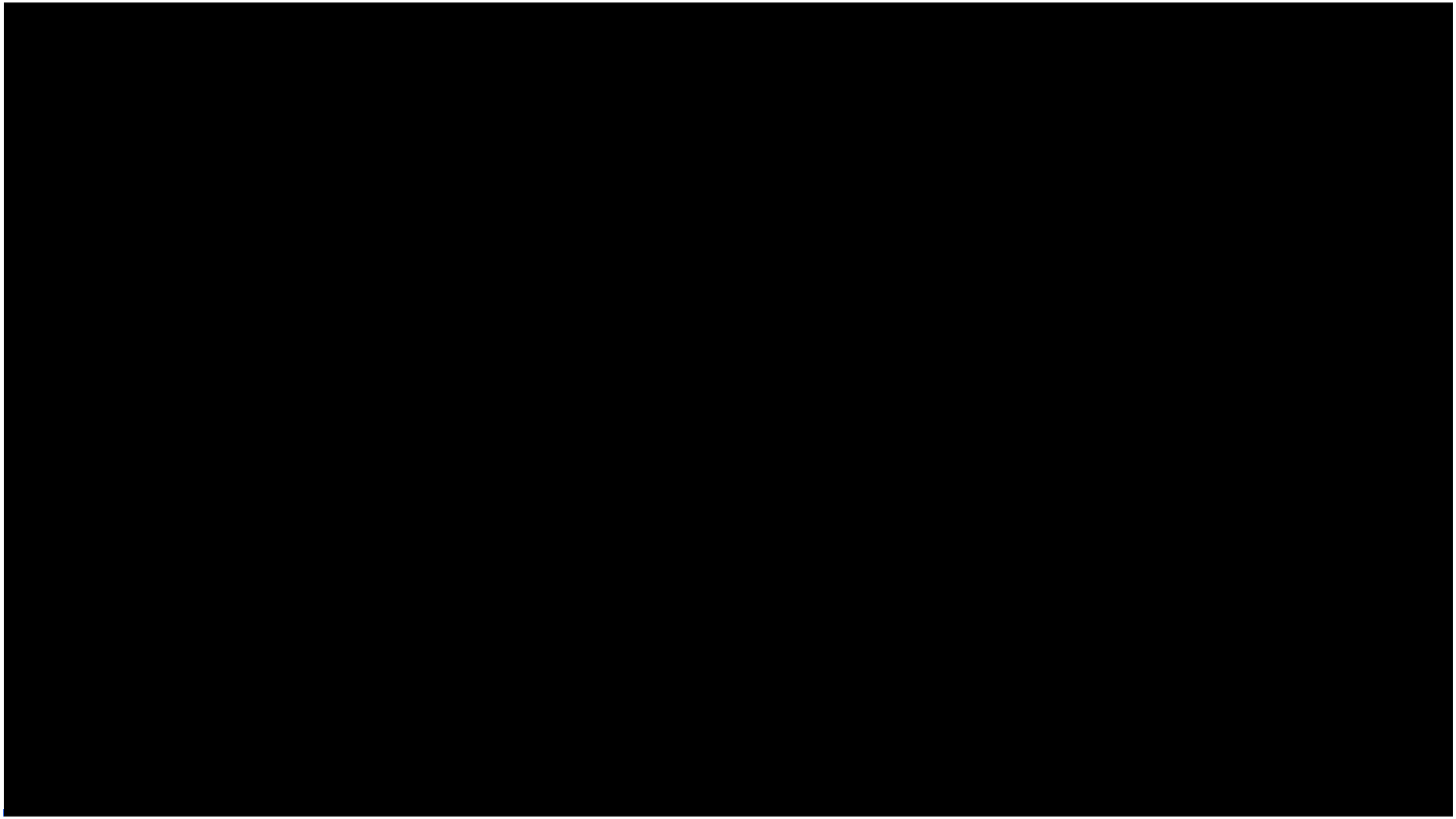
- Performance Interpolation,

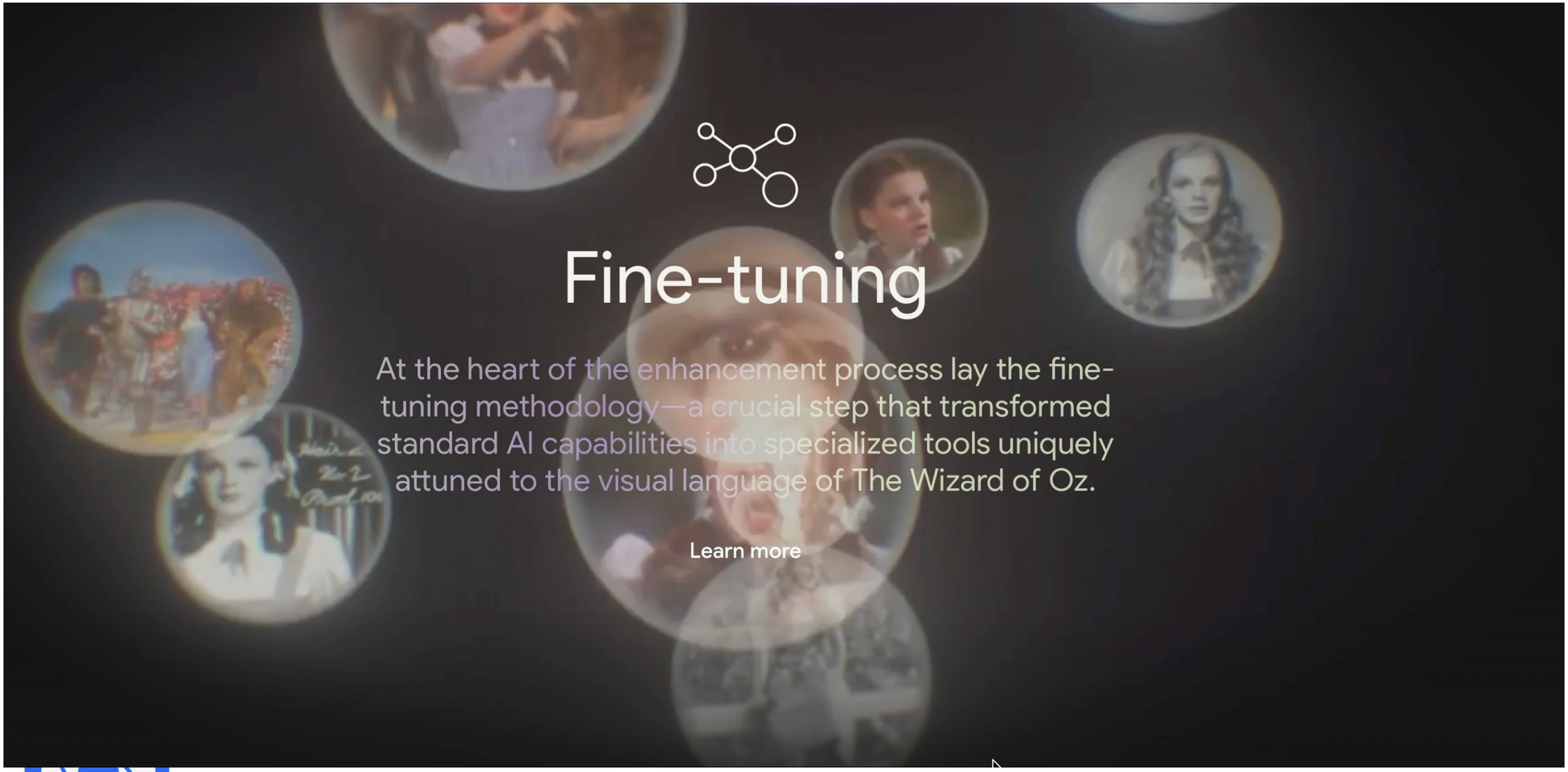


The Wizard of Oz @ The Sphere

- Auto-Director

https://behindthecurtain.withgoogle.com



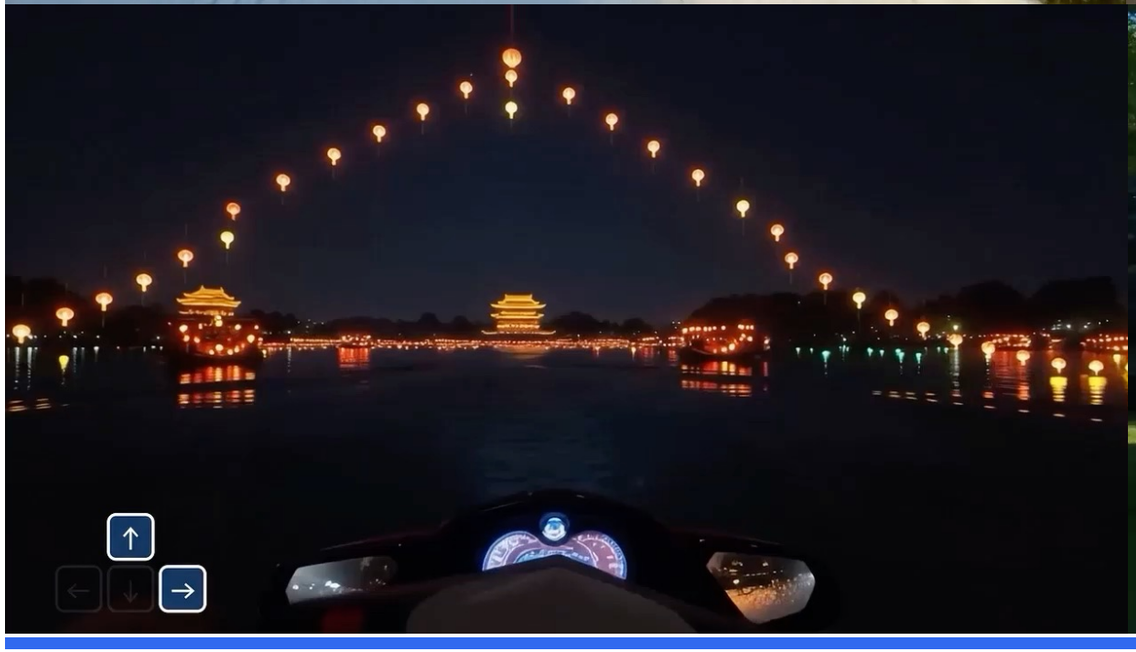


https://behindthecurtain.withgoogle.com

Genie 3

Released Aug 2025

https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/



Genie 3

- A new frontier of world models
- From a text prompt or a starting image/video
- Generate interactive worlds at 24fps and 720px
- Remains consistent for several minutes
- Genie 3’s consistency is an emergent capability

In today's tutorial

Motivation and Datasets in Egocentric Video Understanding

Video Understanding: Data and Tasks

Videos are Multimodal

Connected Videos of One's Life

Video Understanding Out of the Frame

Teaser: The Wizard of Oz & Genie 3

Outlook into the Future of Egocentric Vision

Conclusion

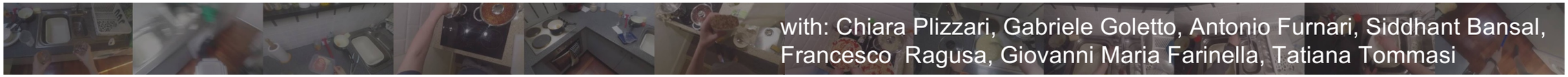
An Outlook into the Future of Egocentric Vision

Chiara Plizzari*, Gabriele Goletto*, Antonino Furnari*, Siddhant Bansal*, Francesco Ragusa*, Giovanni Maria Farinella†, Dima Damen†, Tatiana Tommasi†

Politecnico di Torino

University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA



with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi

Envisioning an Ambitious Future and **Analysing** the Current Status of Egocentric Vision

How did we do this?

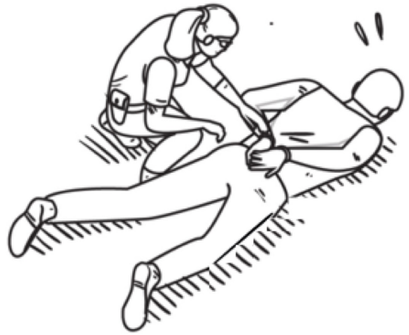
We imagined a device – *EgoAI* and envisioned its utility in multiple scenarios

EGO-Designer

EGO-Worker

EGO-Tourist

EGO-Home

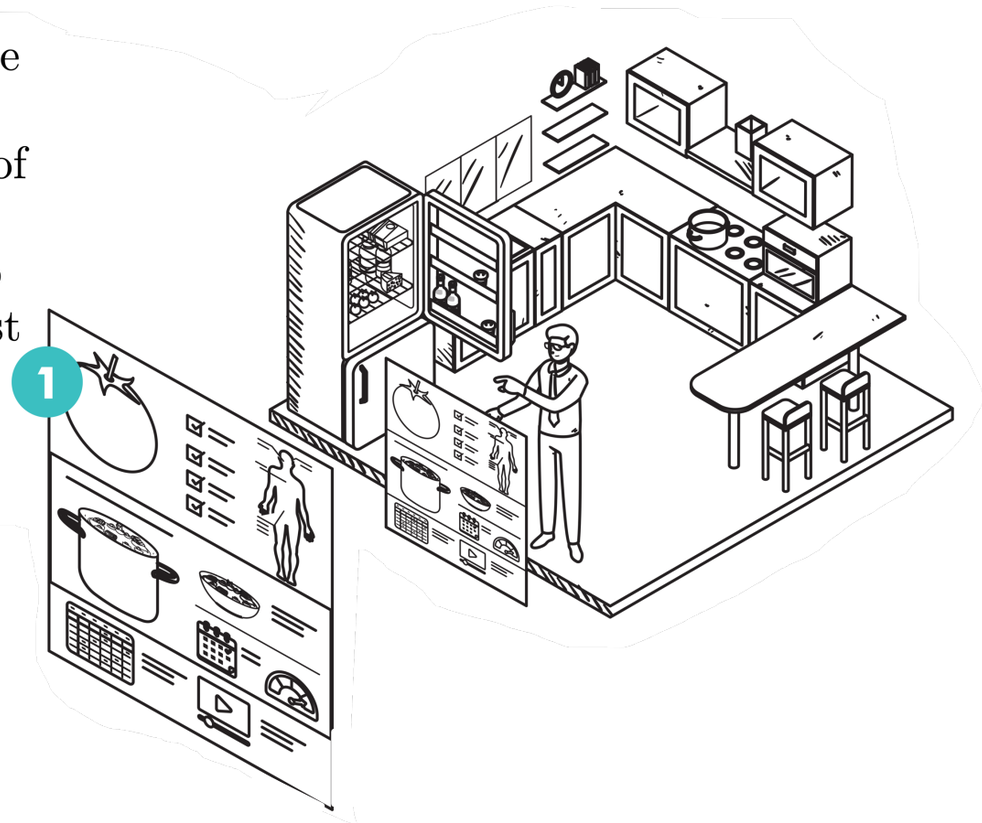


Ego-Police

EGO-Home

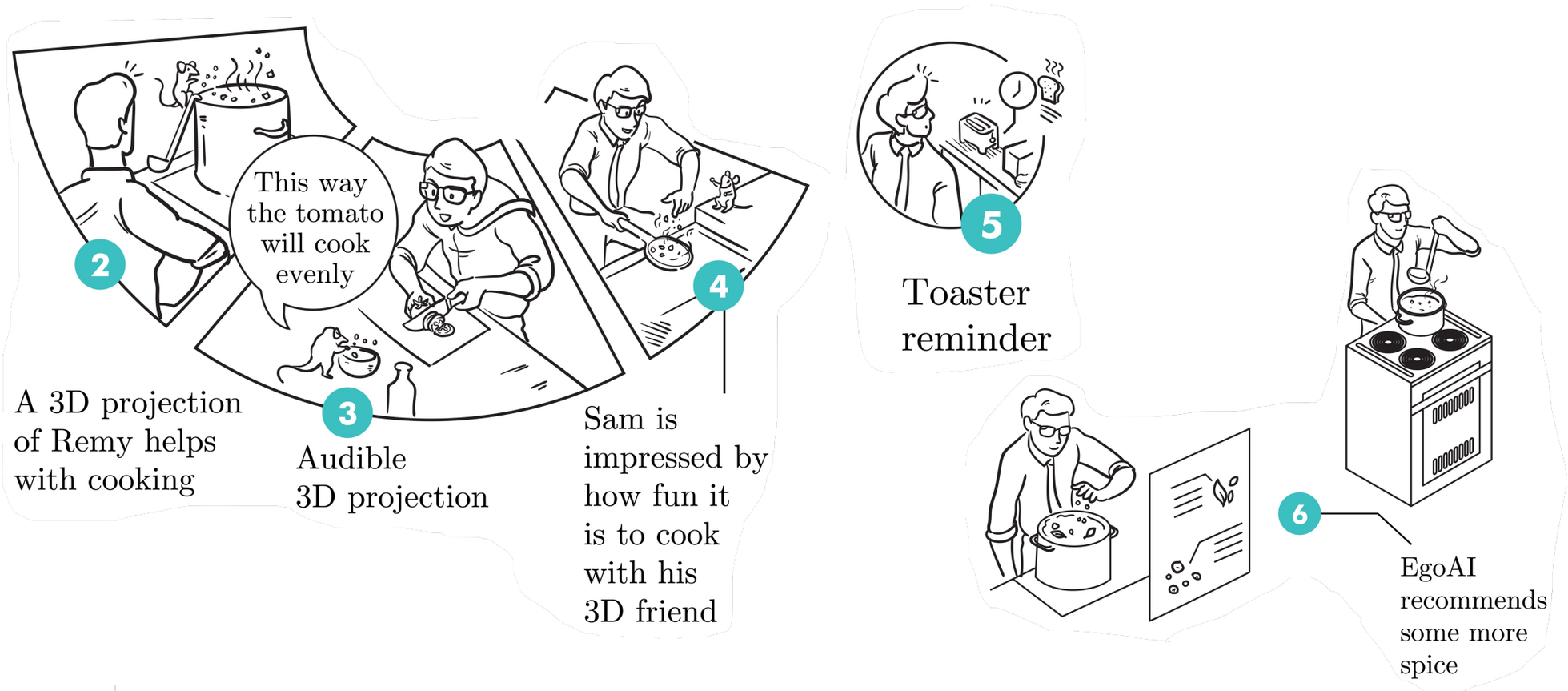
with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi

Sam is finally home after a long day. EgoAI kept track of Sam's food intake and a tomato soup sounds like the best complementary nutrition



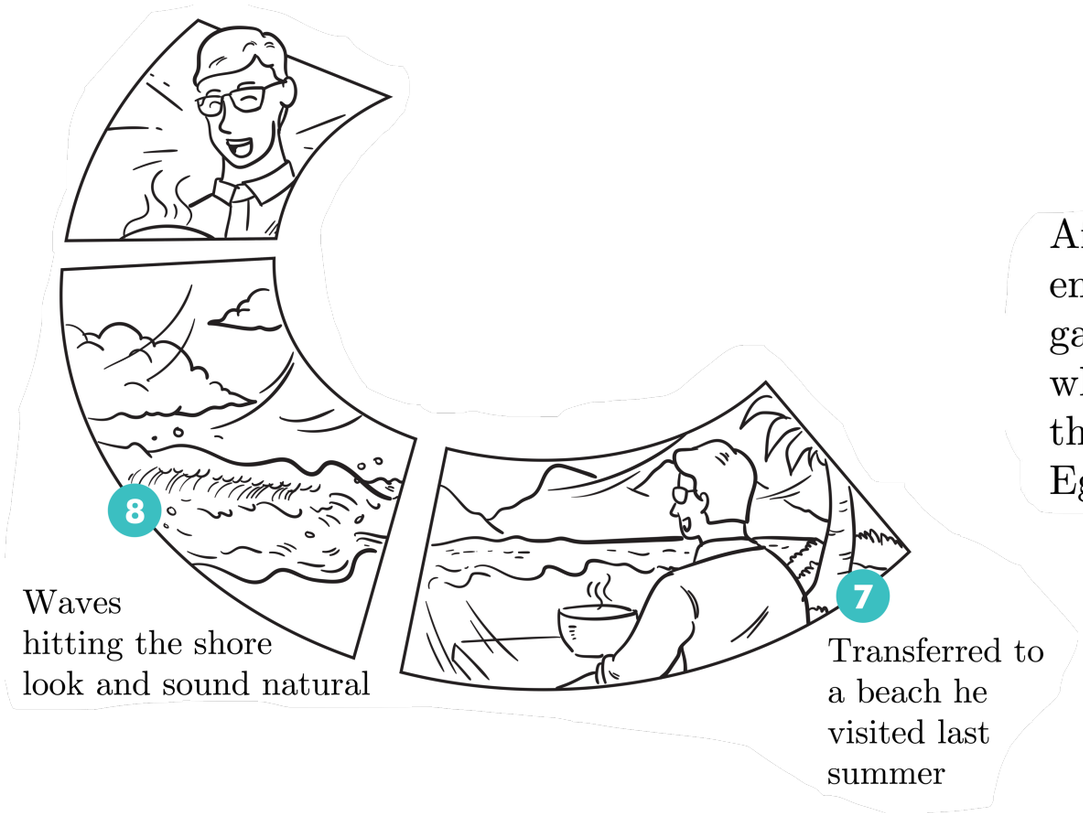
EGO-Home

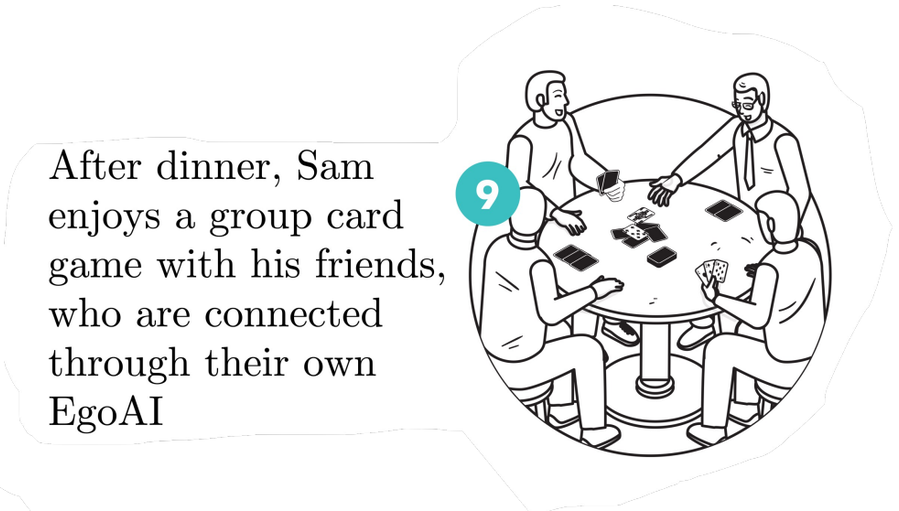
with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi



EGO-Home

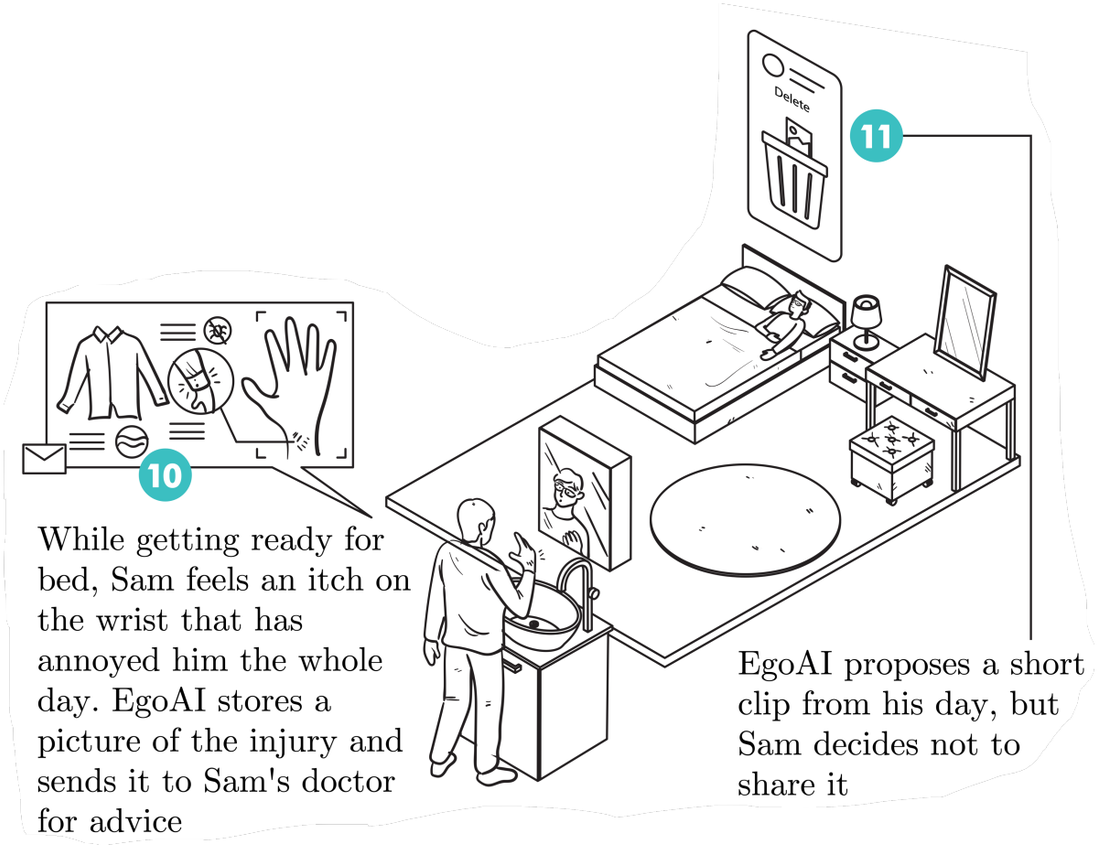
with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi





EGO-Home

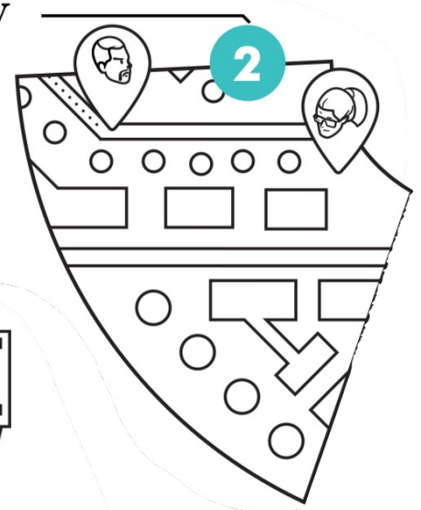
with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi

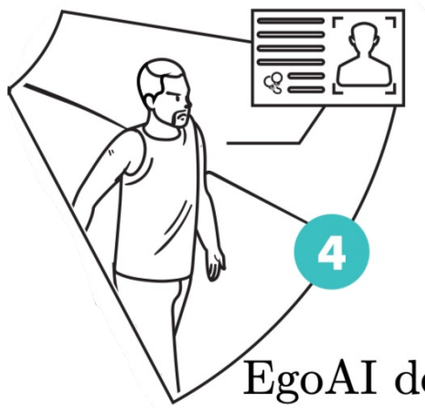


From Stories to Tasks

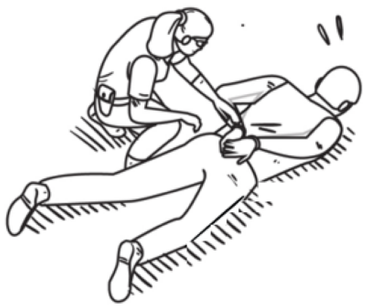
with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi

EgoAI helps Judy navigate through the shortest safe path to target places





EgoAI detected and re-identified the man before he passed Judy

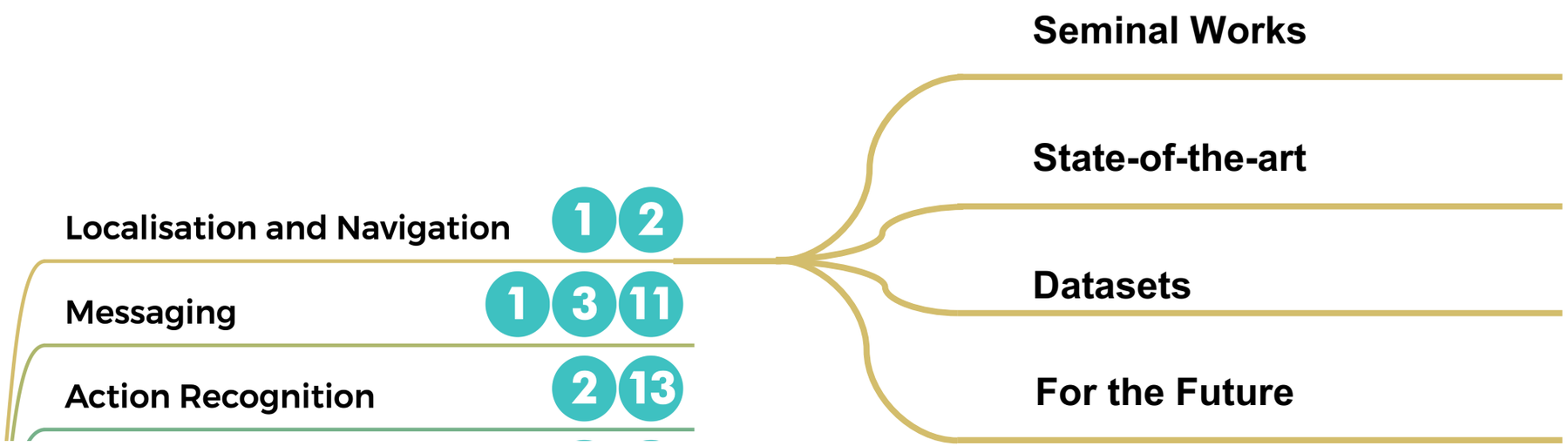


EGO-Police

- Localisation and Navigation 1 2
- Messaging 1 3 11
- Action Recognition 2 13
- Person Re-ID 2 4
- Object Detection and Retrieval 7
- Measuring System 8 9
- Decision Making 9
- 3D Scene Understanding 10
- Hand-Object Interaction 12
- Summarisation 13
- Privacy 14

The Survey Part

with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi



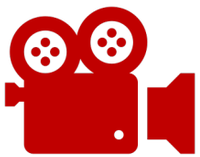
The Survey Part

with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi

- 12 tasks
- 46 pages (excluding references)
- 462 references

In today's tutorial

Motivation and Datasets in Egocentric Video Understanding



Video Understanding: Data and Tasks

Videos are Multimodal

Connected Videos of One's Life

Video Understanding Out of the Frame

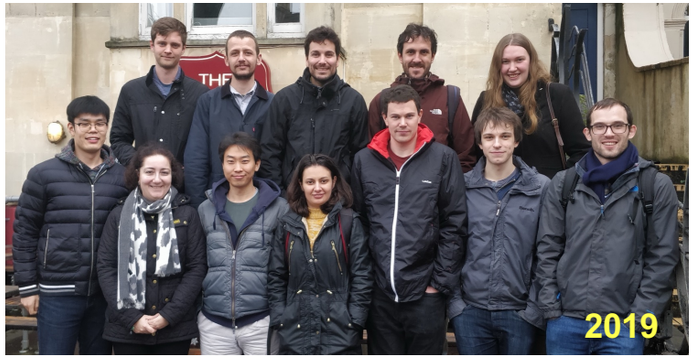
Teaser: The Wizard of Oz & Genie 3

Outlook into the Future of Egocentric Vision

Conclusion

The Team

grateful

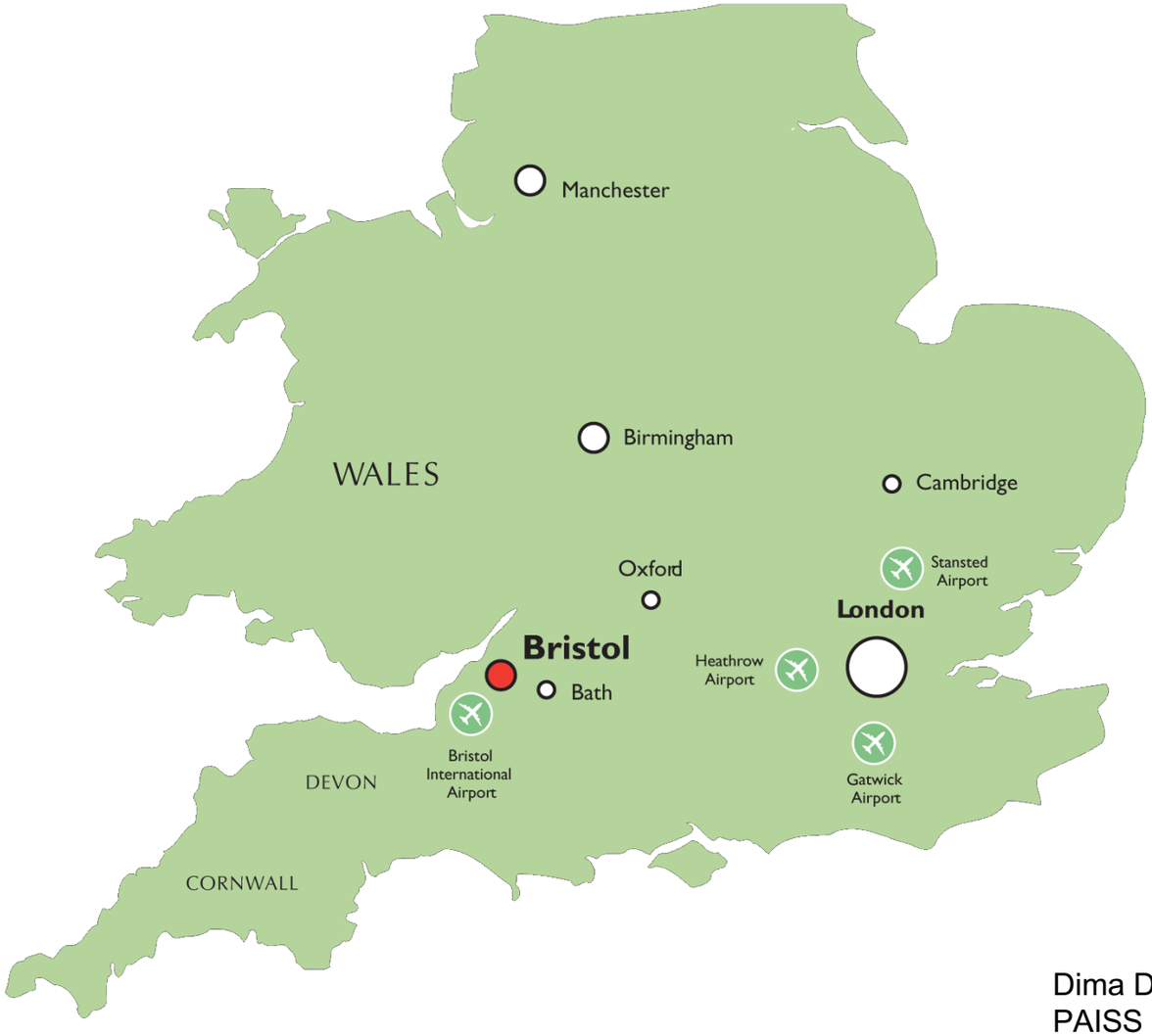


My research team…

grateful

BRISTOL

Visit us…



University of BRISTOL

Dima Damen
PAISS 2025

Thank you

For further info, datasets, code, publications…

http://dimadamen.github.io


@dimadamen


@dimadamen.bsky.social


http://www.linkedin.com/in/dimadamen

Q&A