Video Understanding

University of BRISTOL

Dima Damen
ICVSS2025

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

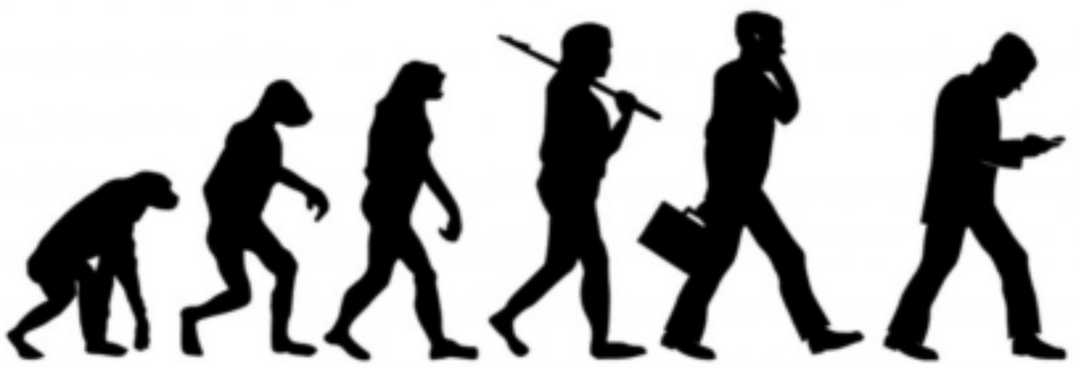
- Motivation and Datasets in Egocentric Video Understanding
- Video Understanding: Data and Tasks
- Videos are Multimodal
- Connected Videos of One's Life
- Video Understanding Out of the Frame
- Teaser: The Wizard of Oz at the Sphere
- Outlook into the Future of Egocentric Vision
- Conclusion

In today's tutorial

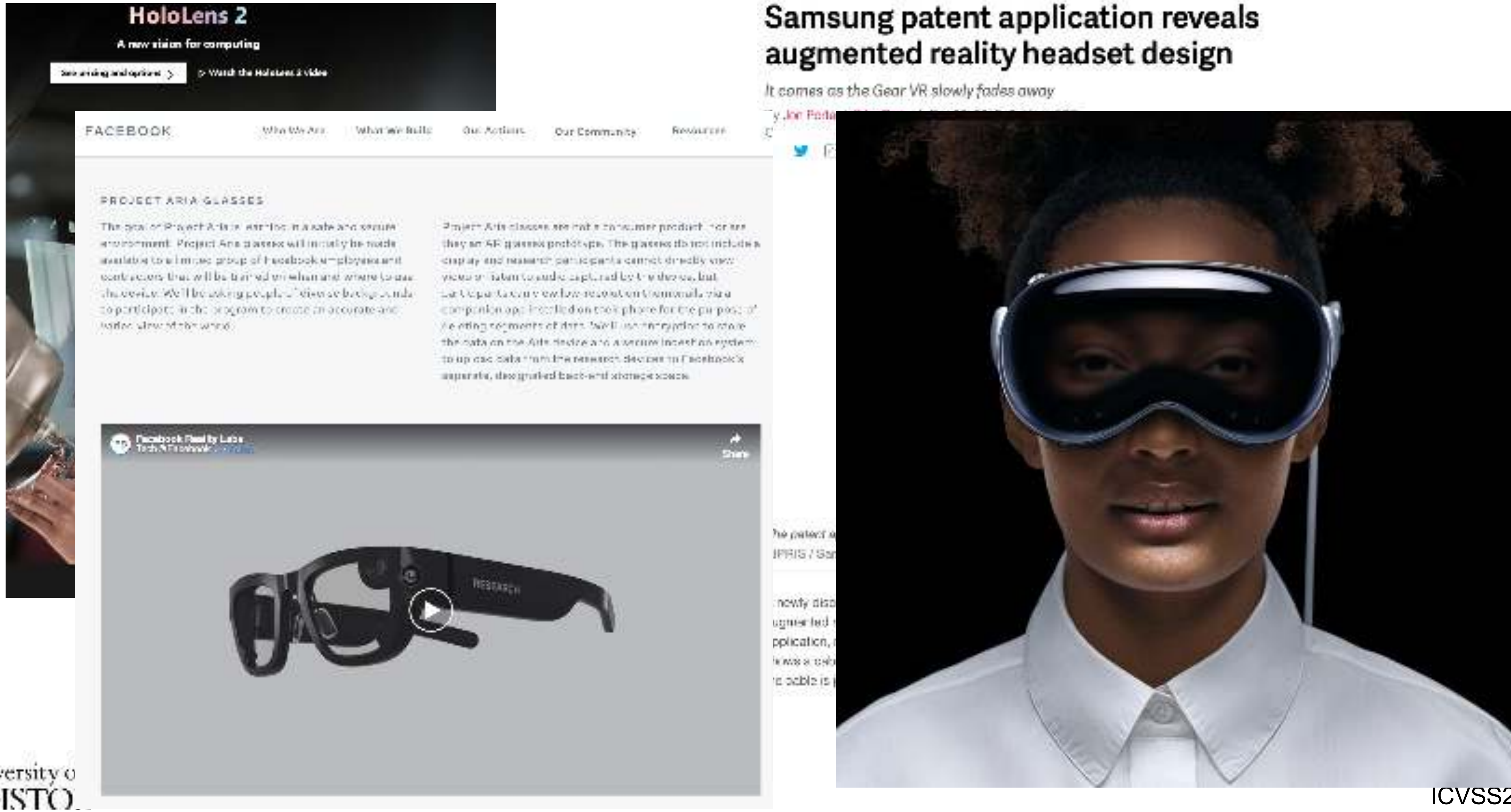
- Motivation and Datasets in Egocentric Video Understanding
- Video Understanding: Data and Tasks
- Videos are Multimodal
- Connected Videos of One's Life
- Video Understanding Out of the Frame
- Teaser: The Wizard of Oz at the Sphere
- Outlook into the Future of Egocentric Vision
- Conclusion

The present…

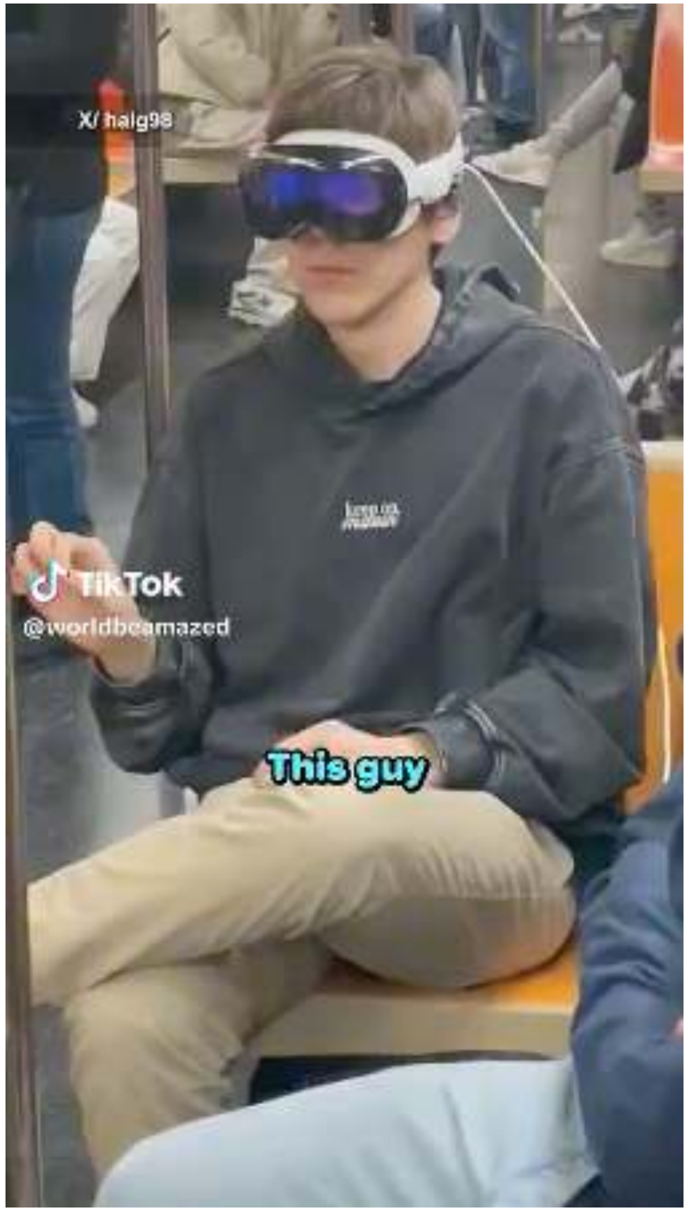
Photo *Illustration* by Pelle Cass



The future…


HoloLens 2
A new vision for computing
FACEBOOK
PROJECT ARIA GLASSES
Samsung patent application reveals augmented reality headset design
It comes as the Gear VR slowly fades away

The future is here…

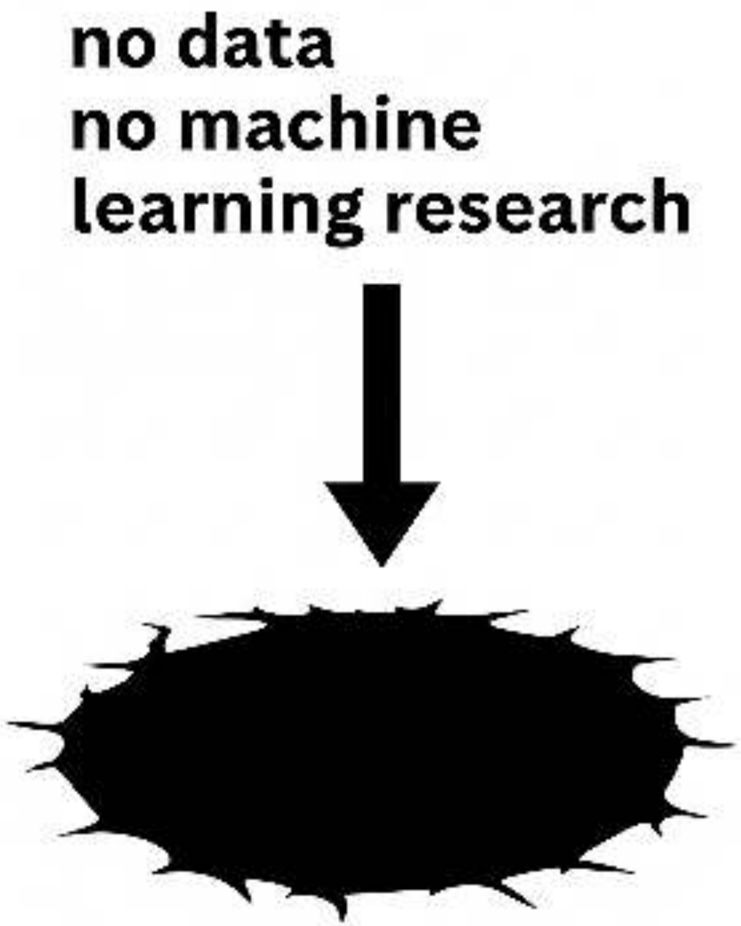


The future can be imagined…

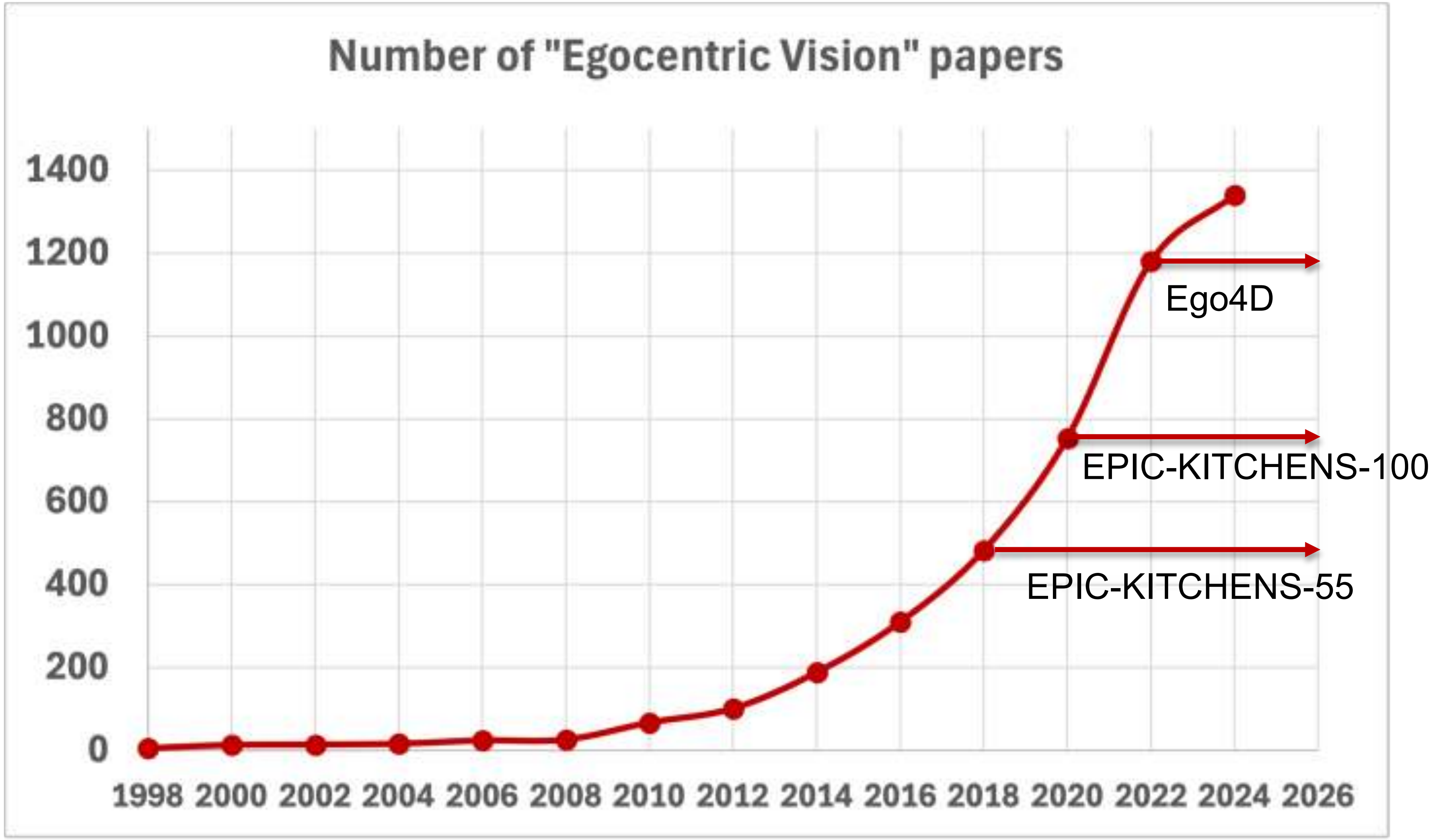
Egocentric Videos?

Egocentric Videos?

Machine Learning in Practice



Is research catching up?

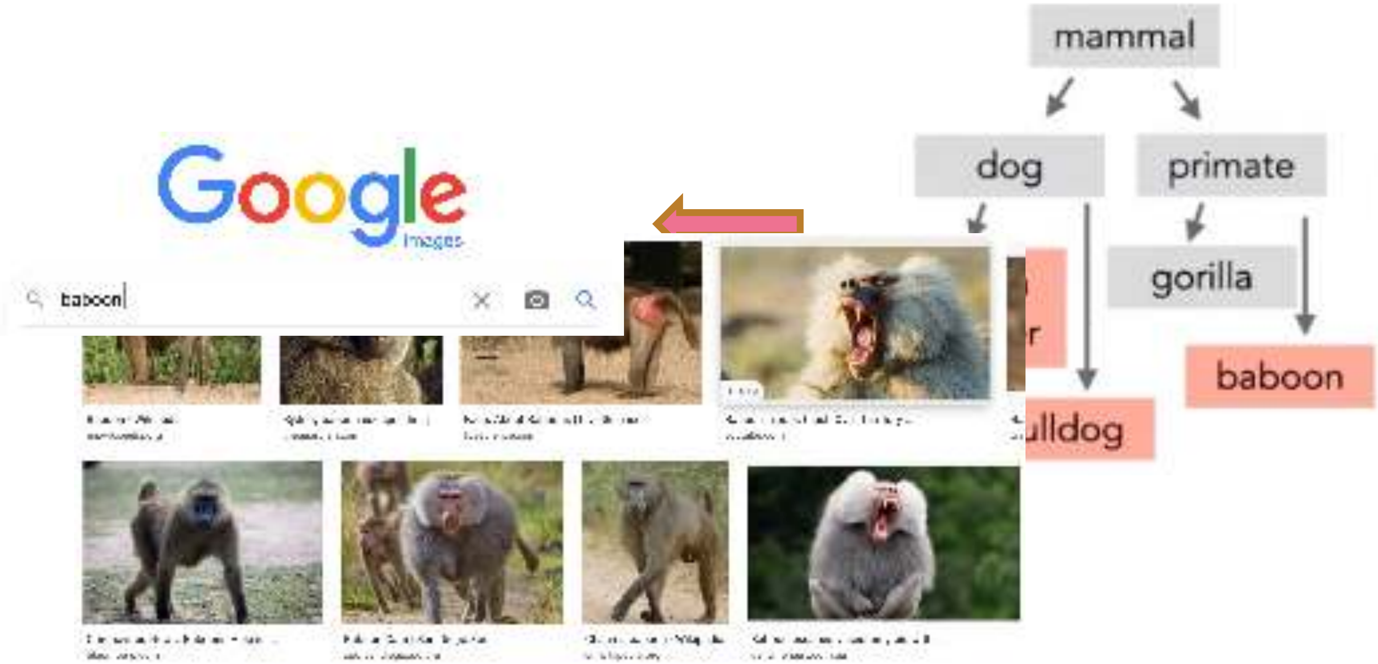


Statistics from Google Scholar – June 2025

Machine Learning in Practice

Real World Task….

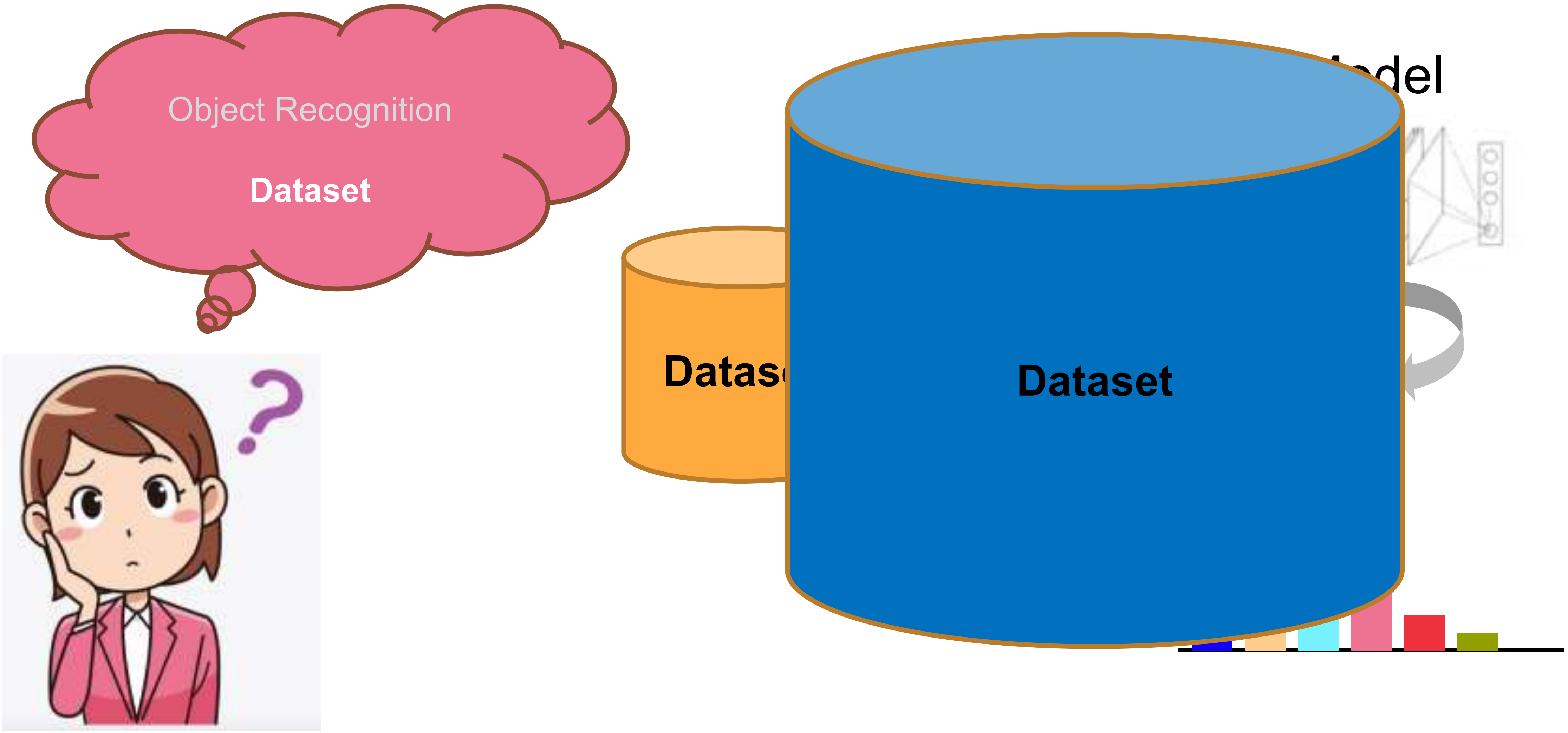
Object Recognition



Dataset

Ideas from: Aleksander Madry, CVPR 2020 workshop on Adversarial machine learning

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

University of BRISTOL

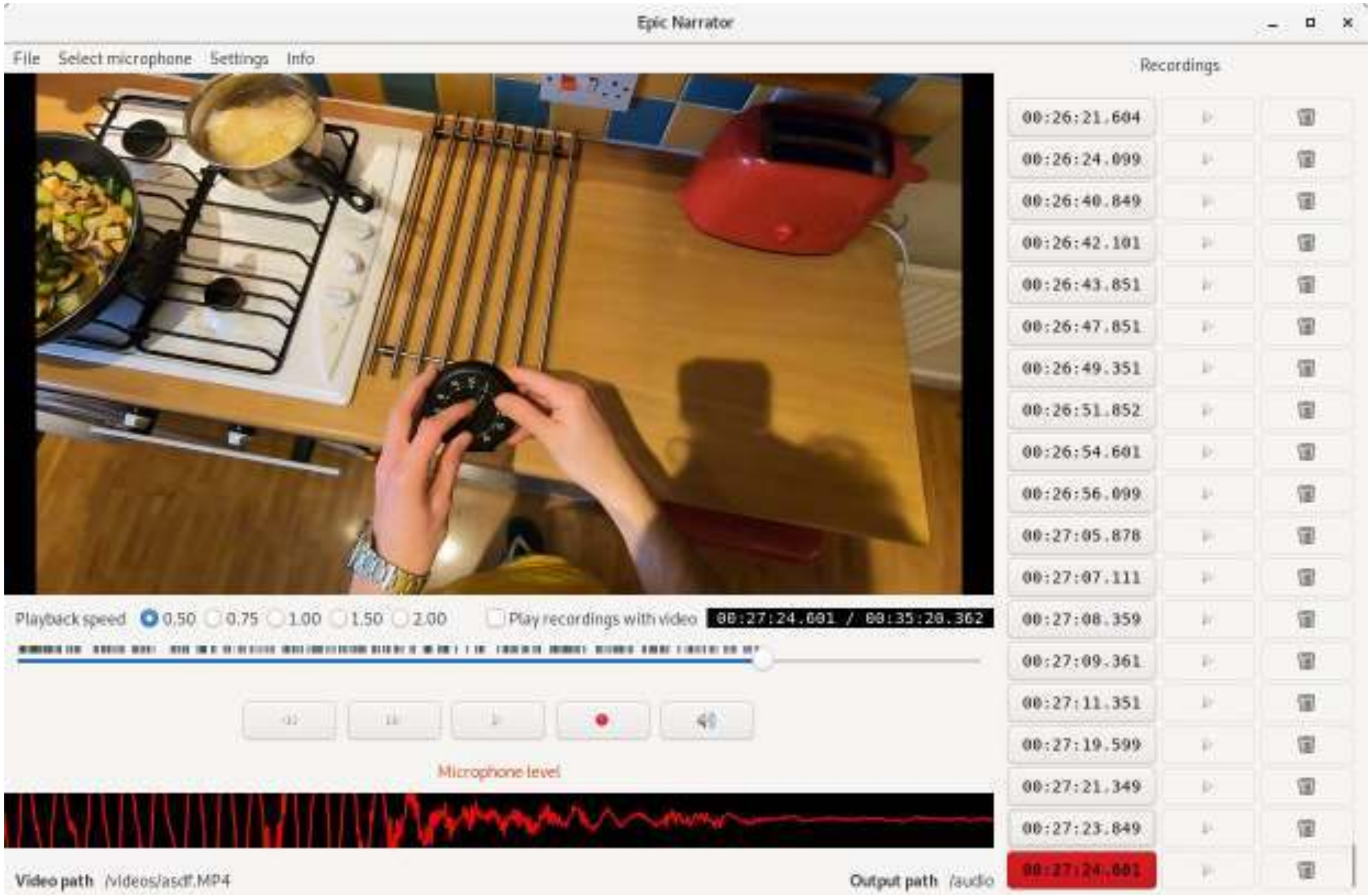
Dima Damen
ICVSS2025

EPIC-KITCHENS

EPIC-KITCHENS


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

EPIC-KITCHENS



EPIC-KITCHENS and Ego4D

Store

Lay-down **PUT**

Place Rest-on

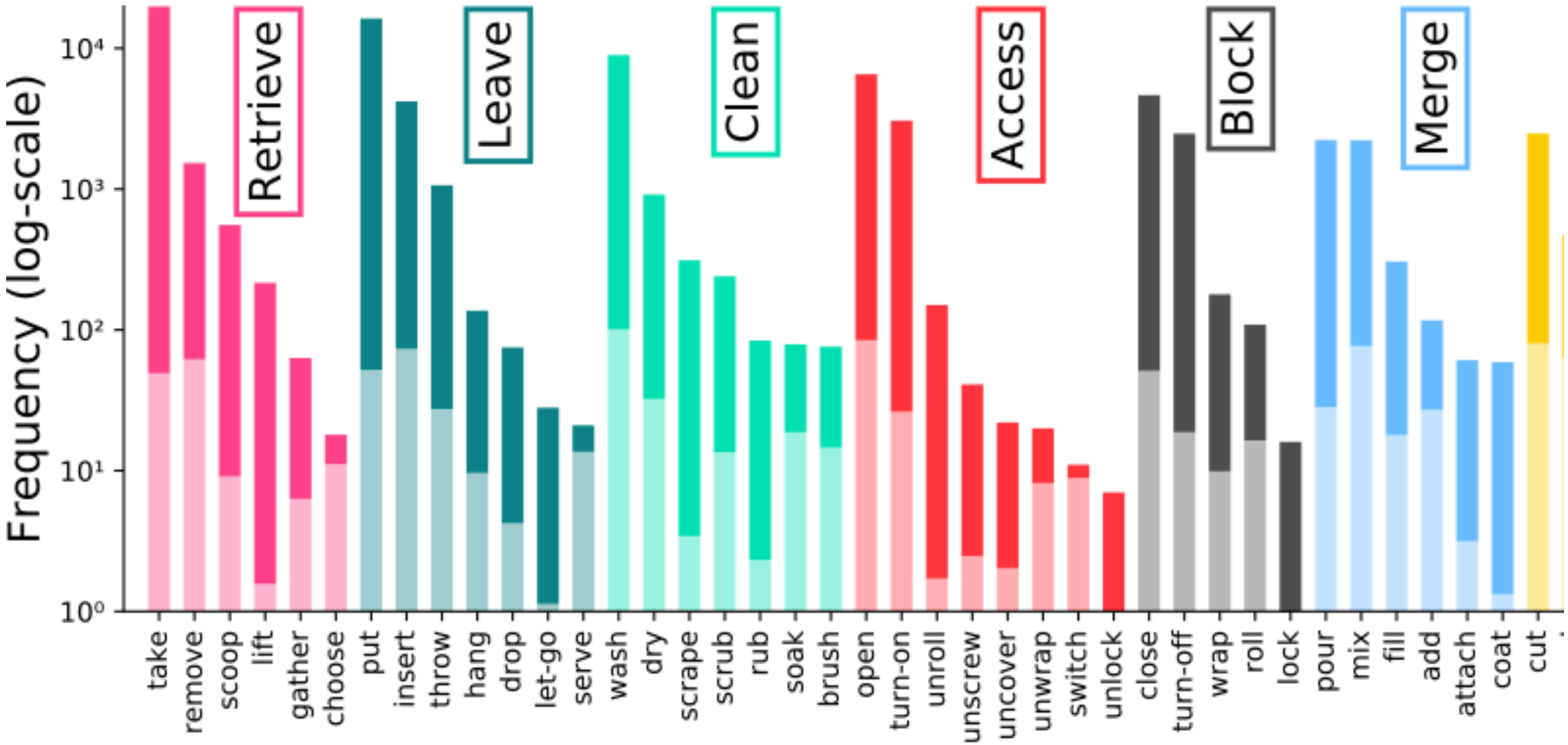
stove burner

hob

stovetop gas

open vocab	lay-down	stovetop
closed vocab	put	hob
category	leave	appliance

EPIC-KITCHENS-100 Statistics



Ego4D

with: Kristen Grauman
+83 authors

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

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 at the Sphere

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

The history of VIDEO understanding

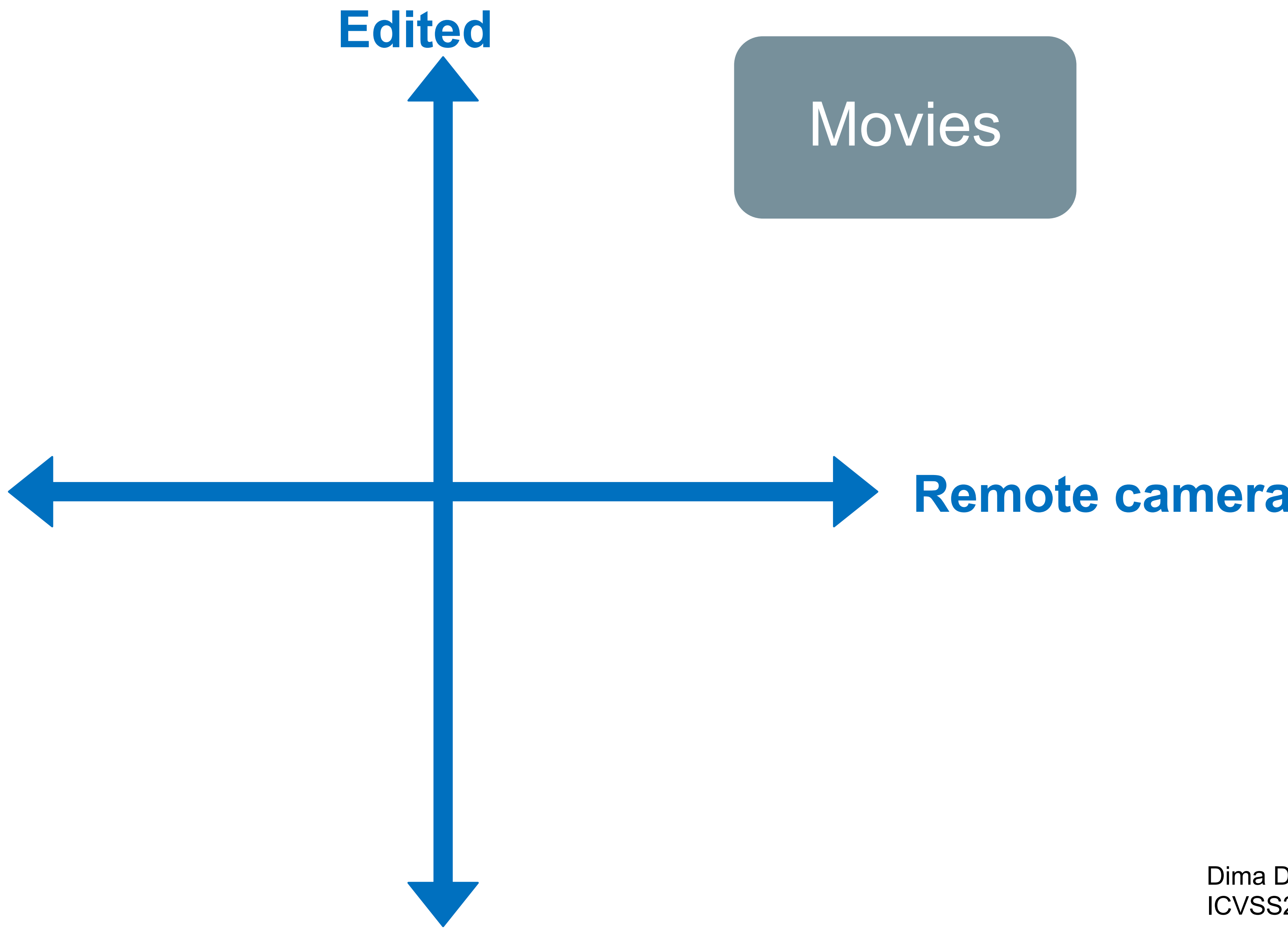


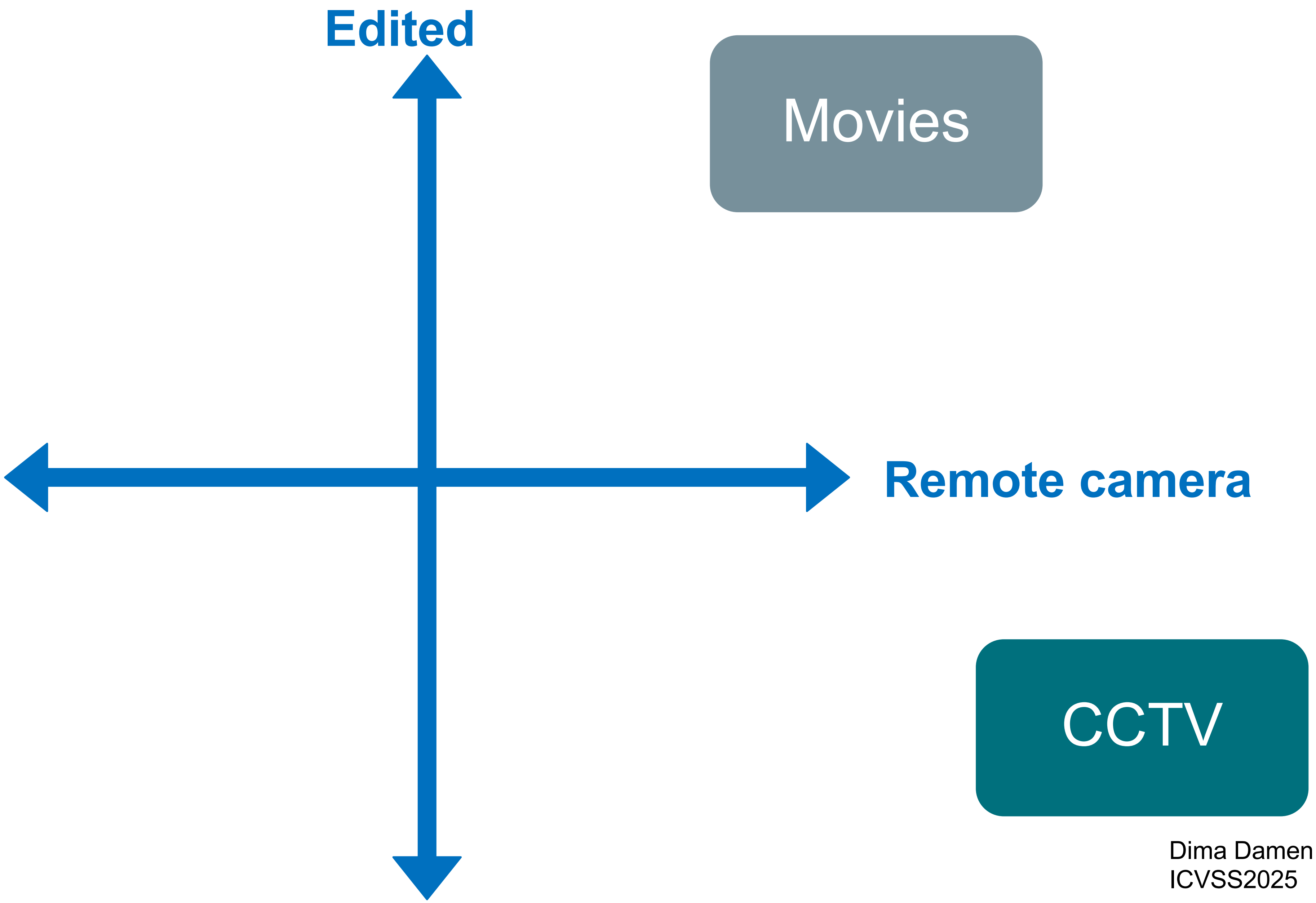
Figure 1. Examples of two action classes (drinking and smoking) from the movie "Coffee and Cigarettes". Note the high within-

Laptev and Perez (2007)

The history of VIDEO understanding

Laptev et al (2008). Learning realistic human actions from movies. CVPR

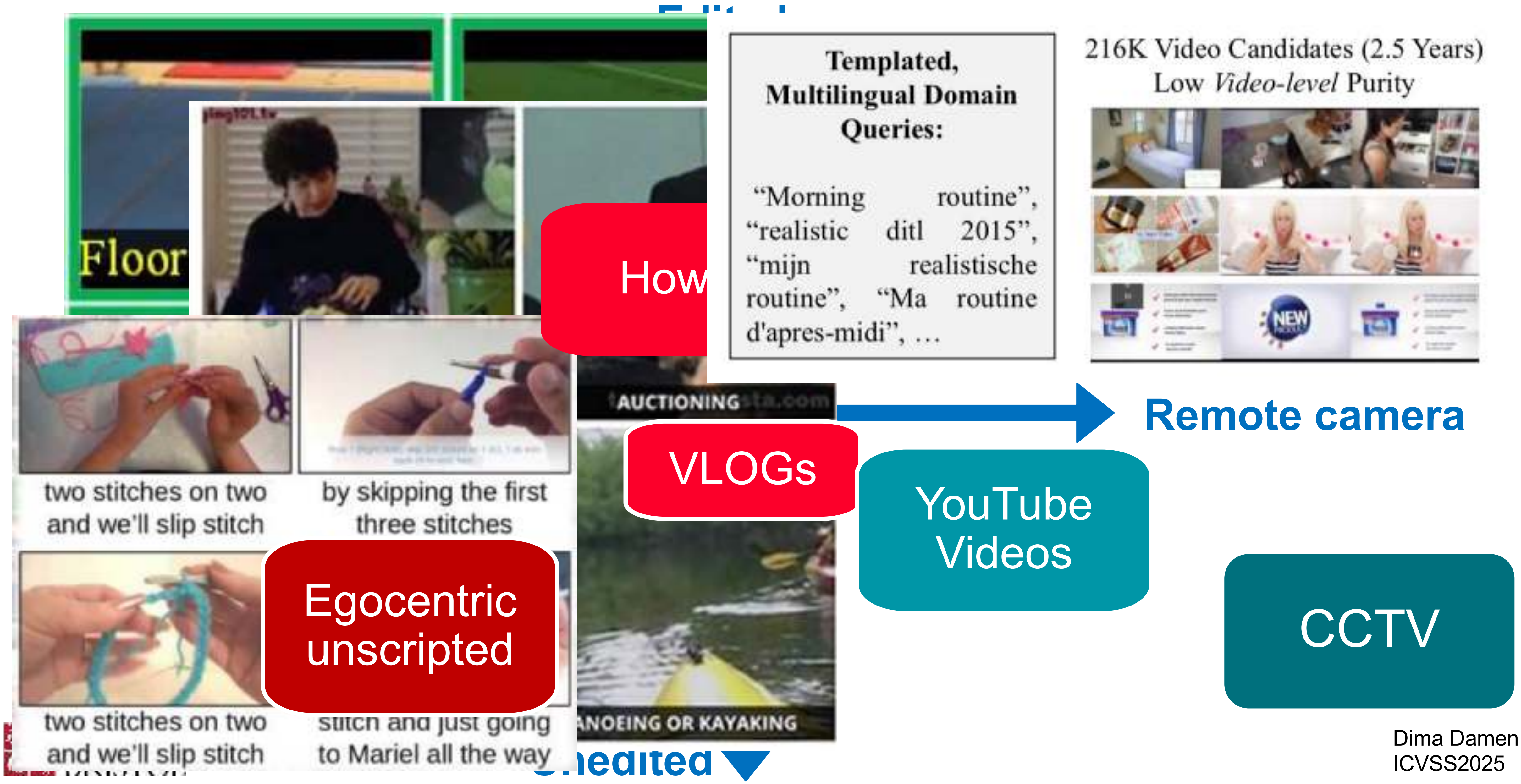
The history of VIDEO understanding



The history of VIDEO understanding

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

The history of VIDEO understanding



The history of VIDEO understanding

Movies

TV (sports)

HowTo

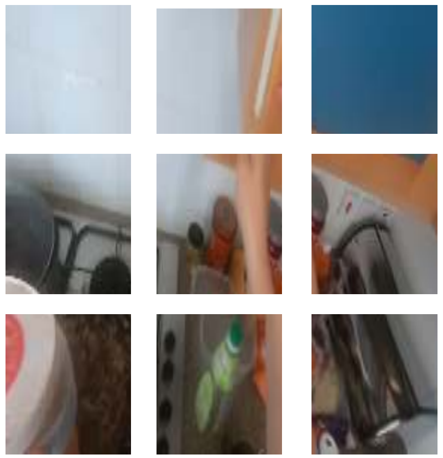
VLOGs

YouTube Videos

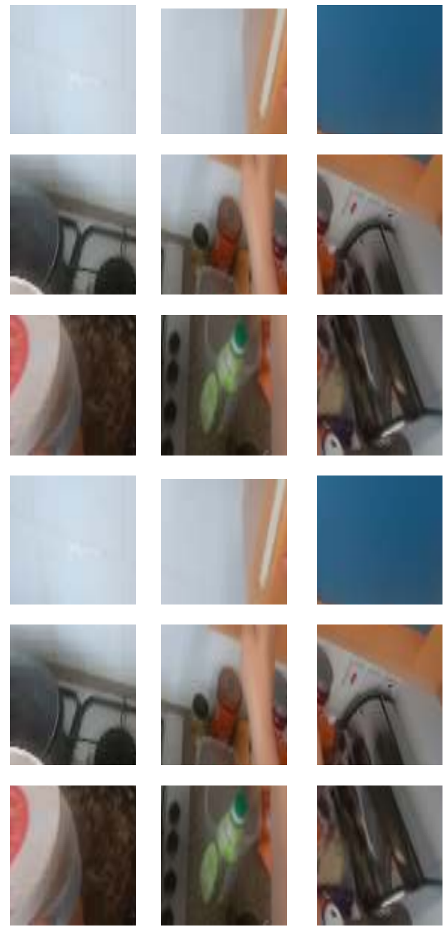
Egocentric unscripted

CCTV

How to Patch-ify a Video?



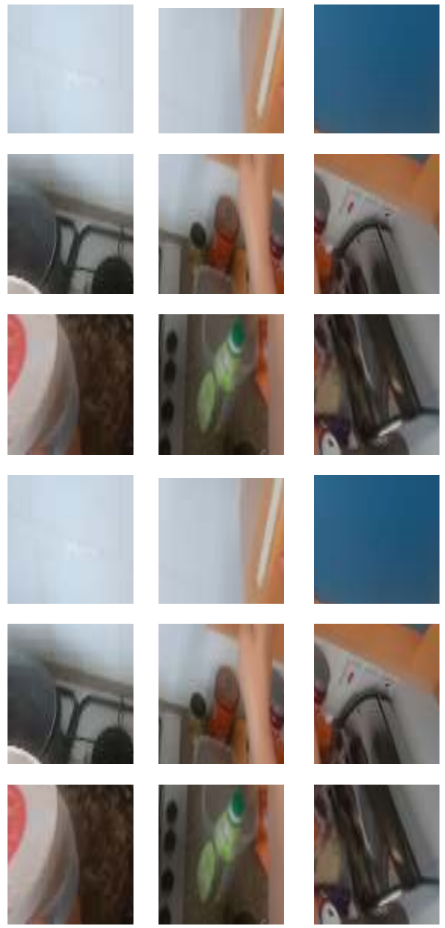
H x W x C

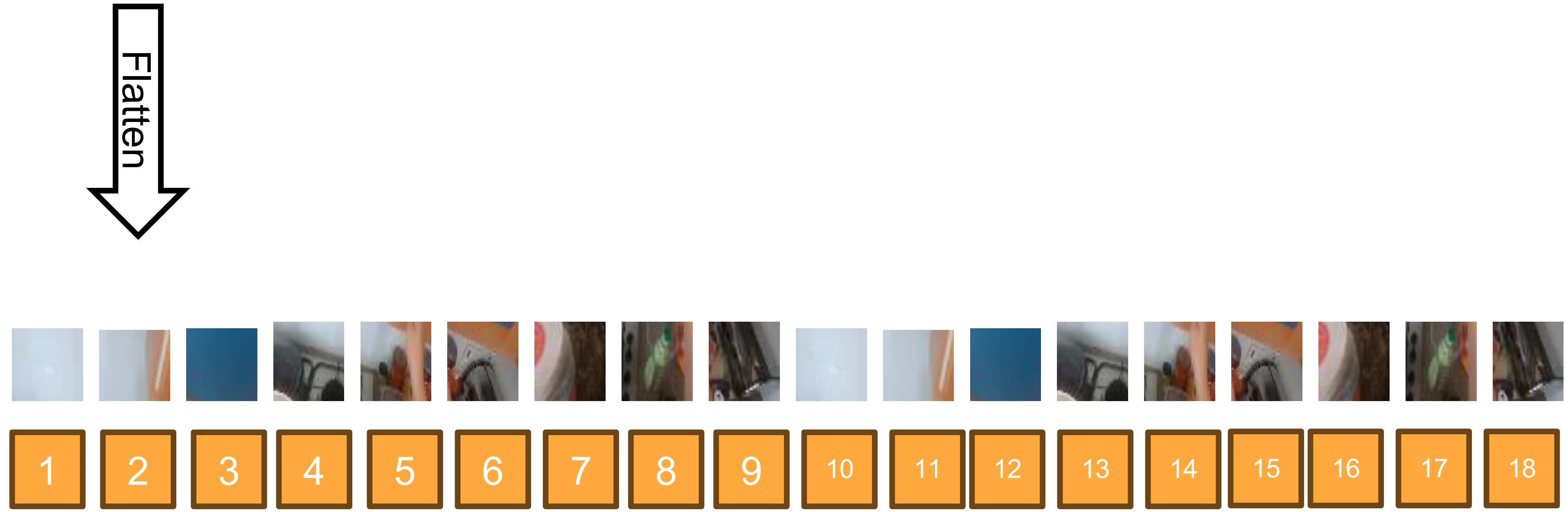


[TH] x W x C

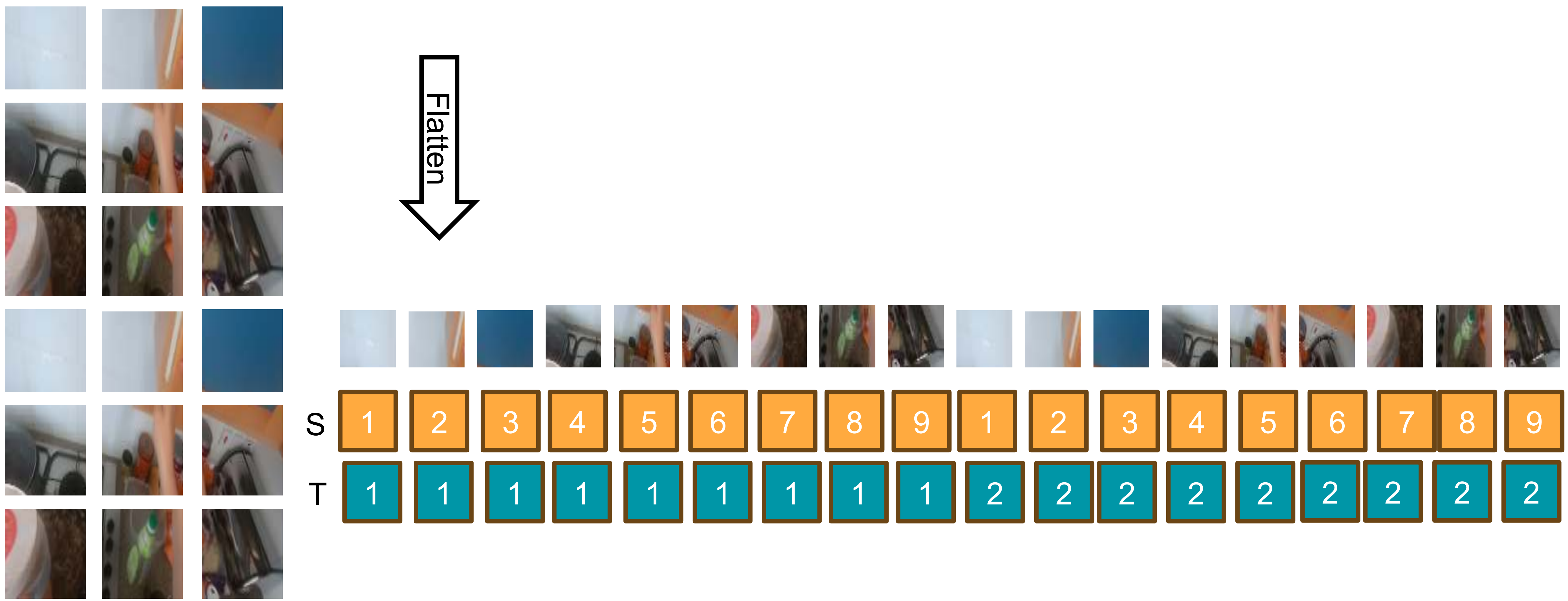
H x W x [CT]

How to Patch-ify a Video?



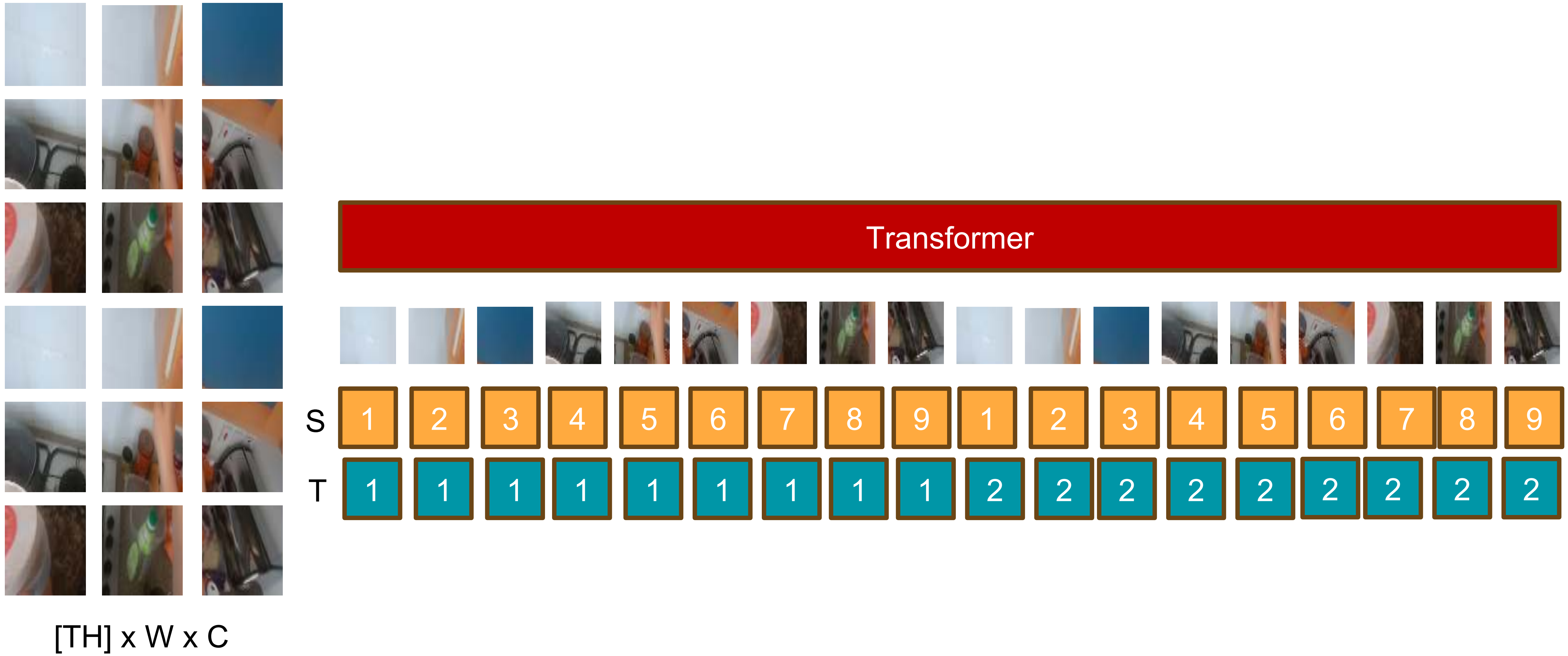


How to Patch-ify a Video?

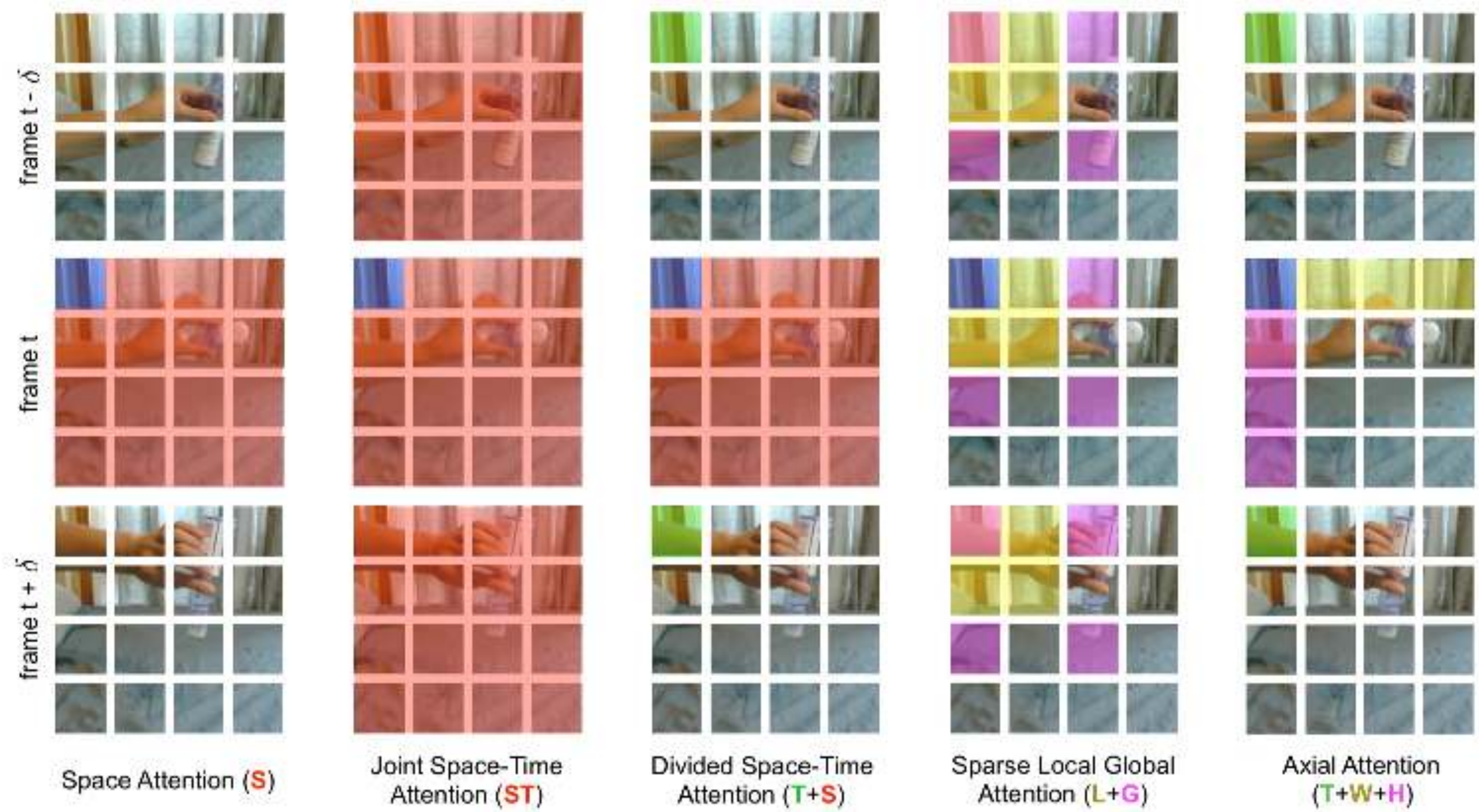


[TH] x W x C

How to Patch-ify a Video?

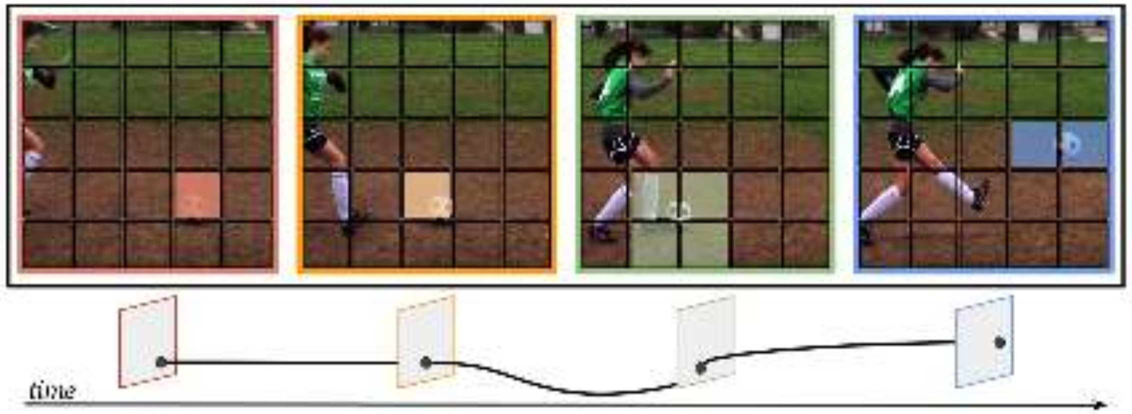


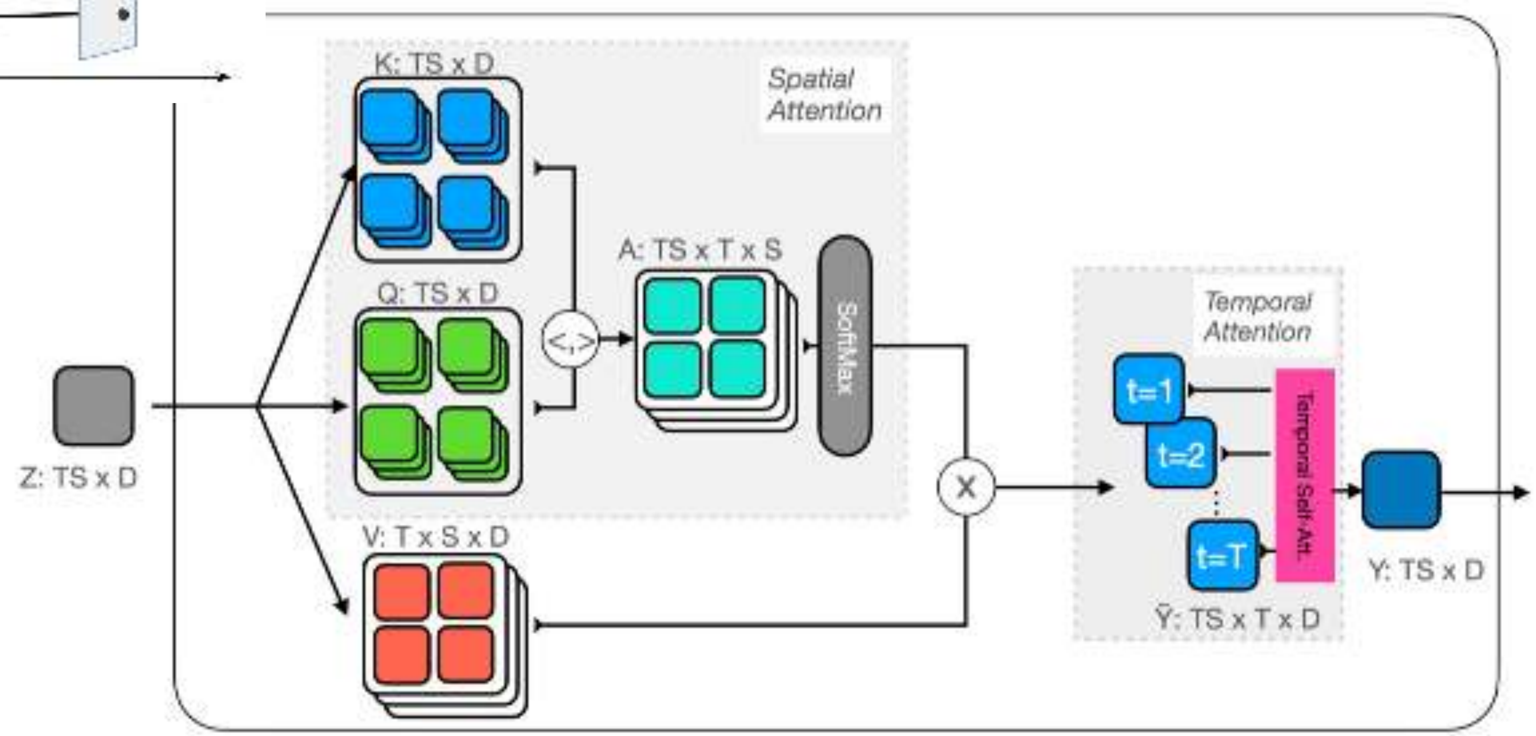
TimeSFormer



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

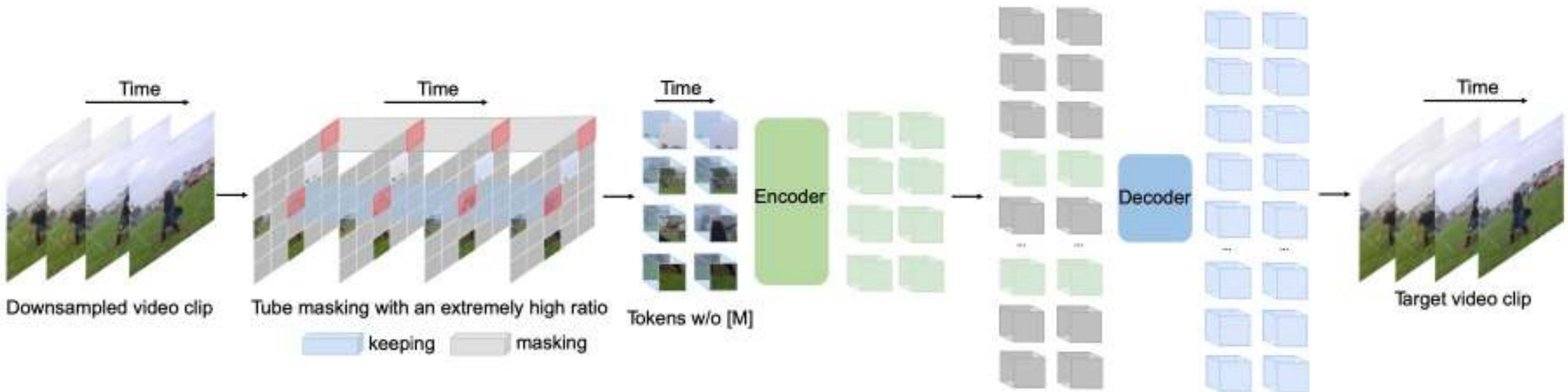
MotionFormer





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

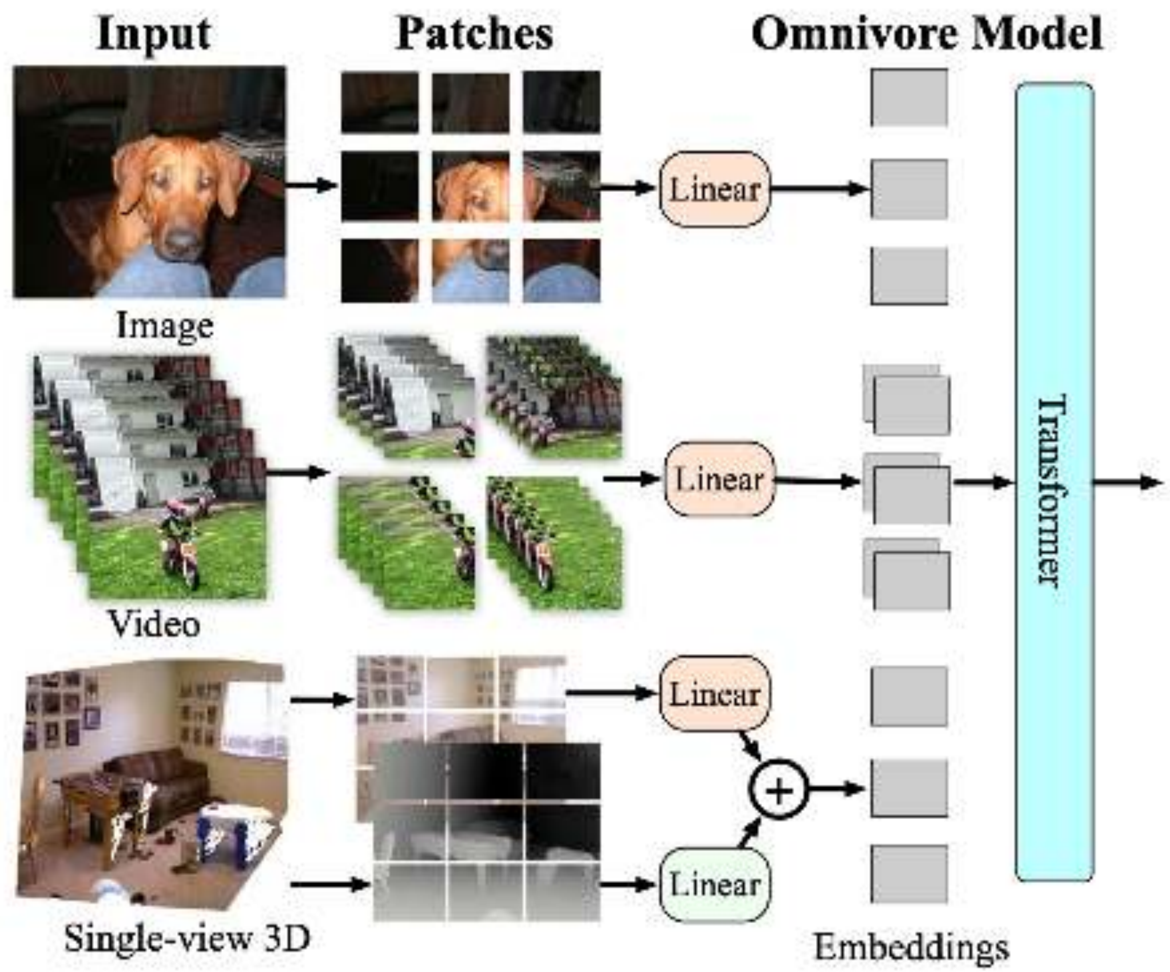
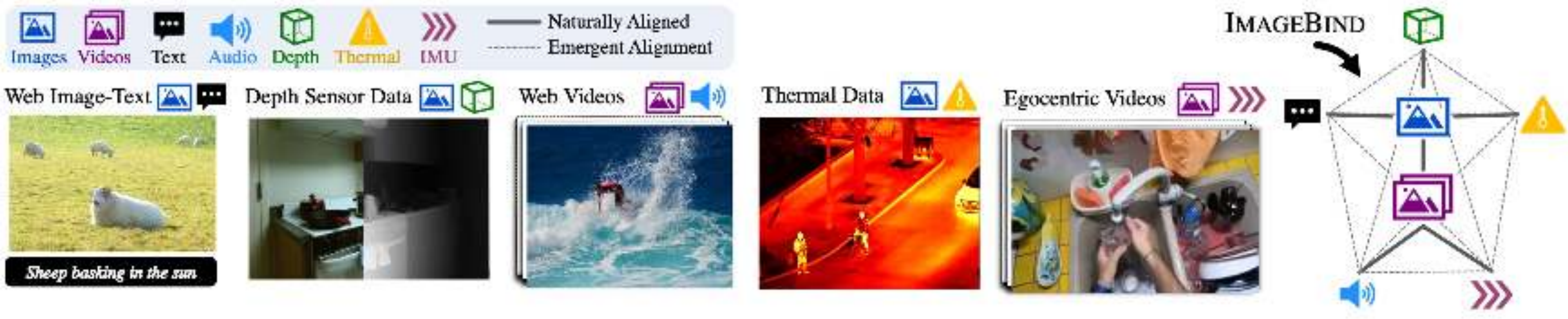


Figure 2. Multiple visual modalities in the OMNIVORE model.

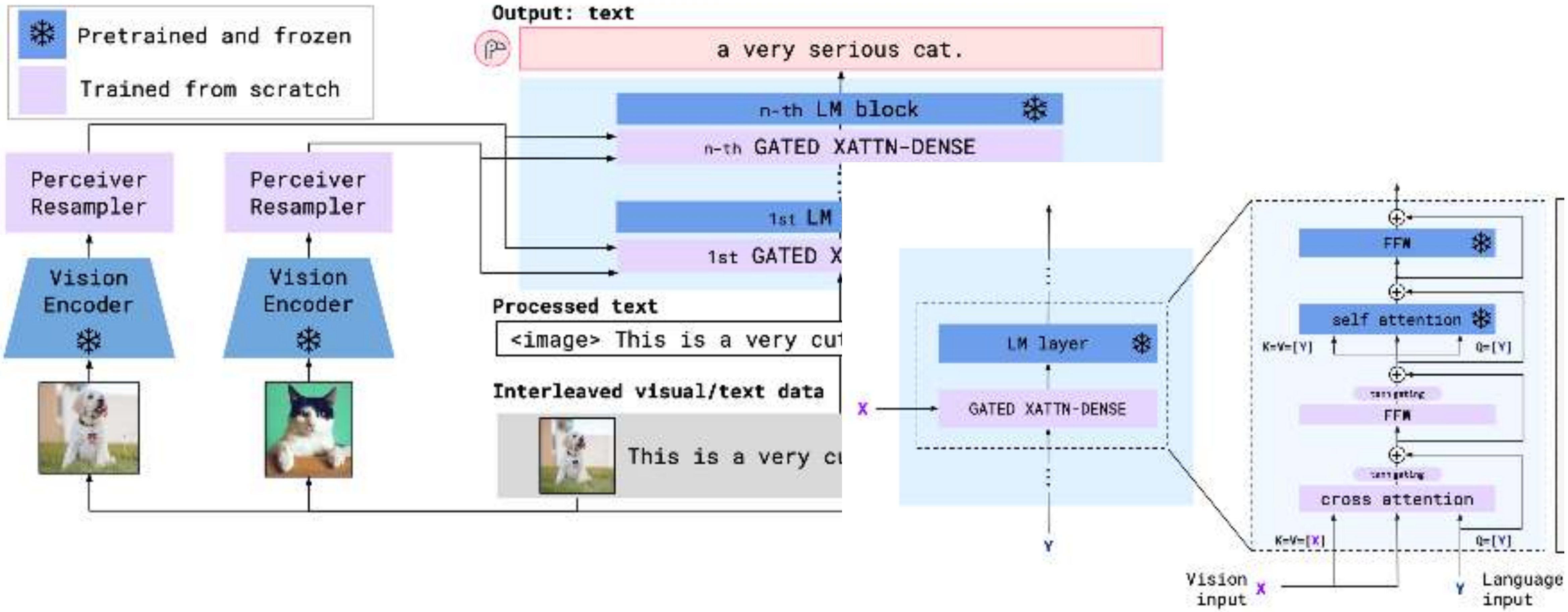
ImageBind



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

Gridhar et al (2023). IMAGEBIND: One Embedding Space To Bind Them All. CVPR

Flamingo



Alayrac et al (2022). Flamingo: a Visual Language Model for Few-Shot Learning. NeurIPS

InternVideo

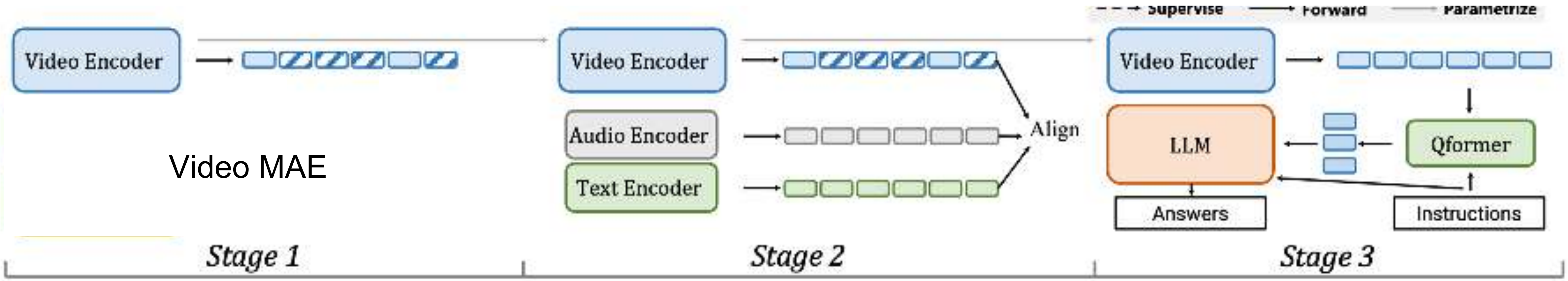
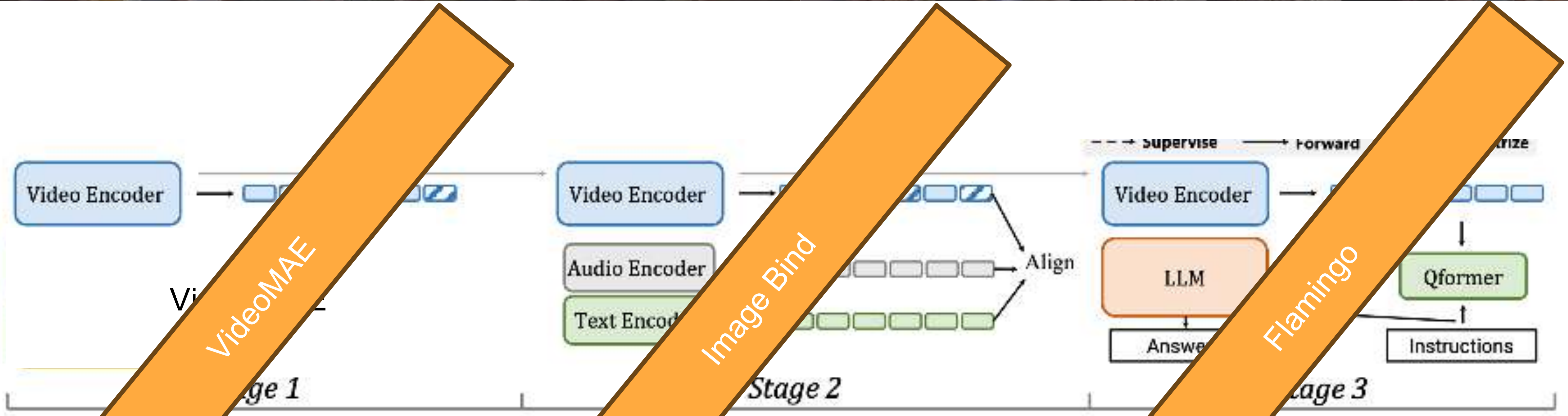


Figure 2: Framework of **InternVideo2**. It consists of three consecutive training phases: unmasked video token reconstruction, multimodal contrastive learning, and next token prediction. In stage 1, the video encoder is trained from scratch, while in stages 2 and 3, it is initialized from the version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

InternVideo



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

Two types of video understanding tasks

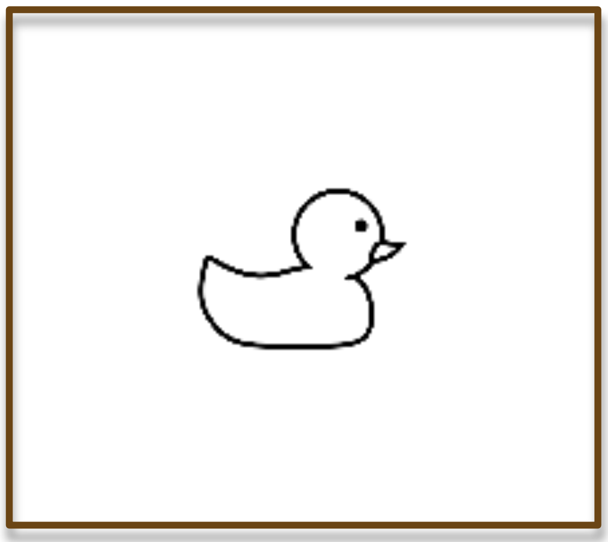
Video Understanding Tasks

- Analogous to Image-based Tasks
- Novel Video Tasks

Analogous Tasks

Image

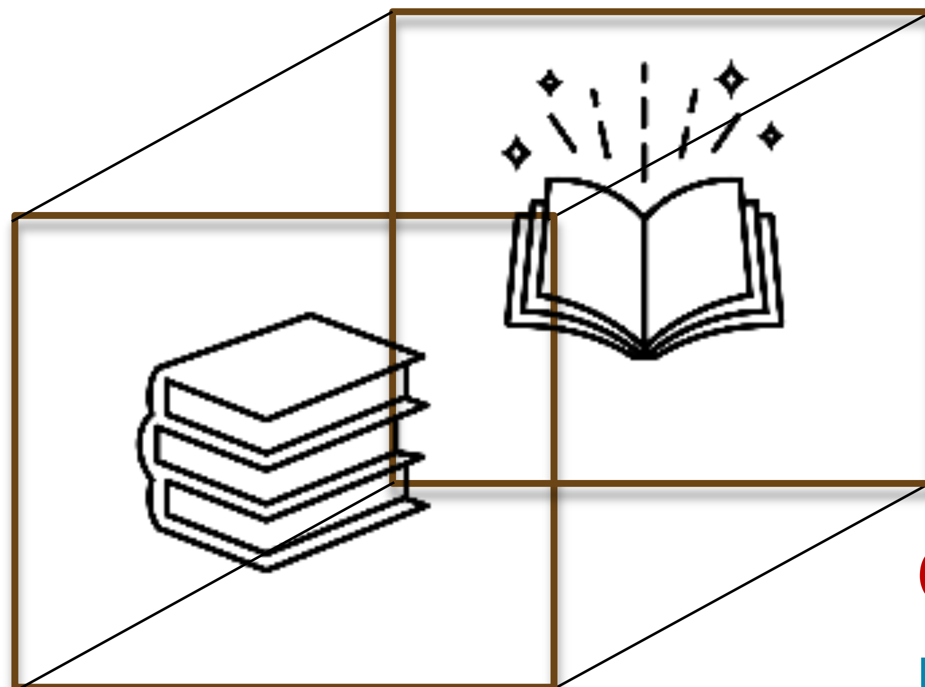
- Object Recognition



Duck

Video

- Action Recognition

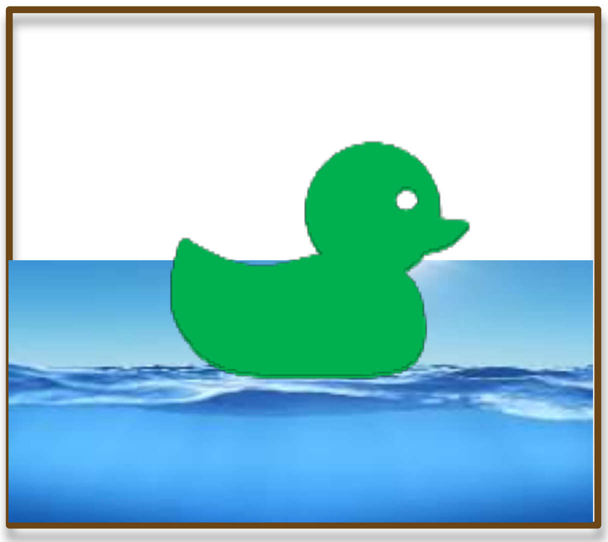


Open
Book

Analogous Tasks

Image

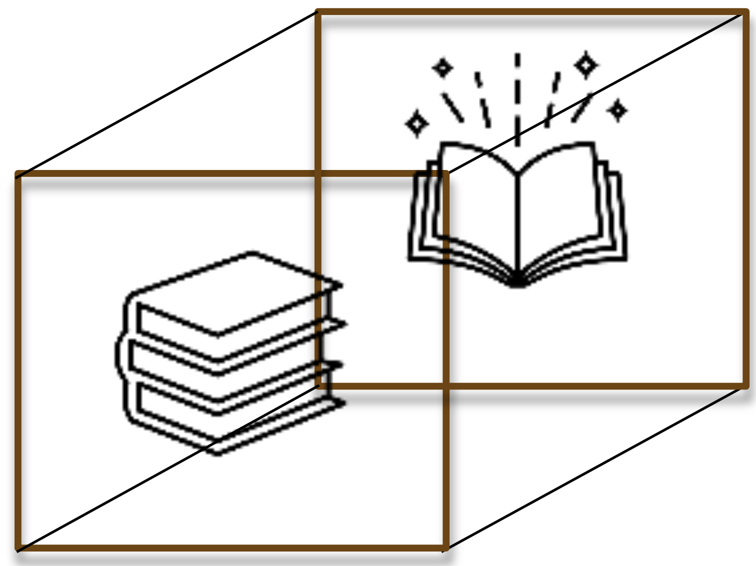
- Image Captioning



A green duck swimming
In clear water

Video

- Video Captioning

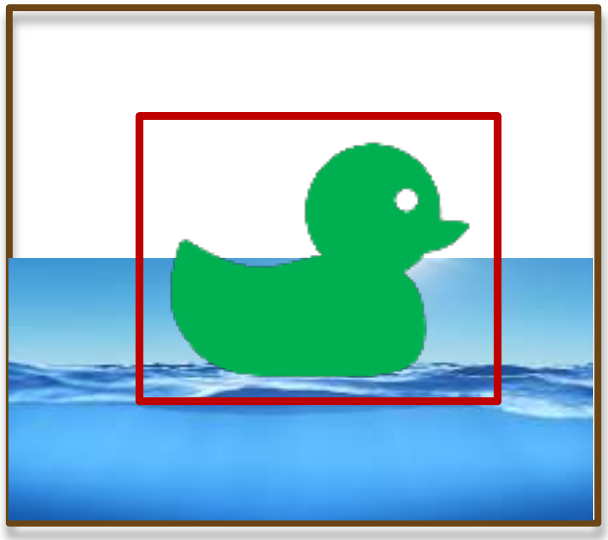


A book picked from top of the pile
and opened to a page in the middle

Analogous Tasks

Image

- Object Detection



Duck

Video

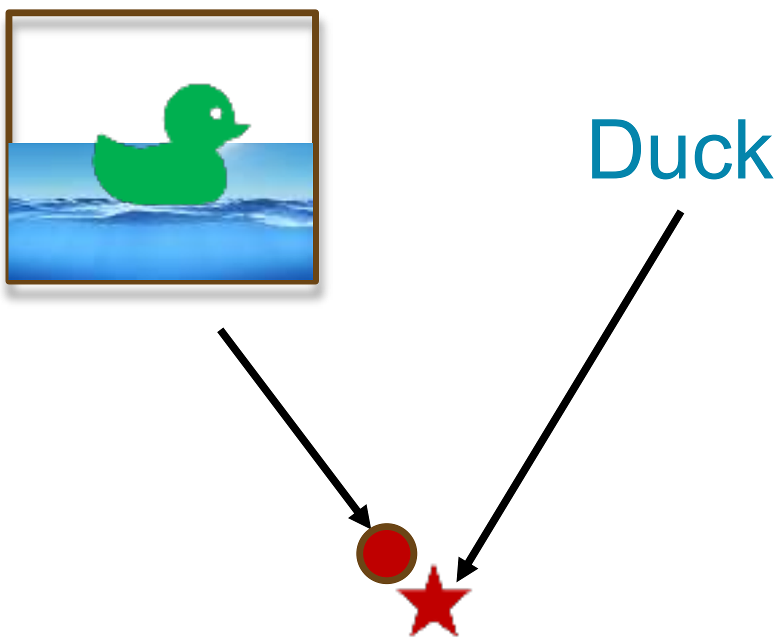
- Action Detection

Open Book

Analogous Tasks

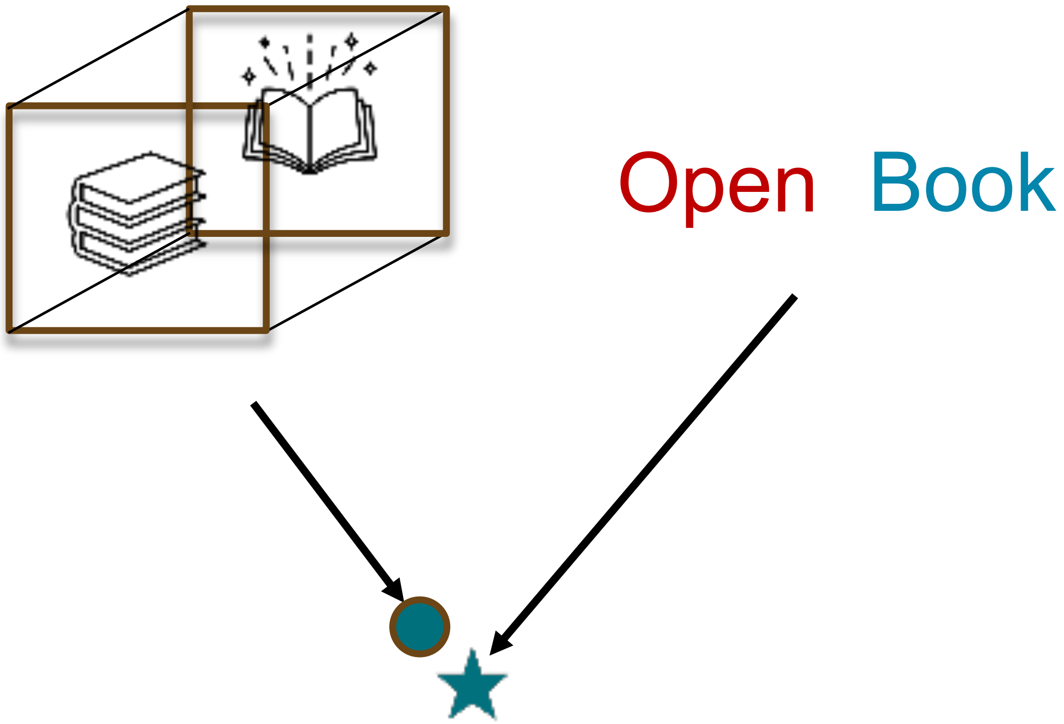
Image

- Image Retrieval



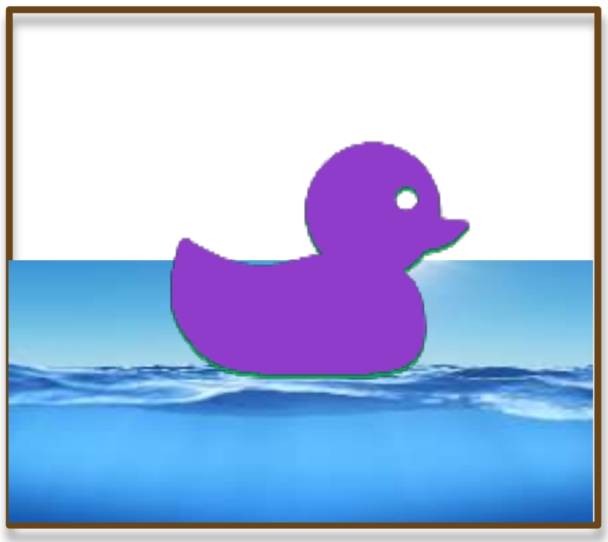
Video

- Video Retrieval



Image

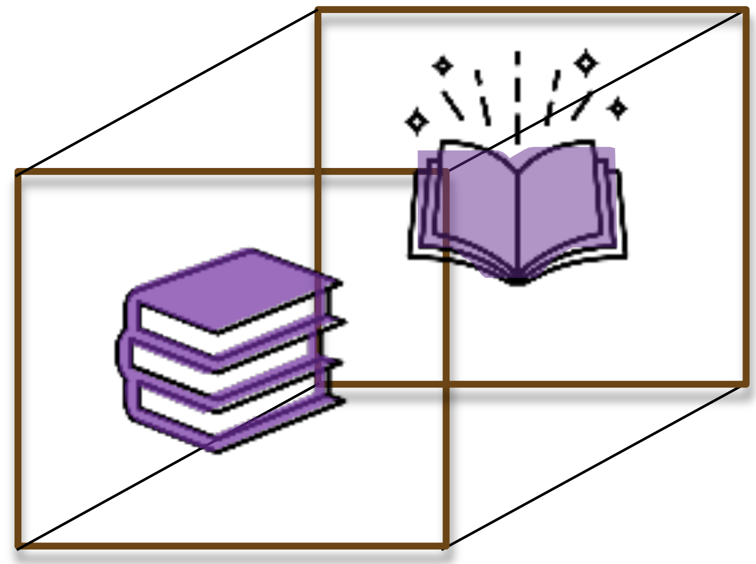
- Object Segmentation



Duck

Video

- Video Object Segmentation



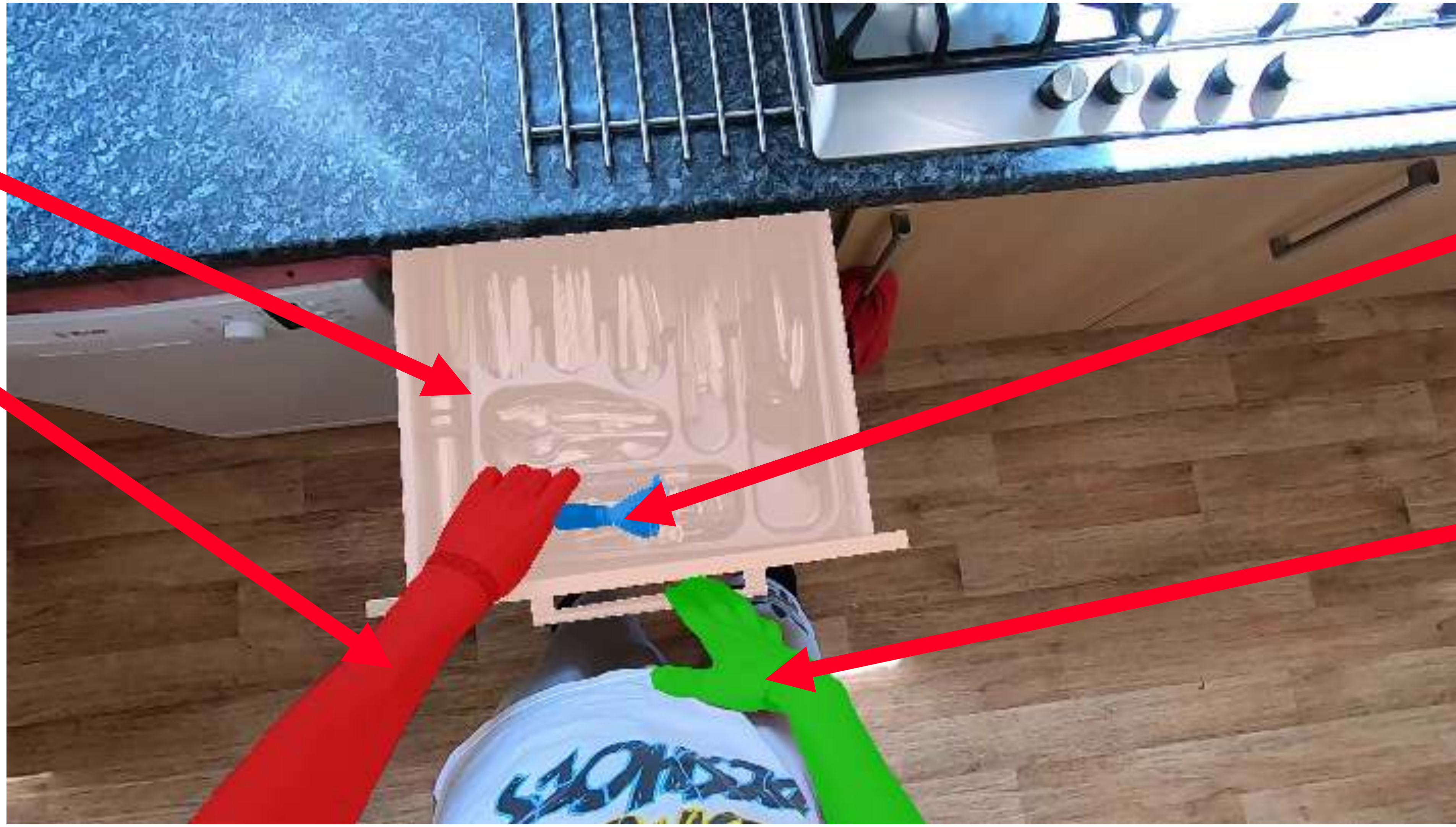
Book

EPIC-KITCHENS VISOR

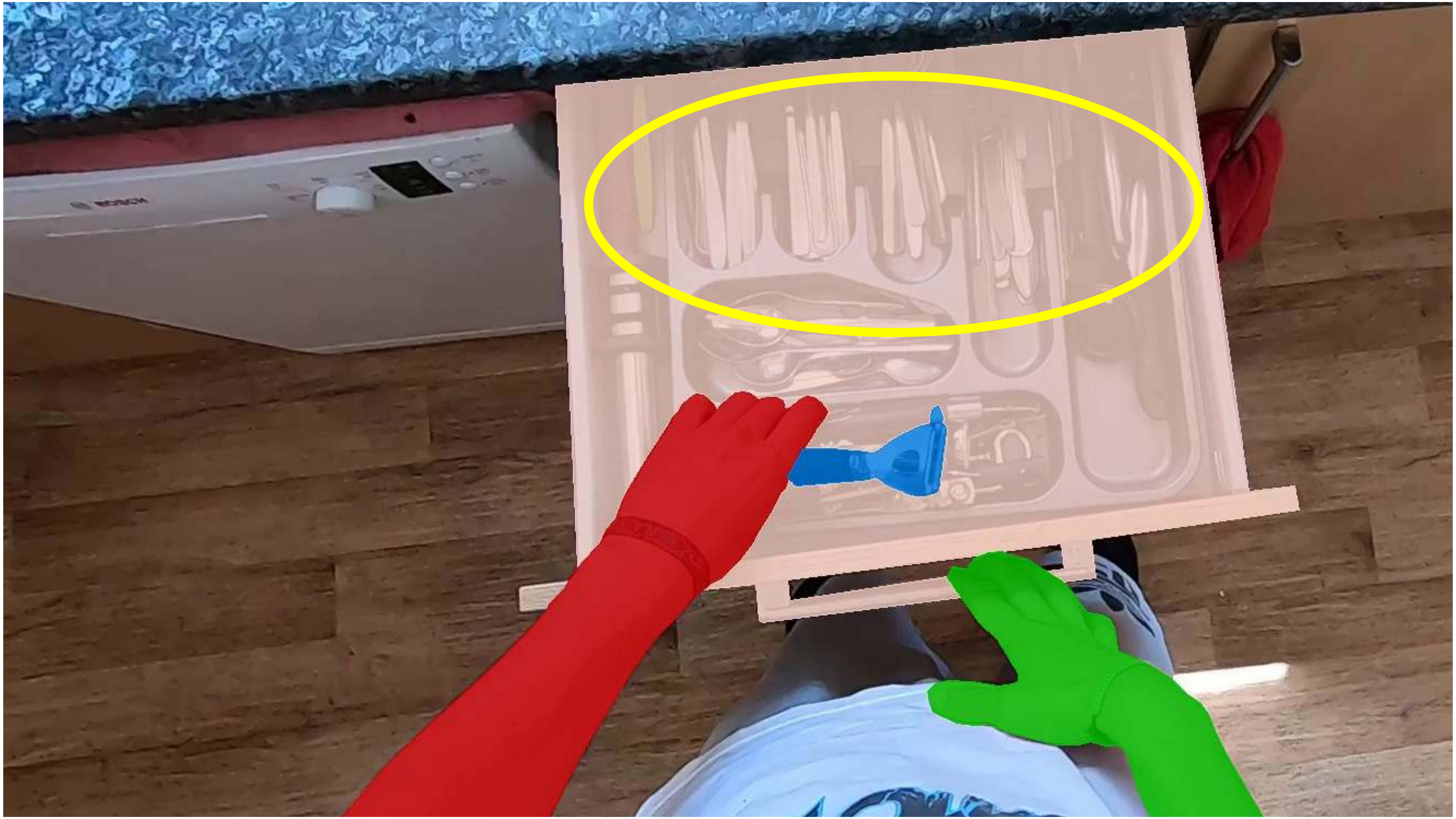
with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma,
Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler

EPIC-KITCHENS VISOR

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

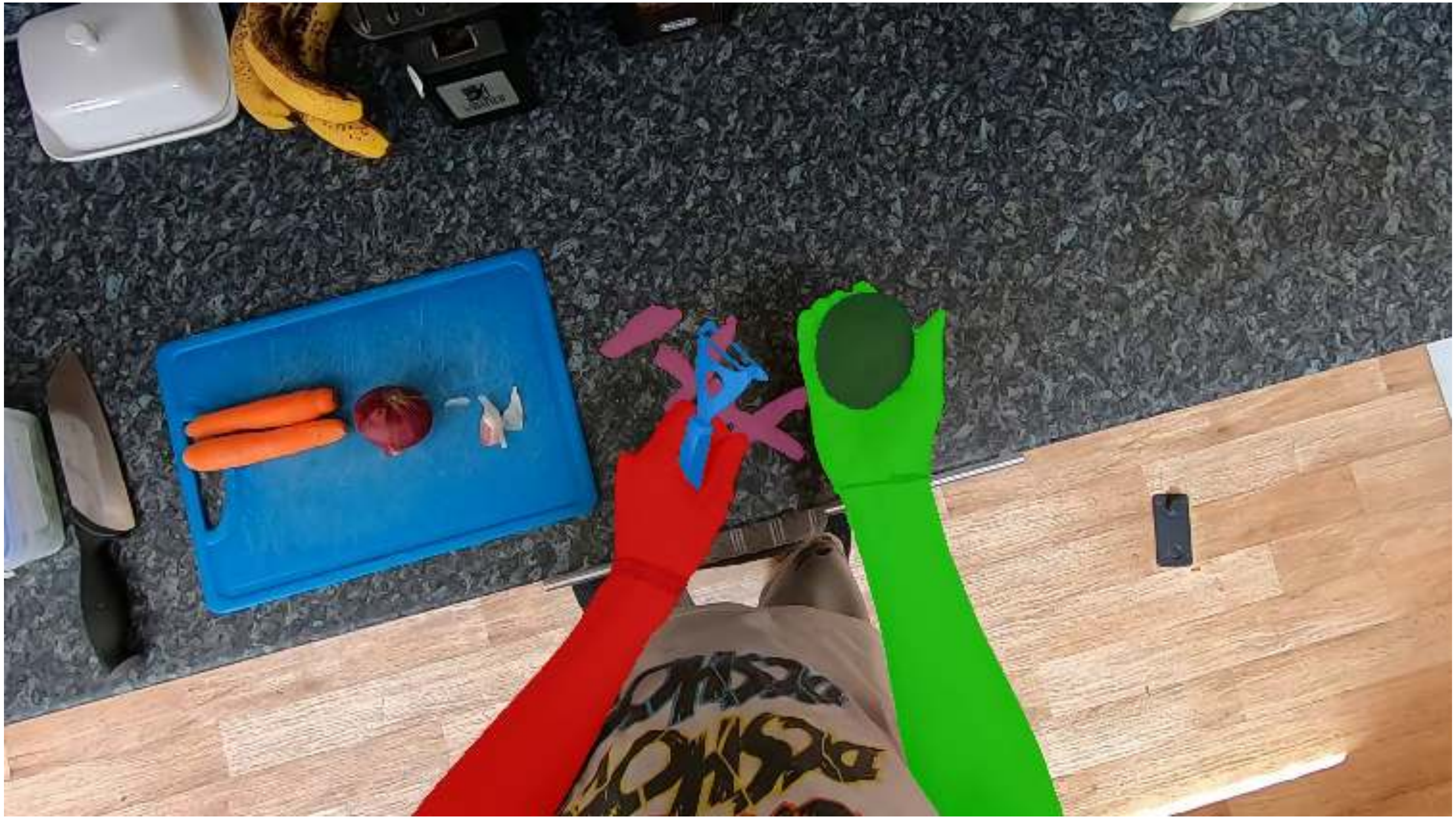


A Darkhalil et al (2022). EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. *NeurIPS*



EPIC-KITCHENS VISOR

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



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

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*

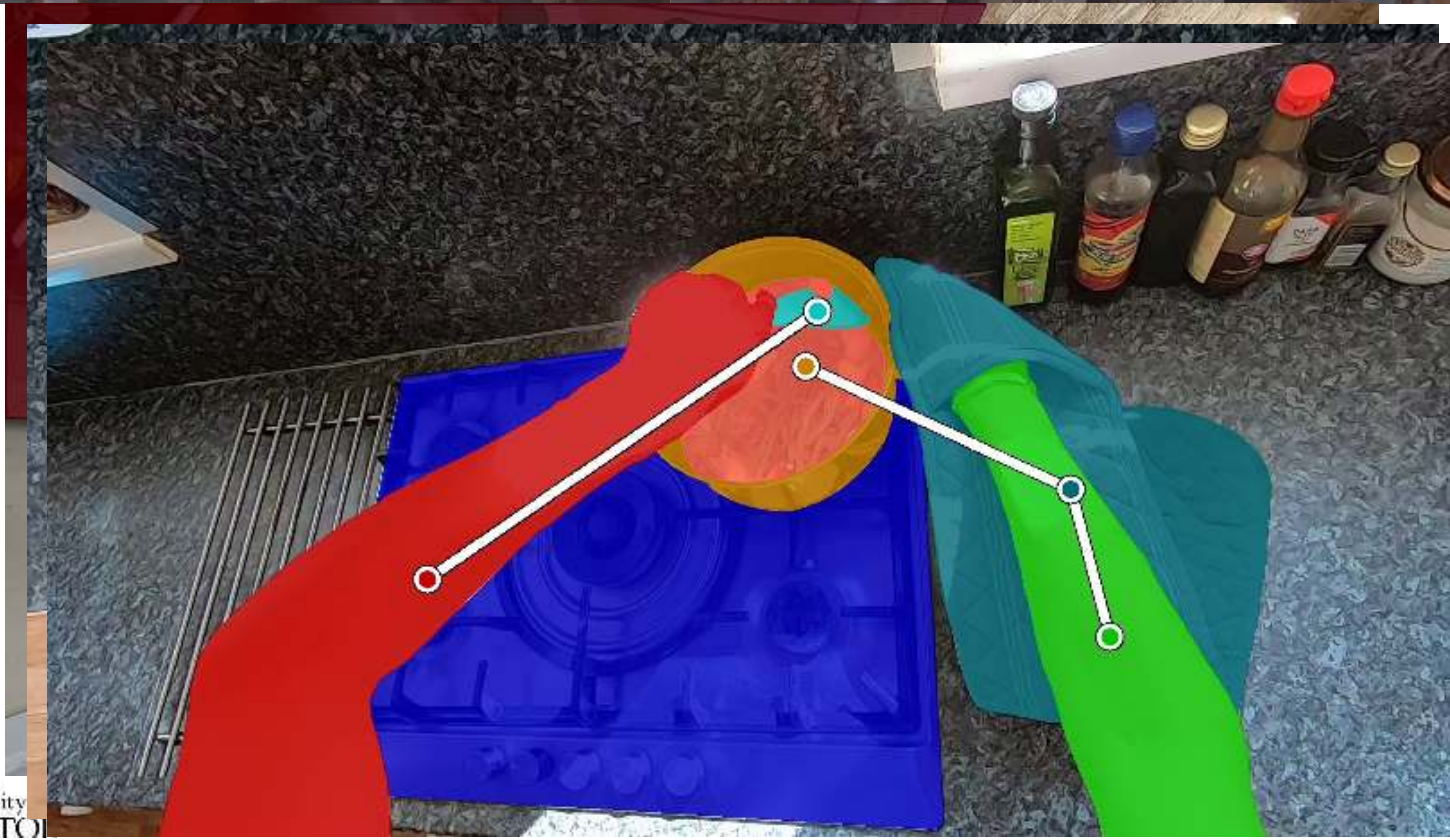
EPIC-KITCHENS VISOR

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*

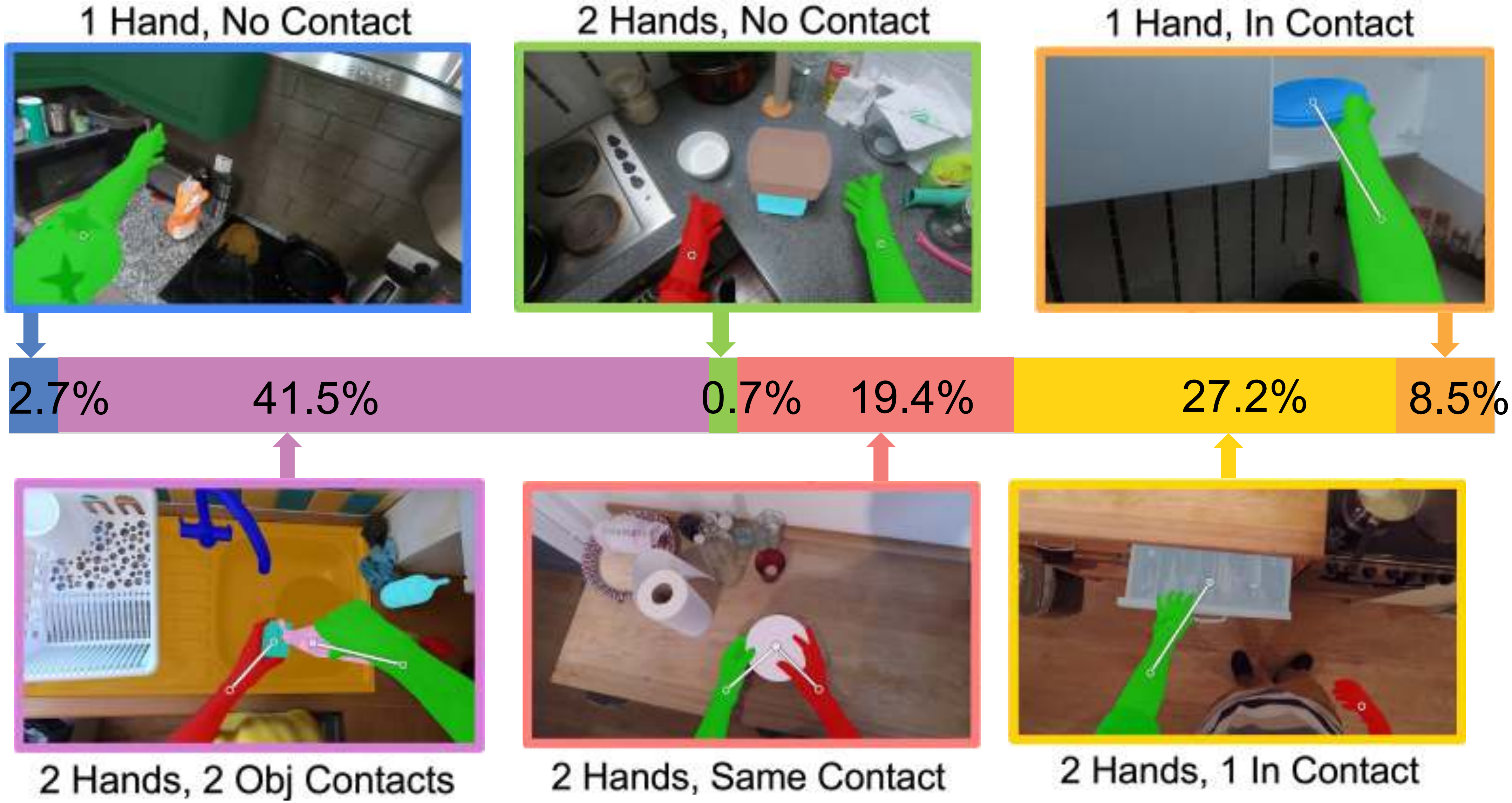
VISOR Relations

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



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*

Analogous Tasks

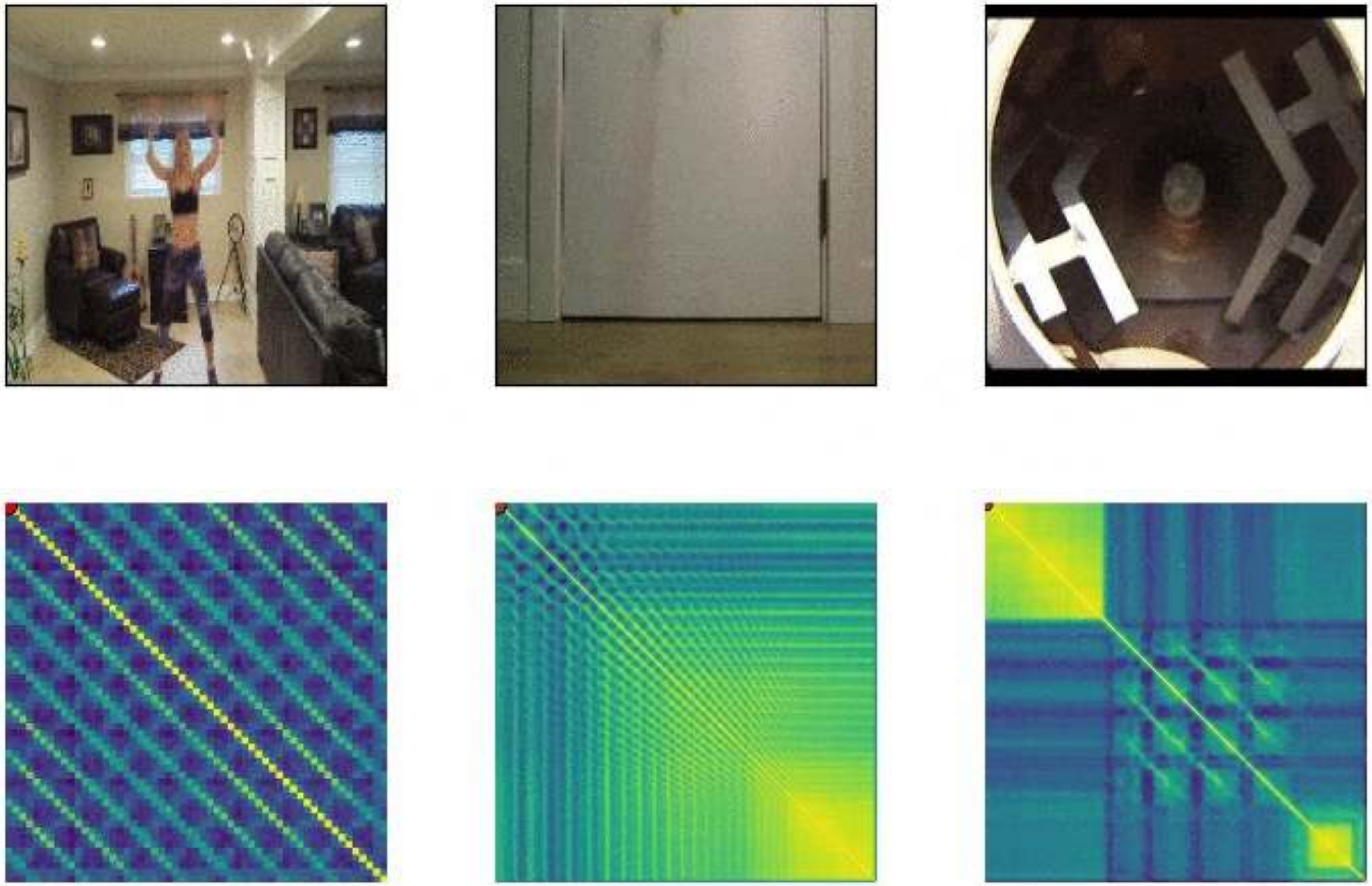
Image

- Object Counting

Video

- Action Counting

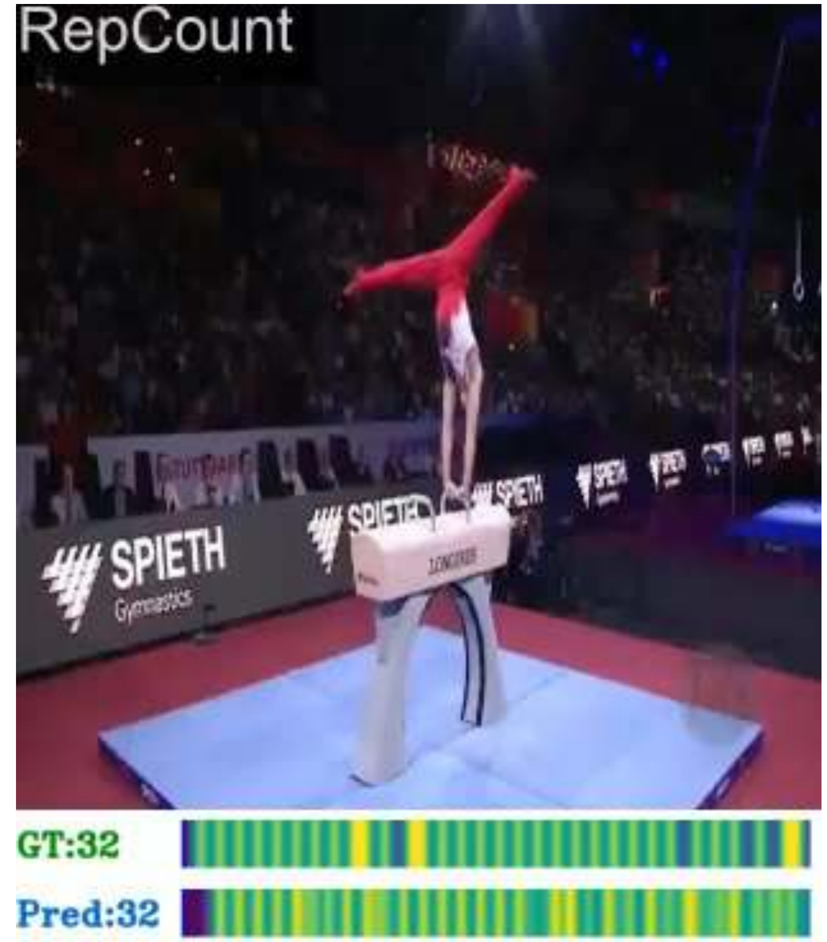
Countix



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



Sinha et al (2024). Every Shot Counts: Using Exemplars for Repetition Counting in Videos. ACCV

Analogous Tasks

Image

- Text-to-image Generation

Stable Diffusion

Video

- Text-to-Video Generation

SORA

Text-to-Video Generation

Prompt: Historical footage of California during the gold rush.

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

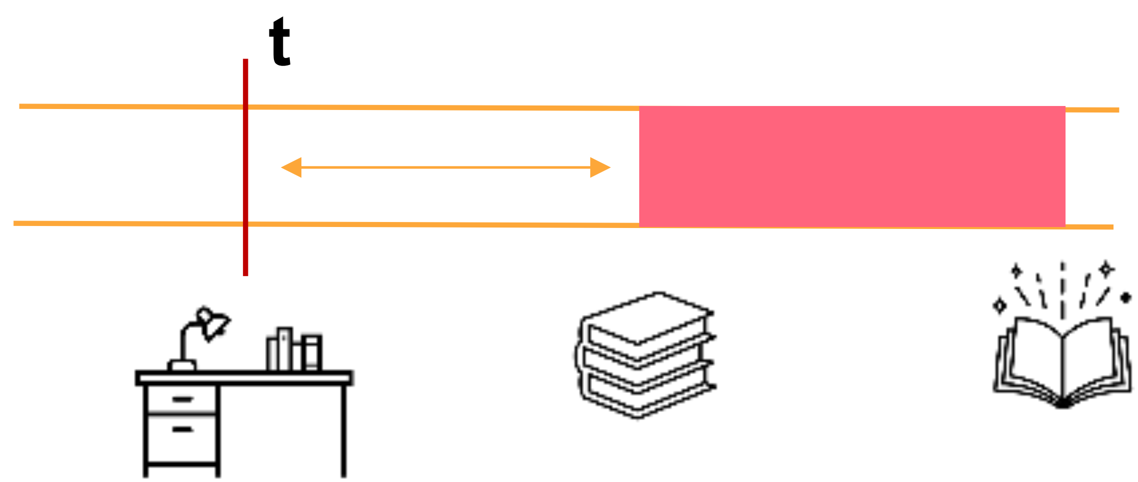
Non-Analogous Tasks

Image

Video

- Action Anticipation

What will happen after 1 second?



Open Book

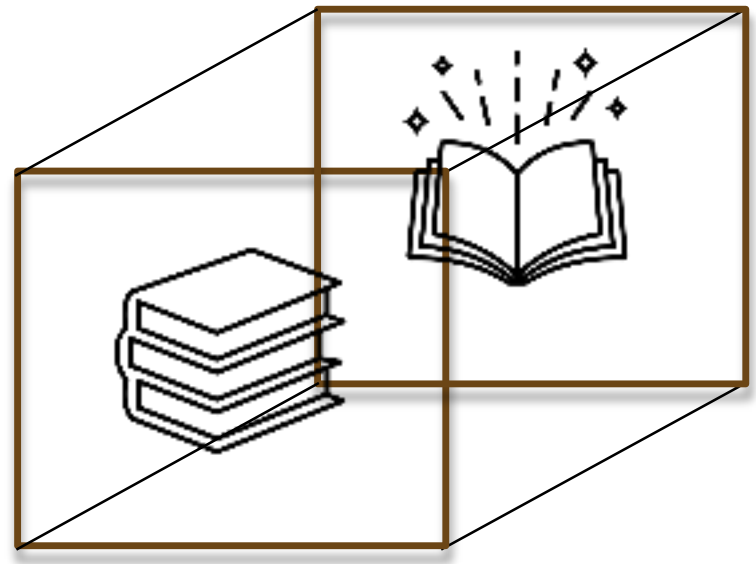
Non-Analogous Tasks

Image

Video

- Manner of action

How did you open the book?



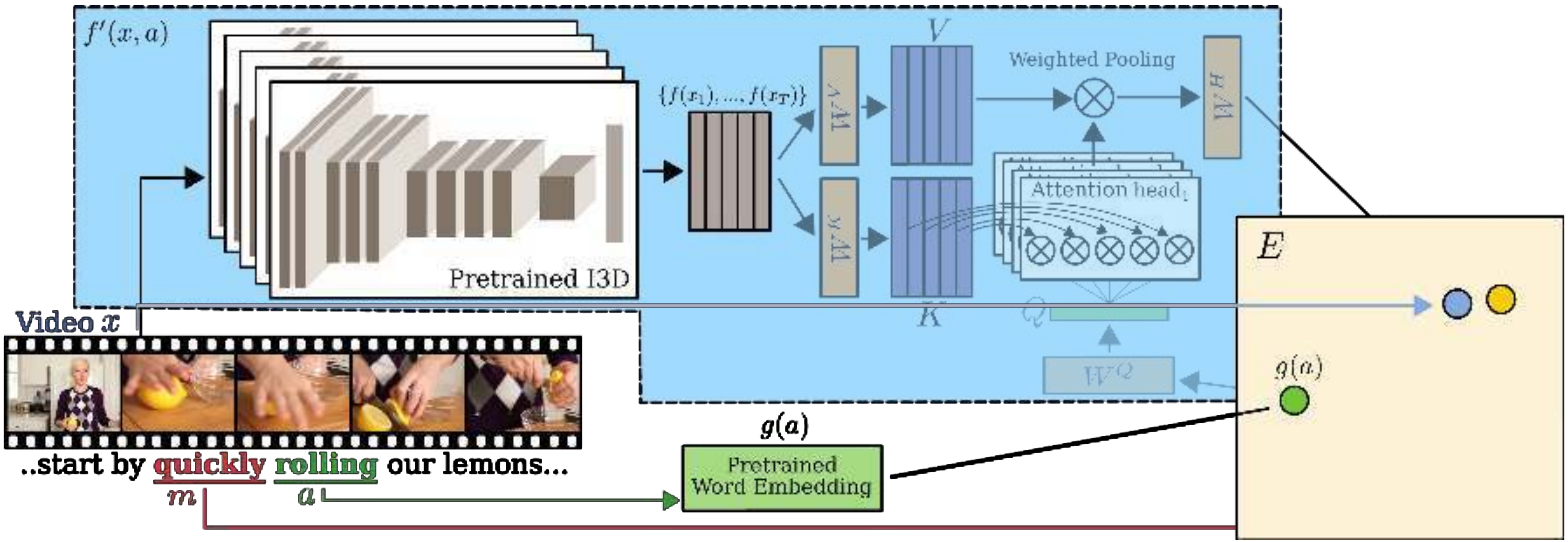
How?

with: Hazel Doughty
Ivan Laptev
Walterio Mayol-Cuevas

Doughty et al (2020). Action Modifiers: Learning from Adverbs in Instructional Videos. CVPR

How?

with: Hazel Doughty
Ivan Laptev
Walterio Mayol-Cuevas



Doughty et al (2020). Action Modifiers: Learning from Adverbs in Instructional Videos. CVPR

How?

with: Hazel Doughty
Ivan Laptev
Walterio Mayol-Cuevas

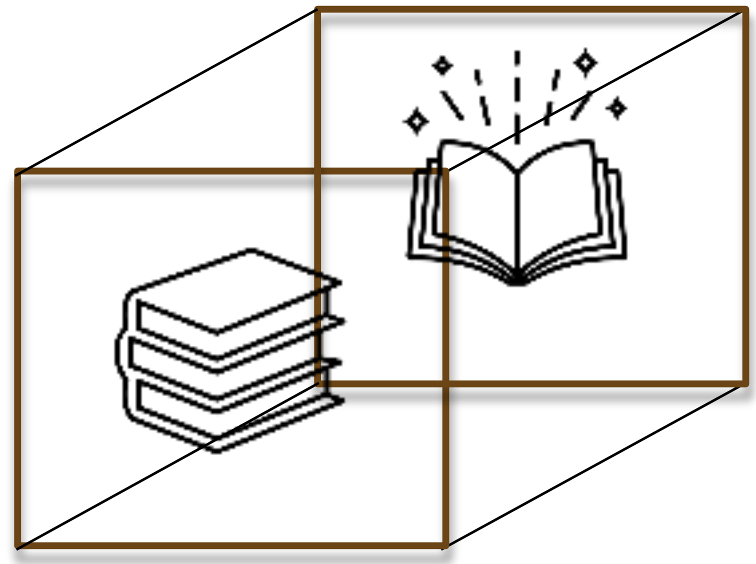
... we're going to mix these up real **quick**...

Non-Analogous Tasks

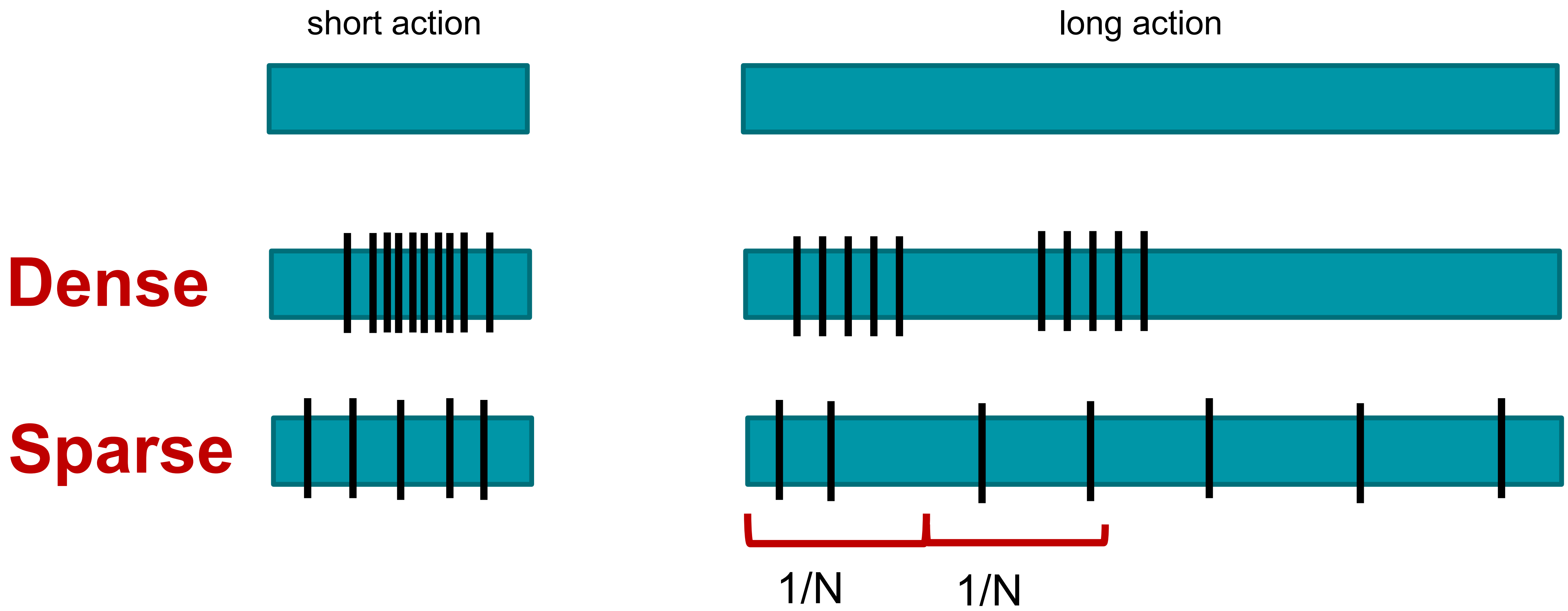
Image

Video

- Frame Selection and Attribution

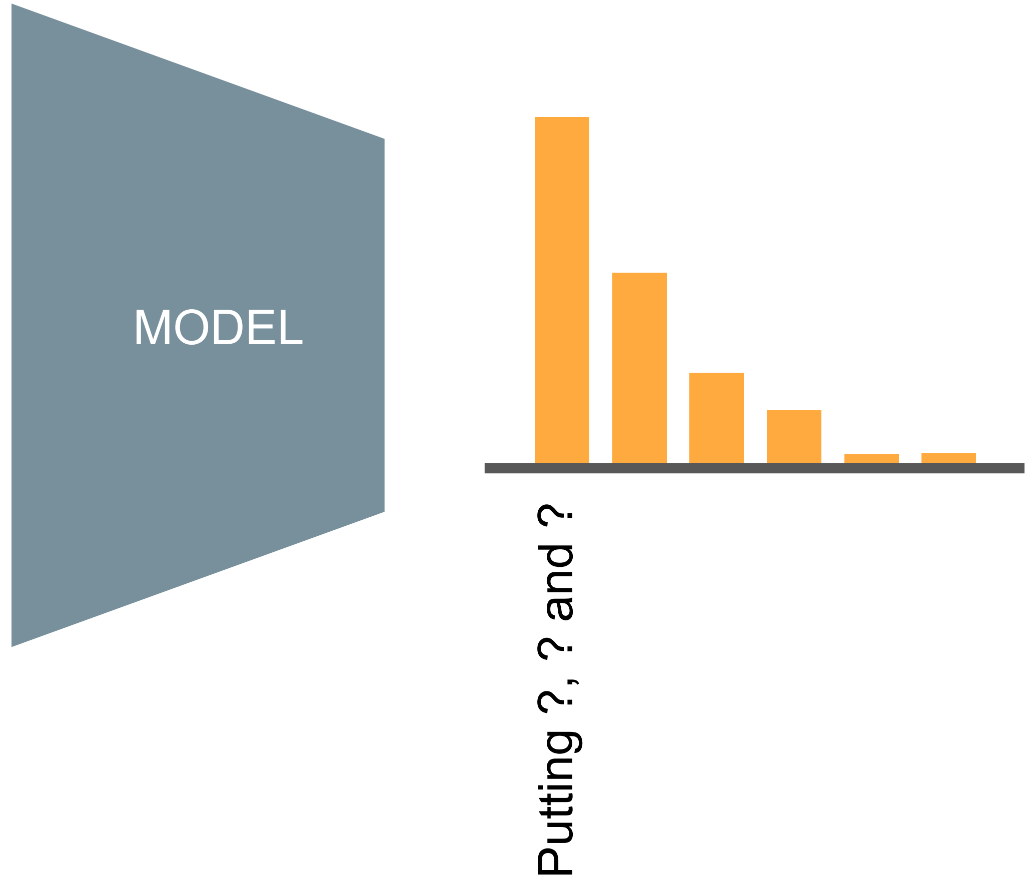


Two sampling approaches



Frame Attributions in Video Models

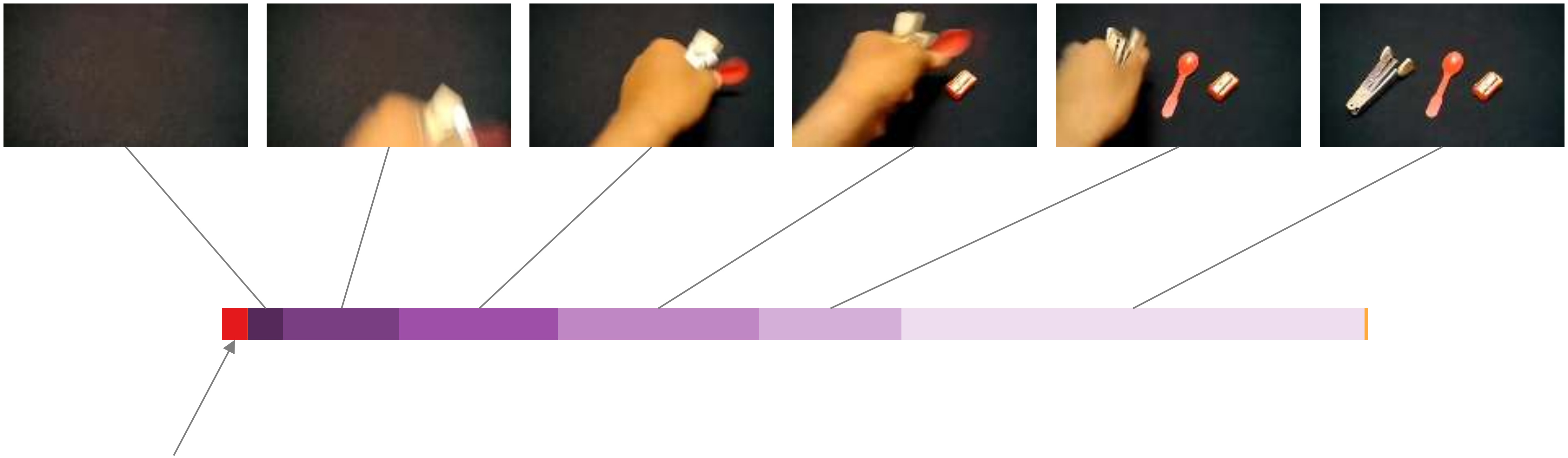
with: Will Price



W Price, D Damen (2020). Play Fair: Frame Attribution in Video Models. ACCV

Frame Attributions in Video Models

with: Will Price

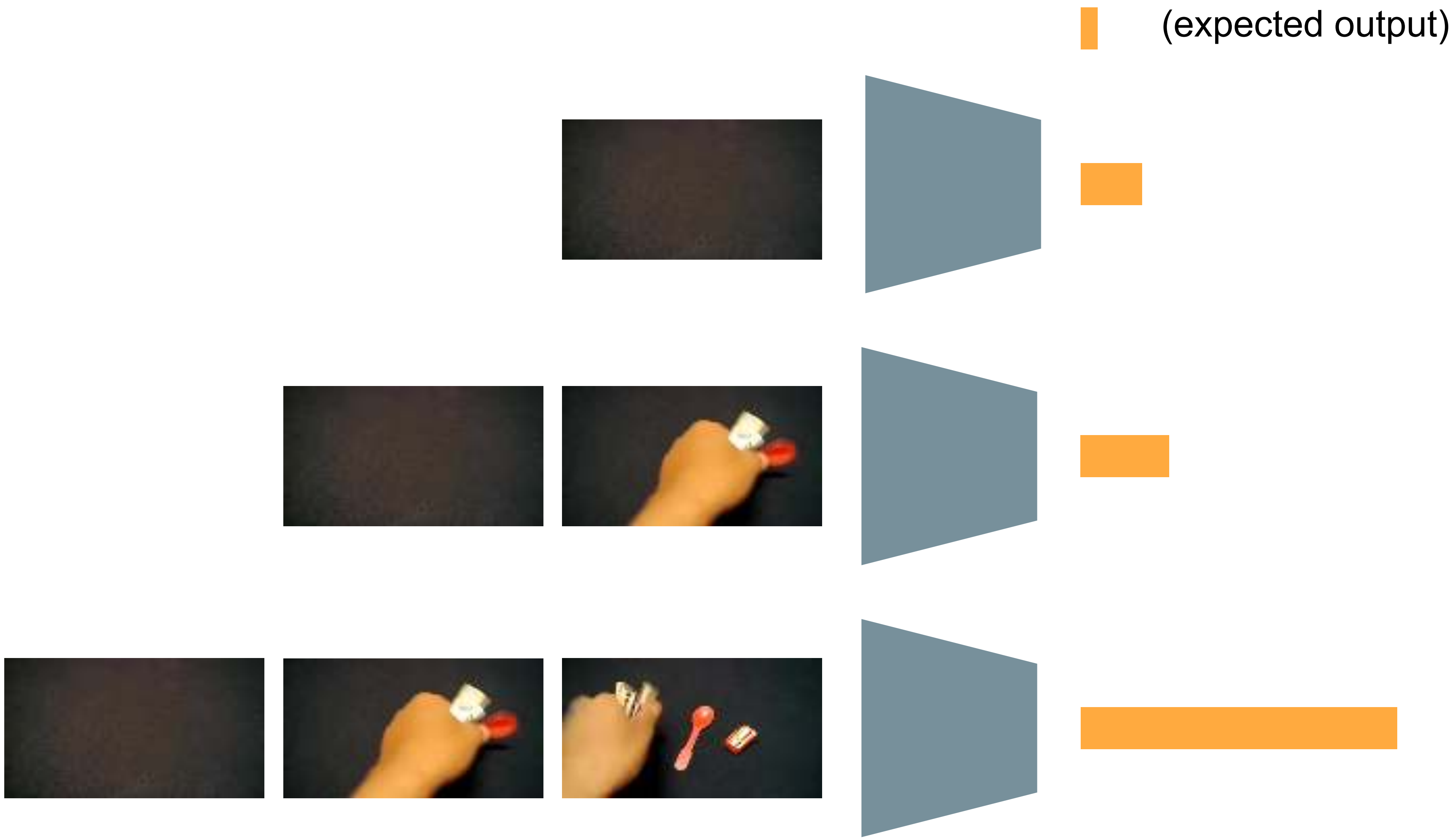


Expected output
(Prior probability for
classification model)

W Price, D Damen (2020). Play Fair: Frame Attribution in Video Models. ACCV

Frame Attributions in Video Models

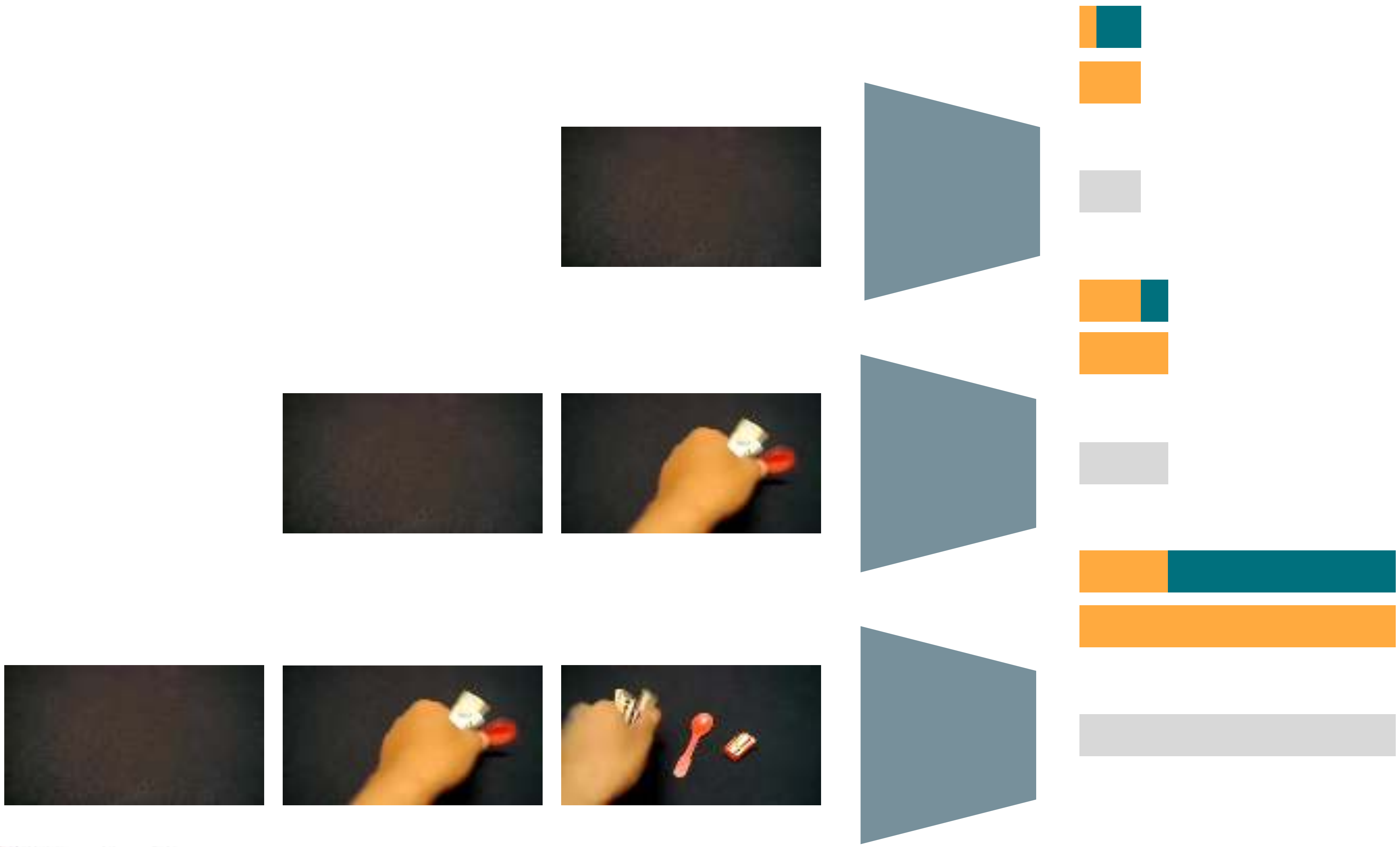
with: Will Price



W Price, D Damen (2020). Play Fair: Frame Attribution in Video Models. ACCV

Frame Attributions in Video Models

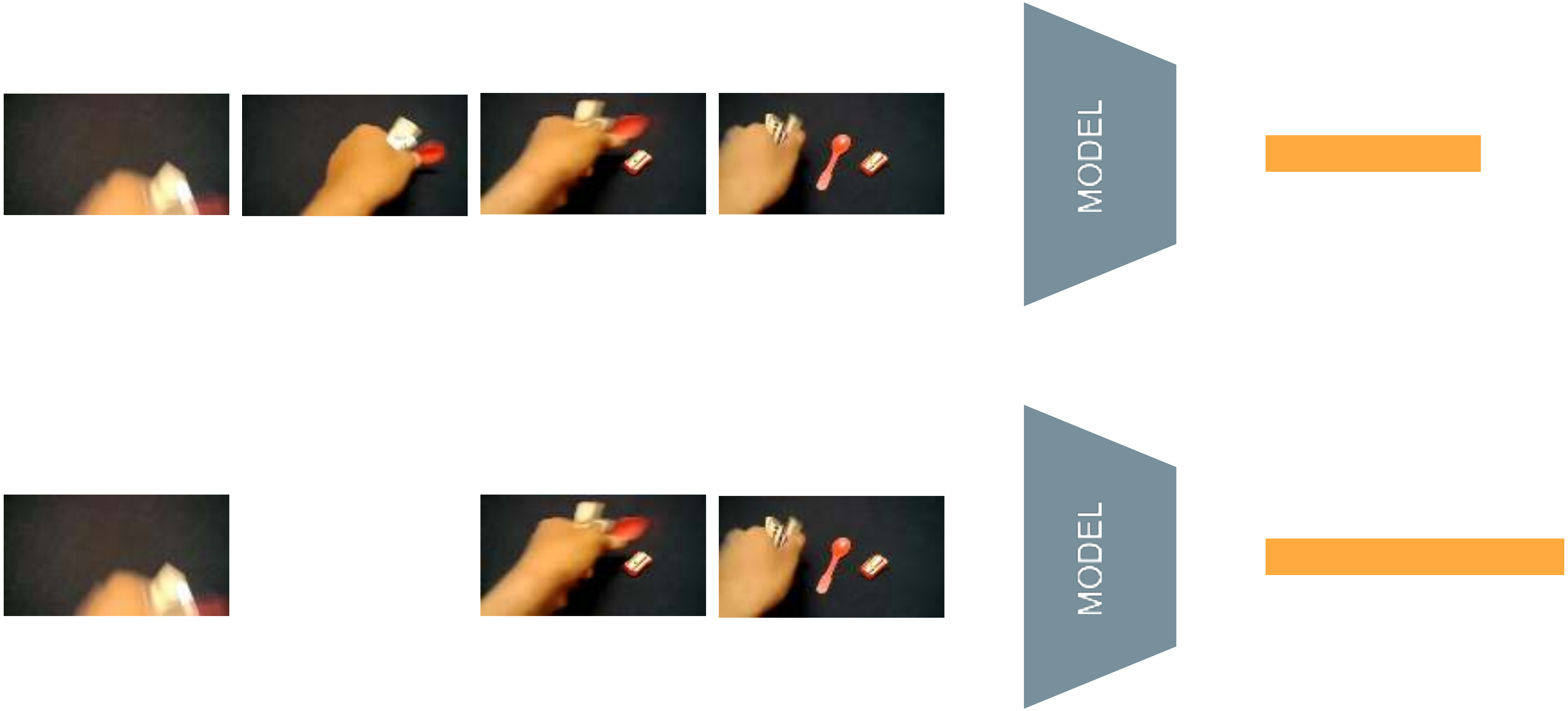
with: Will Price



W Price, D Damen (2020). Play Fair: Frame Attribution in Video Models. ACCV

Frame Attributions in Video Models

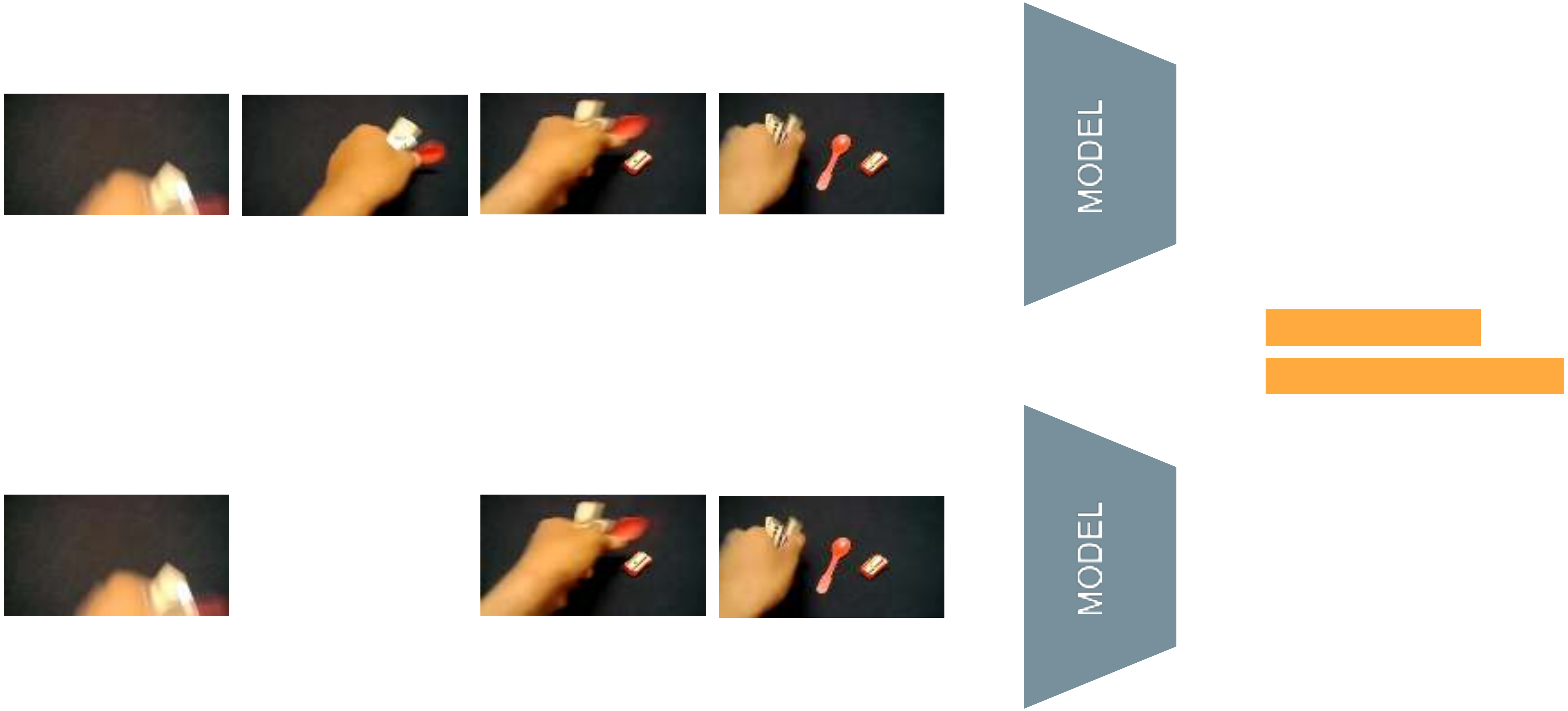
with: Will Price



W Price, D Damen (2020). Play Fair: Frame Attribution in Video Models. ACCV

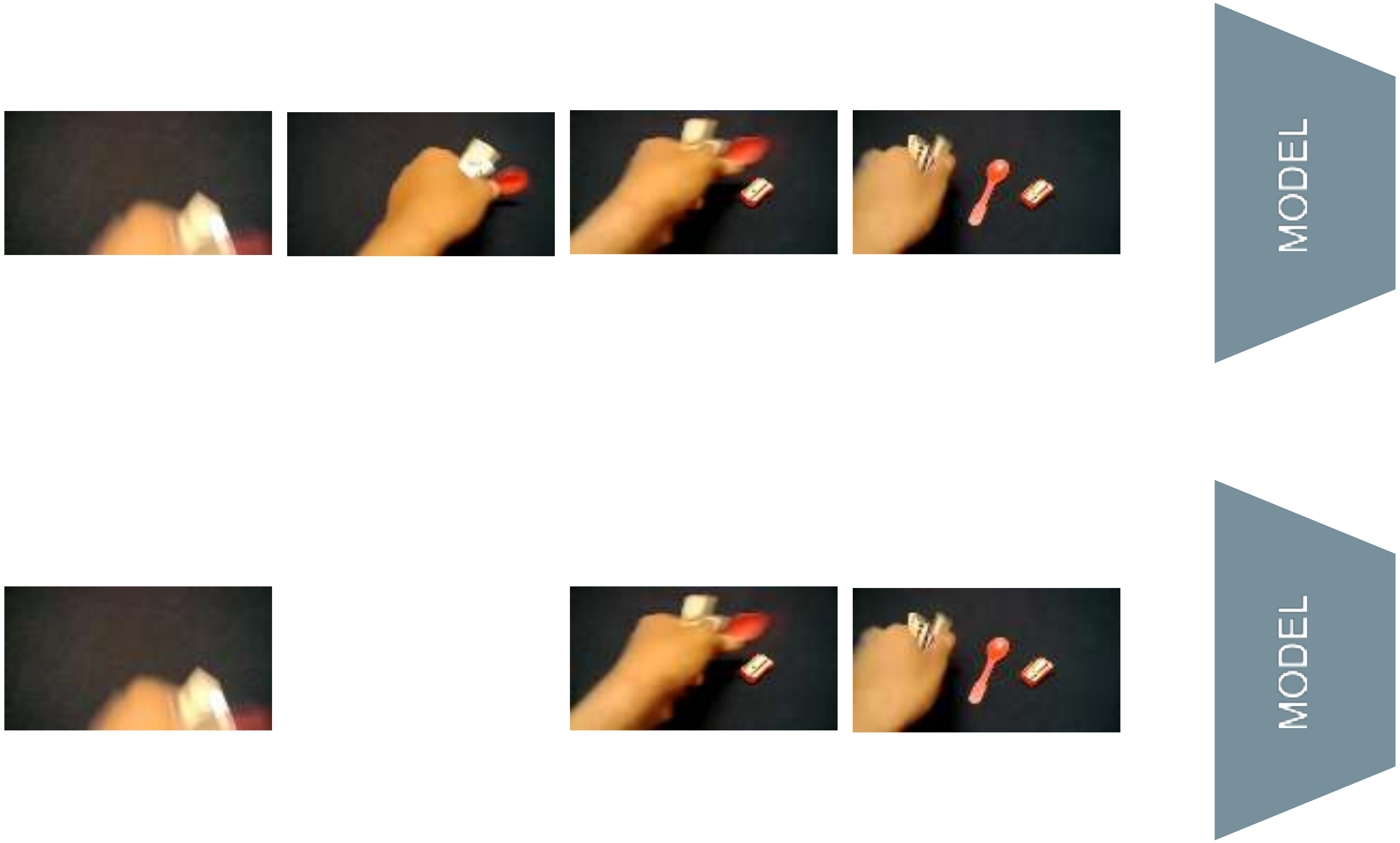
Frame Attributions in Video Models

with: Will Price



Frame Attributions in Video Models

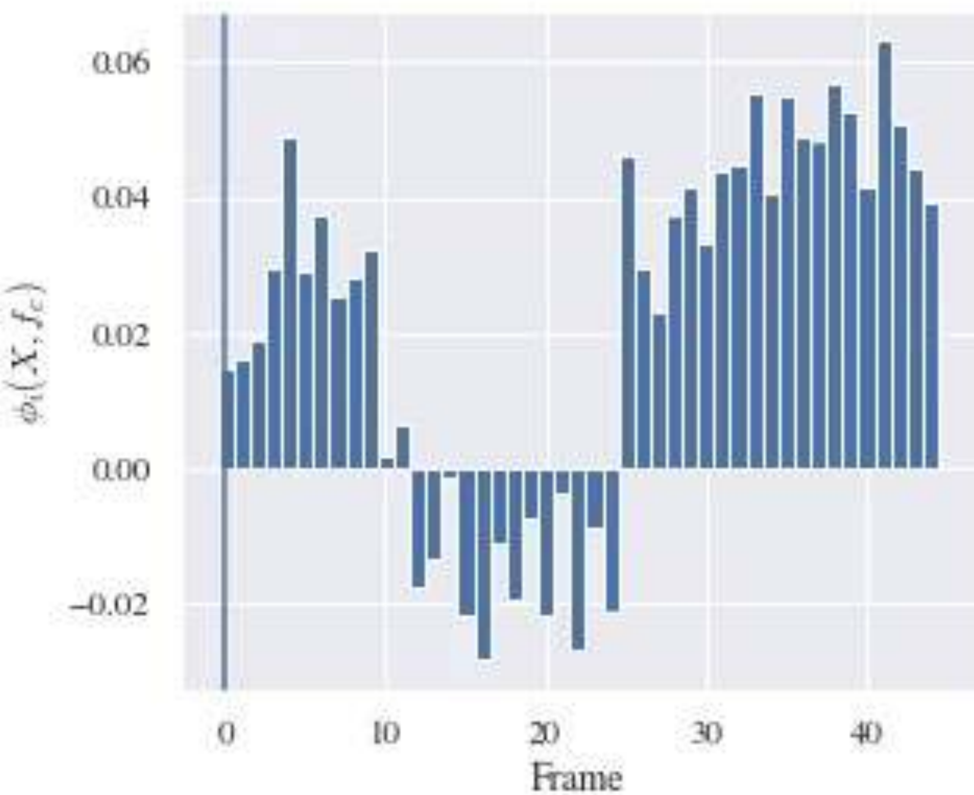
with: Will Price



$\Delta_3(\{1,2,4,5\}) = -.2$

Frame Attributions in Video Models

with: Will Price



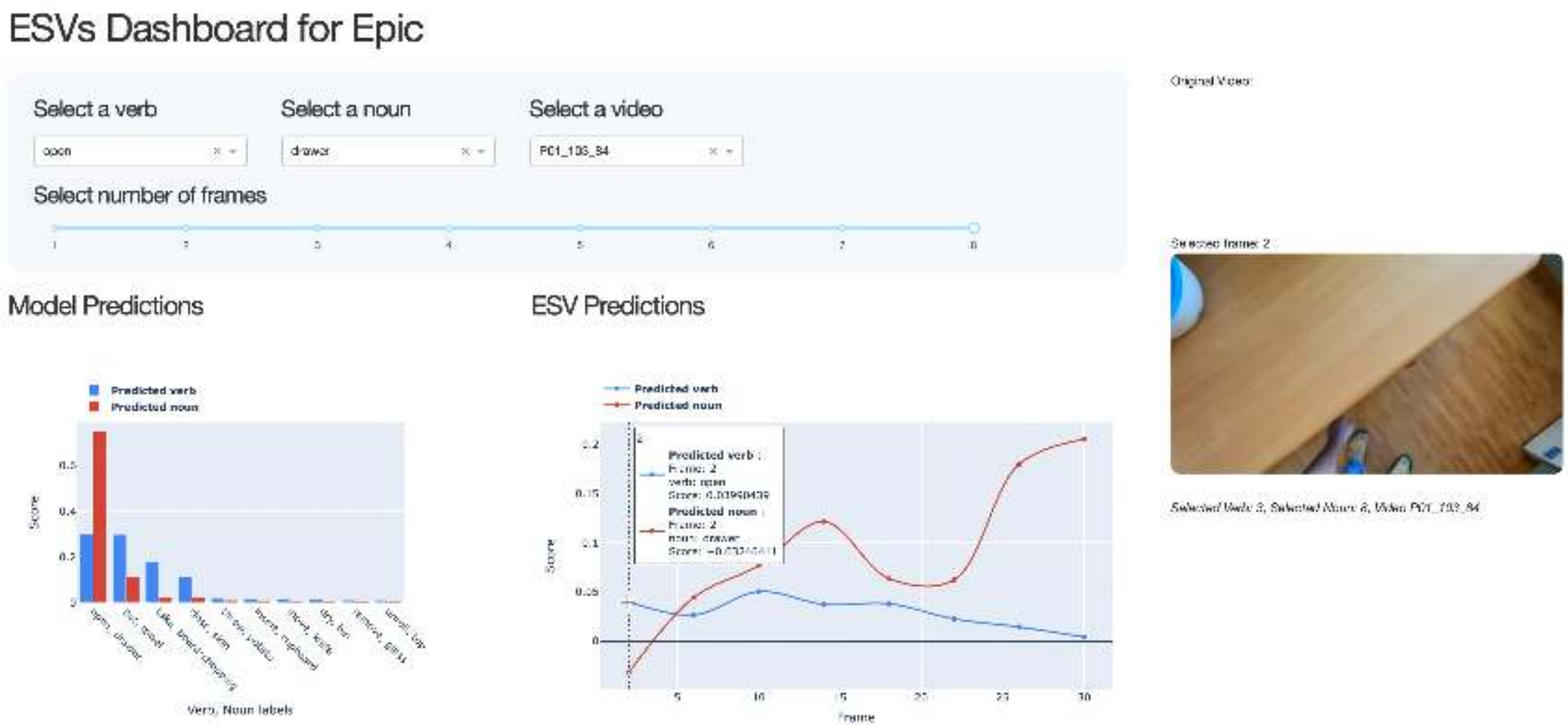
Showing that something is empty

W Price, D Damen (2020). Play Fair: Frame Attribution in Video Models. ACCV

Dashboard

Frame Attributions in Video Models

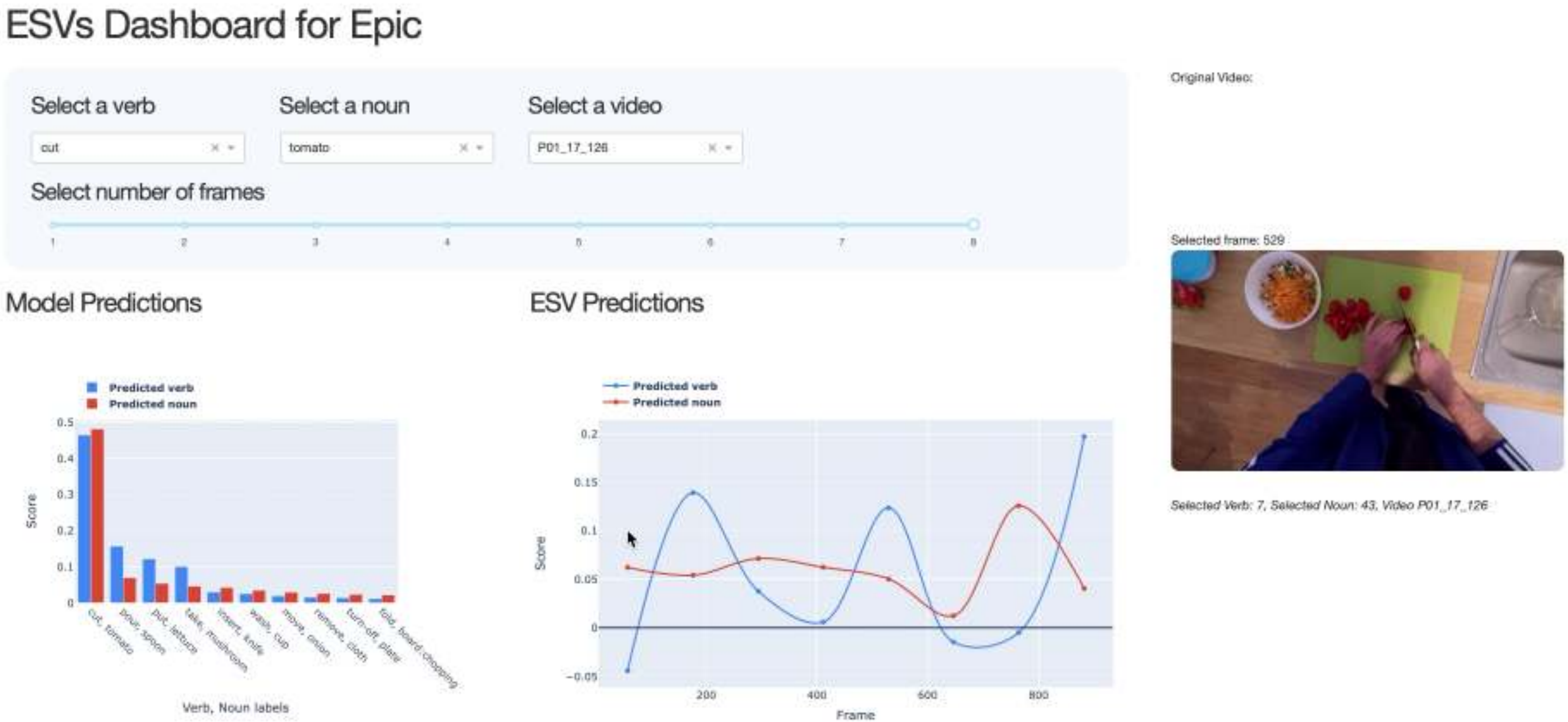
with: Will Price
Tom Stark



W Price, D Damen (2020). Play Fair: Frame Attribution in Video Models. ACCV

Frame Attributions in Video Models

with: Will Price
Tom Stark



W Price, D Damen (2020). Play Fair: Frame Attribution in Video Models. ACCV

In today's tutorial

Motivation and Datasets in Egocentric Video Understanding

Video Understanding: Data and Tasks

Videos are Multimodal

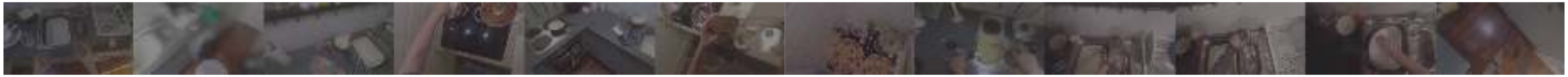
Connected Videos of One's Life

Video Understanding Out of the Frame

Teaser: The Wizard of Oz at the Sphere

Outlook into the Future of Egocentric Vision

Conclusion



Videos are multi-modal

Multi-modal learning…

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

- The magic of audio-visual understanding…

- Object-Object interactions

Multi-modal learning…

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

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

Multi-modal learning…

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

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

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

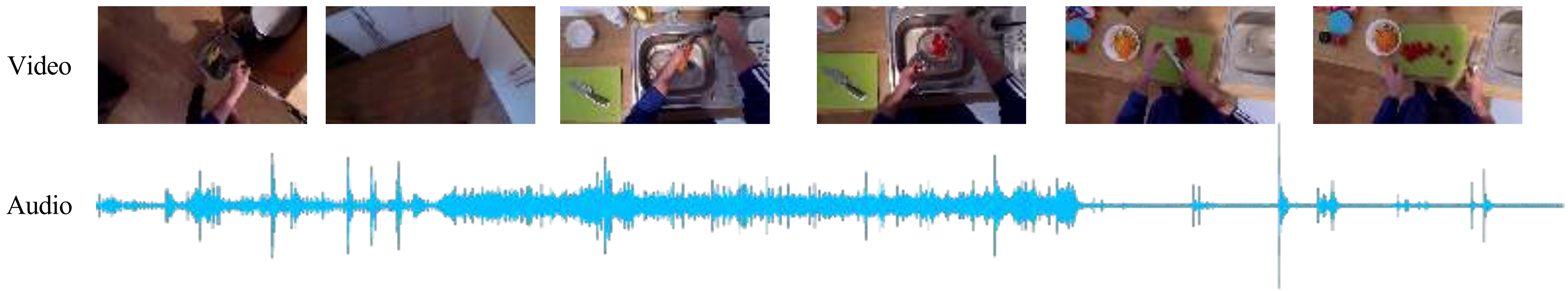
ICASSP 2023
4 - 10 JUNE, RHODES ISLAND, GREECE

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

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

Motivation

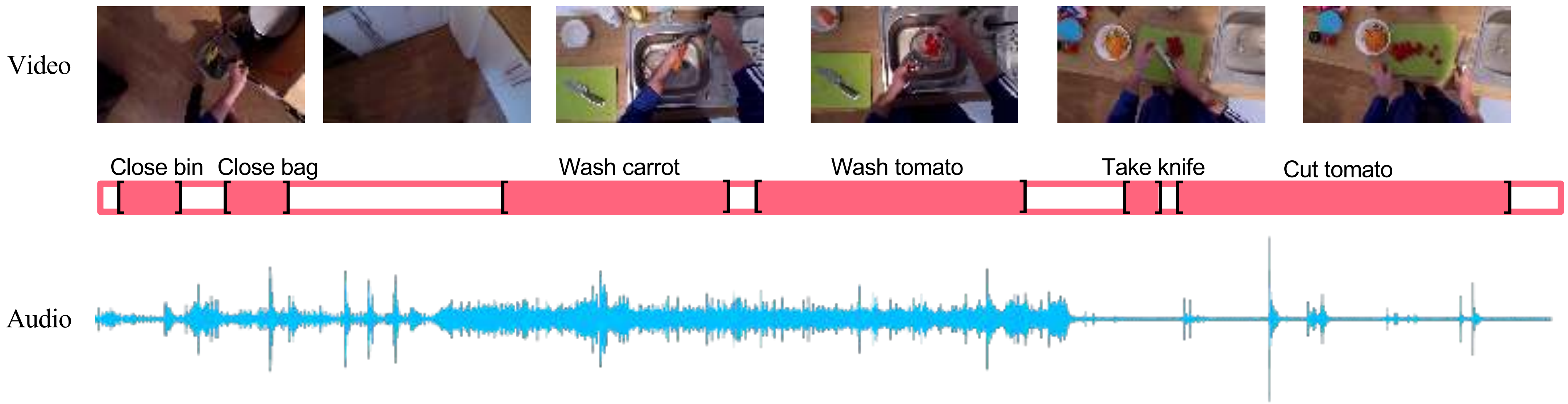
with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman



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

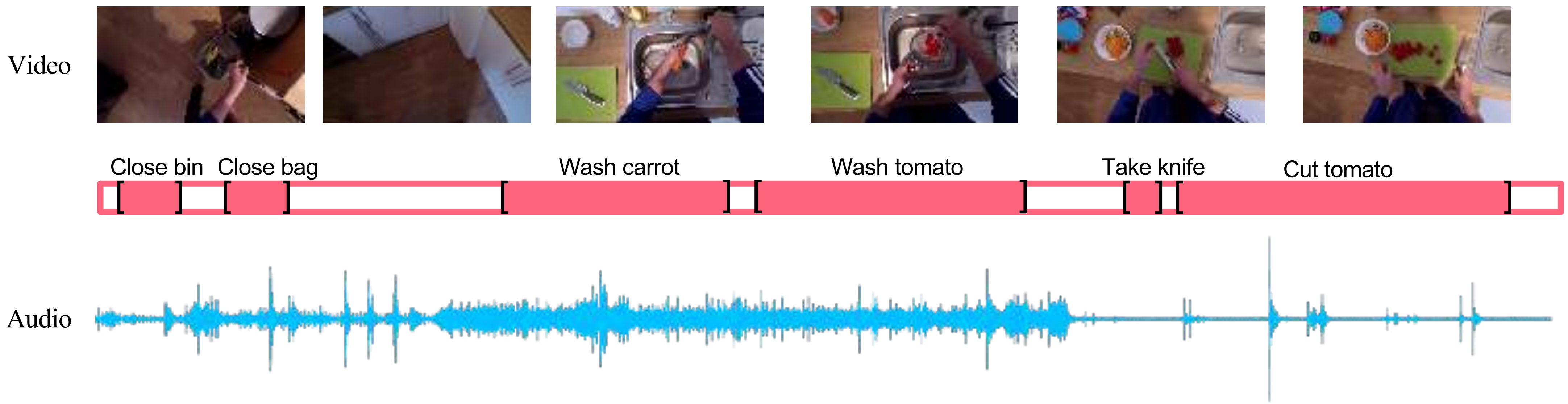
Motivation

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



Motivation

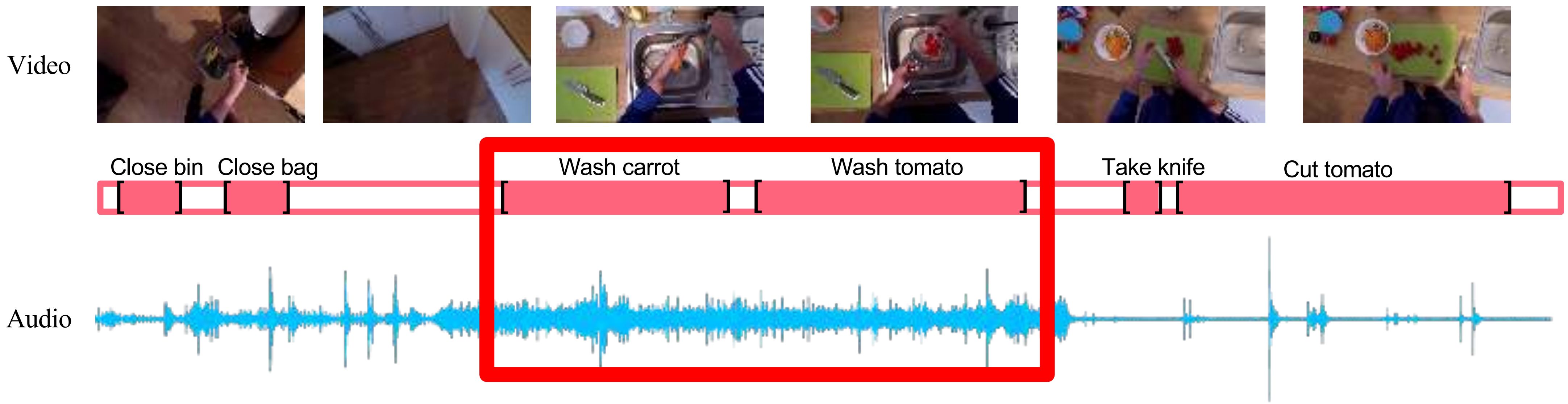
with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman



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

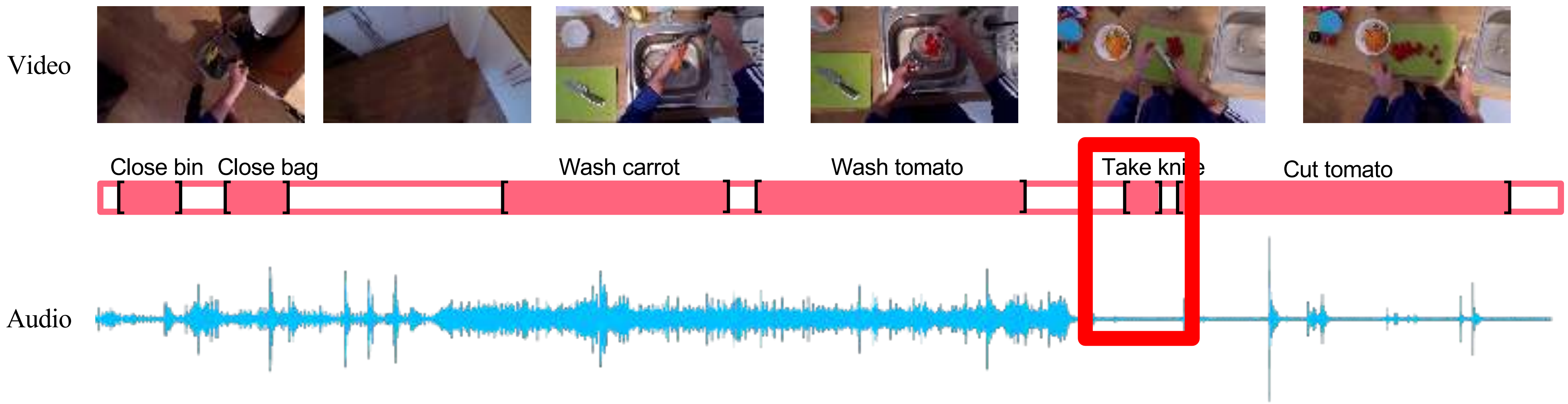


University of BRISTOL

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

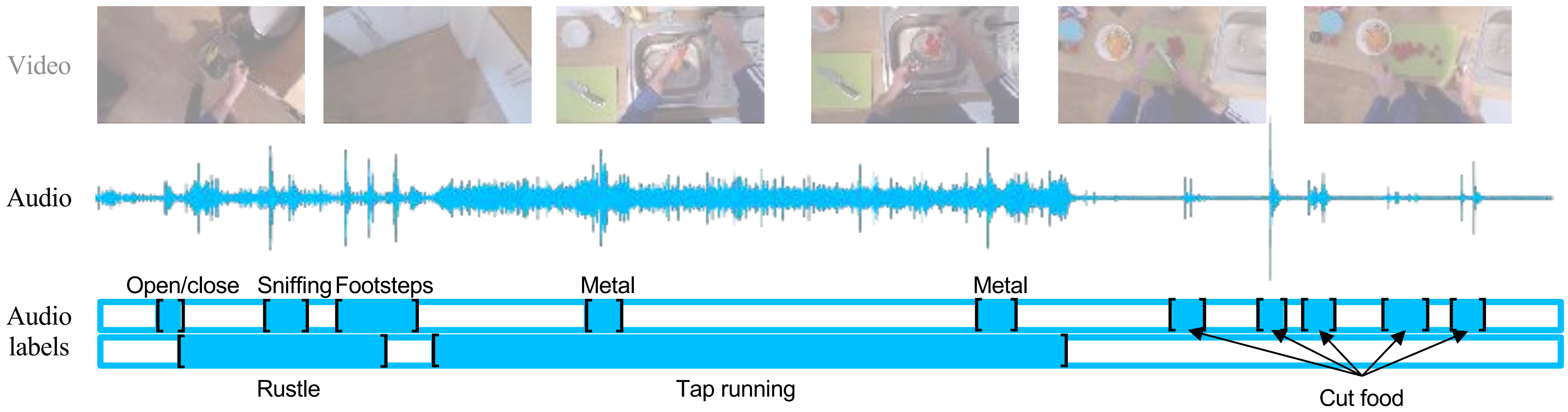
Motivation

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



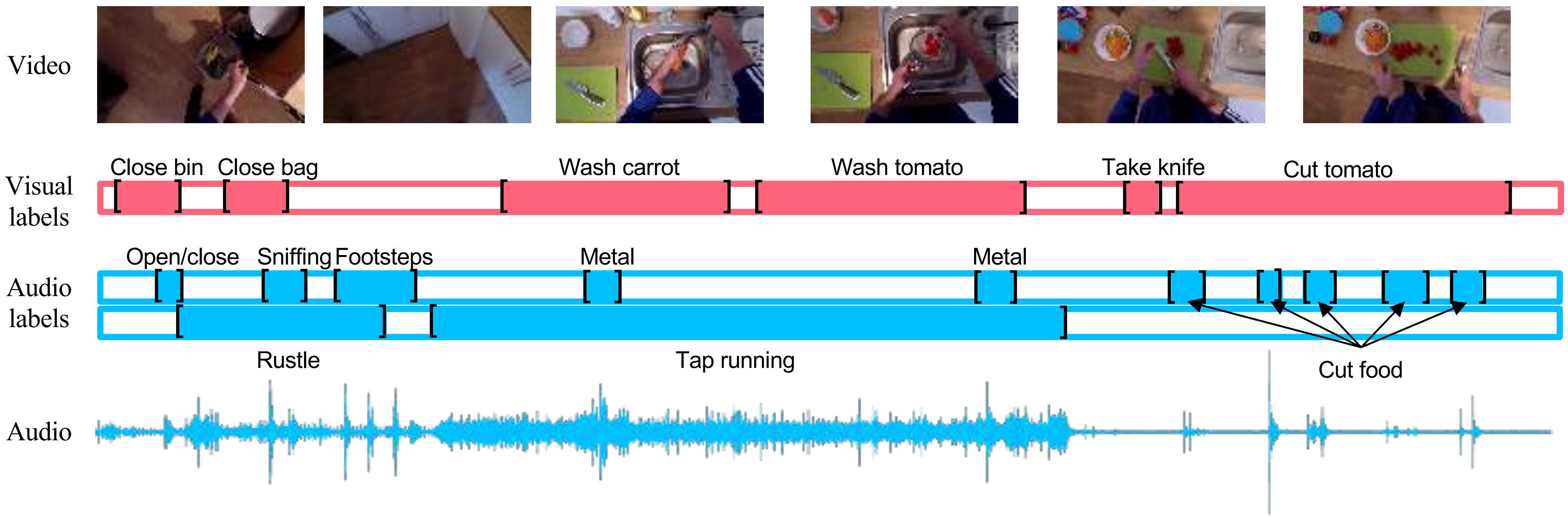
Motivation

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



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

[redacted]

EPIC-SOUNDS

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

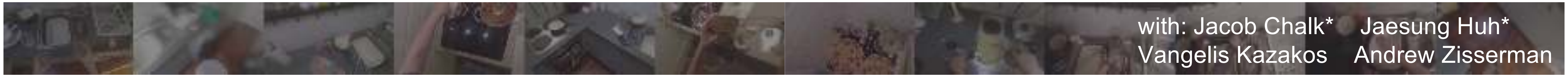
EPIC-KITCHENS VIDEOS
100 hours
45 kitchens

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

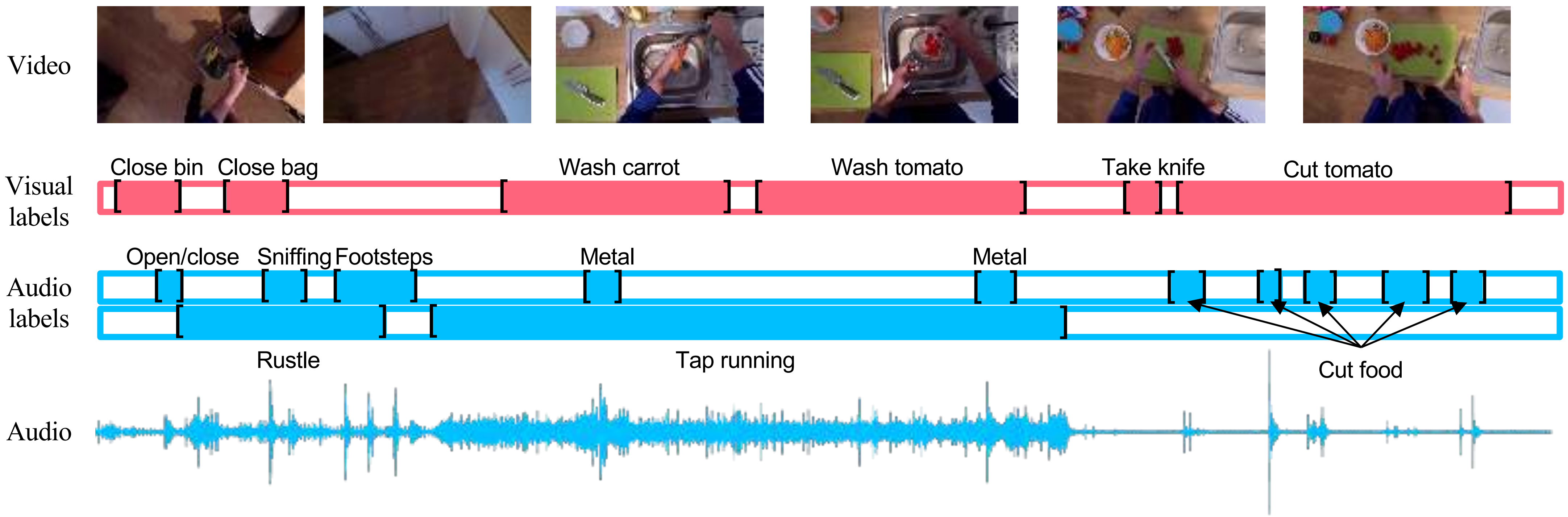
Jacob Chalk*, Jaesung Huh*, Evangelos Kazakos, Andrew Zisserman, Dima Damen

* : Equal contribution

University of BRISTOL
VGG University of Oxford
CZECH TECHNICAL UNIVERSITY IN PRAGUE

Multi-Modal Long-Form Dataset

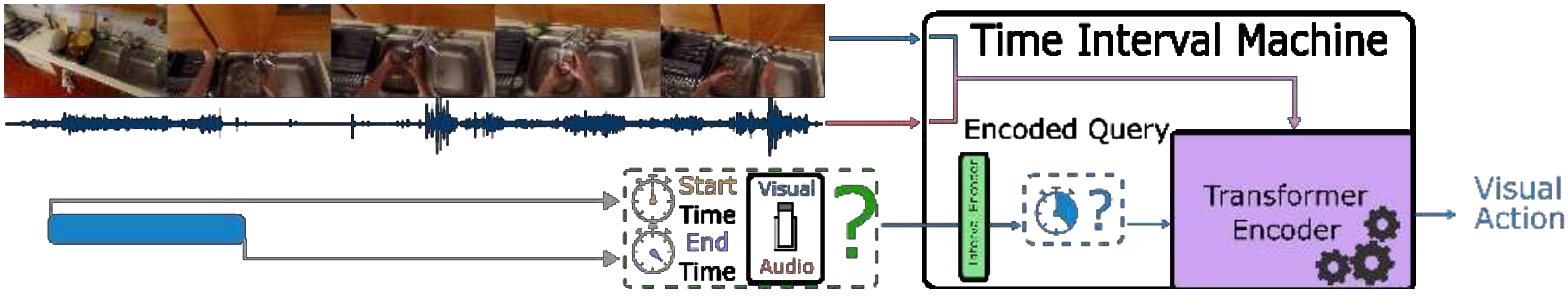
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



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

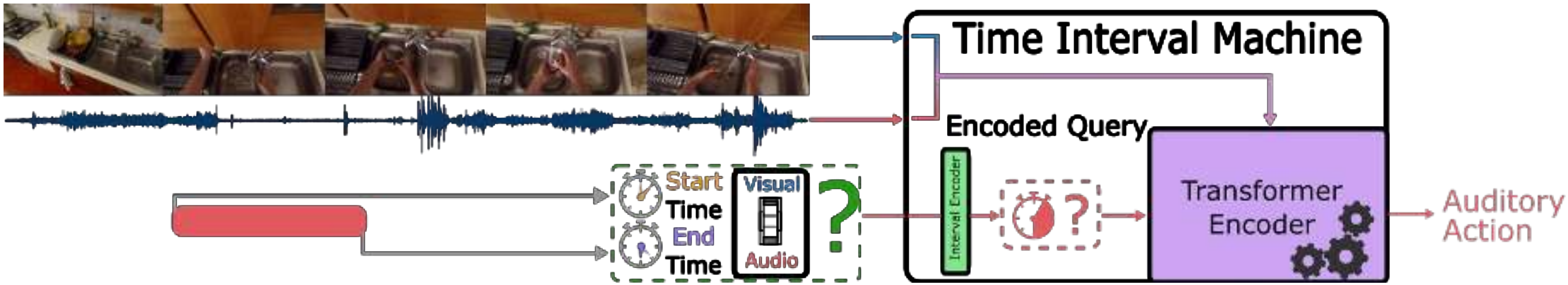
TIM: A Time-Interval Audio-Visual Machine

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



TIM: A Time-Interval Audio-Visual Machine

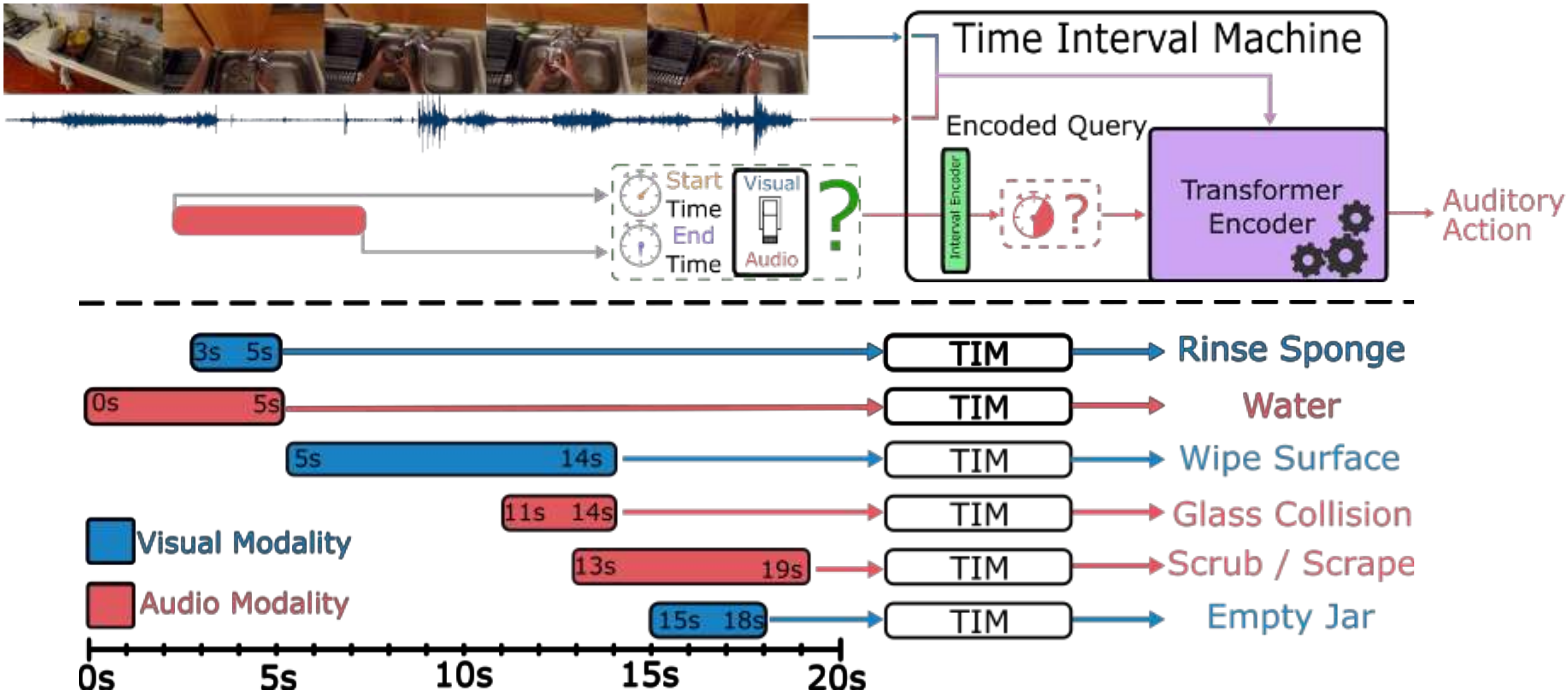
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



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

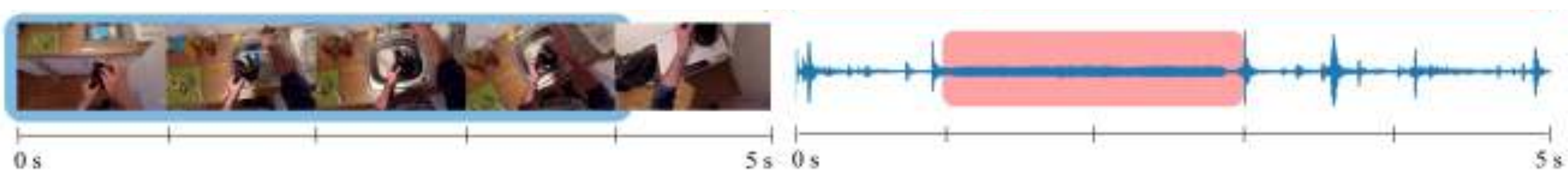


University of BRISTOL

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

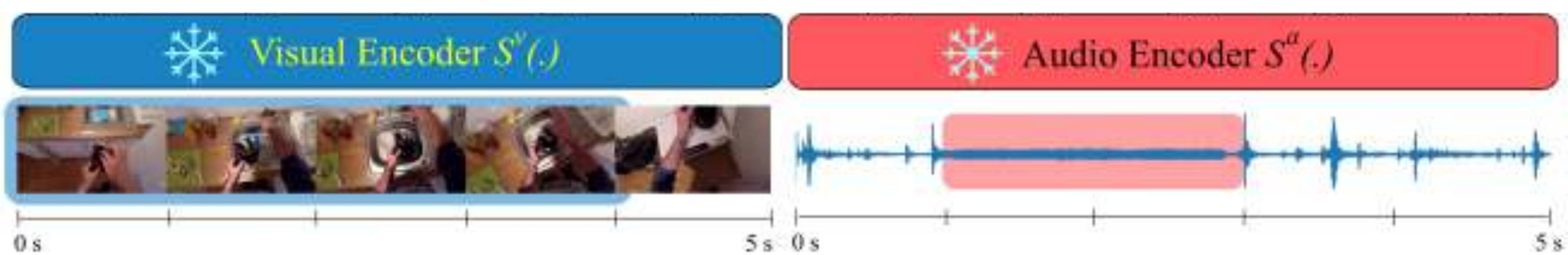
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



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

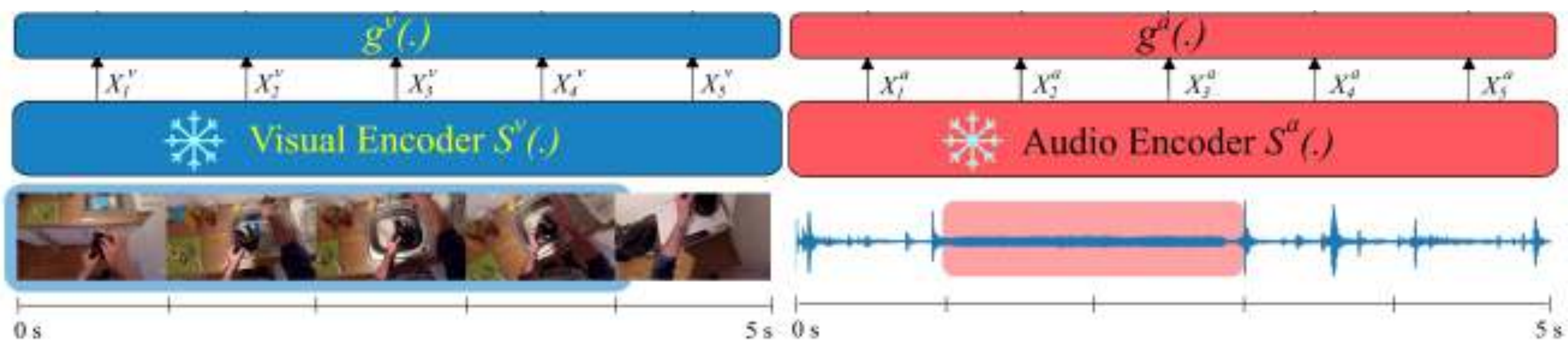
TIM: A Time-Interval Audio-Visual Machine

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



TIM: A Time-Interval Audio-Visual Machine

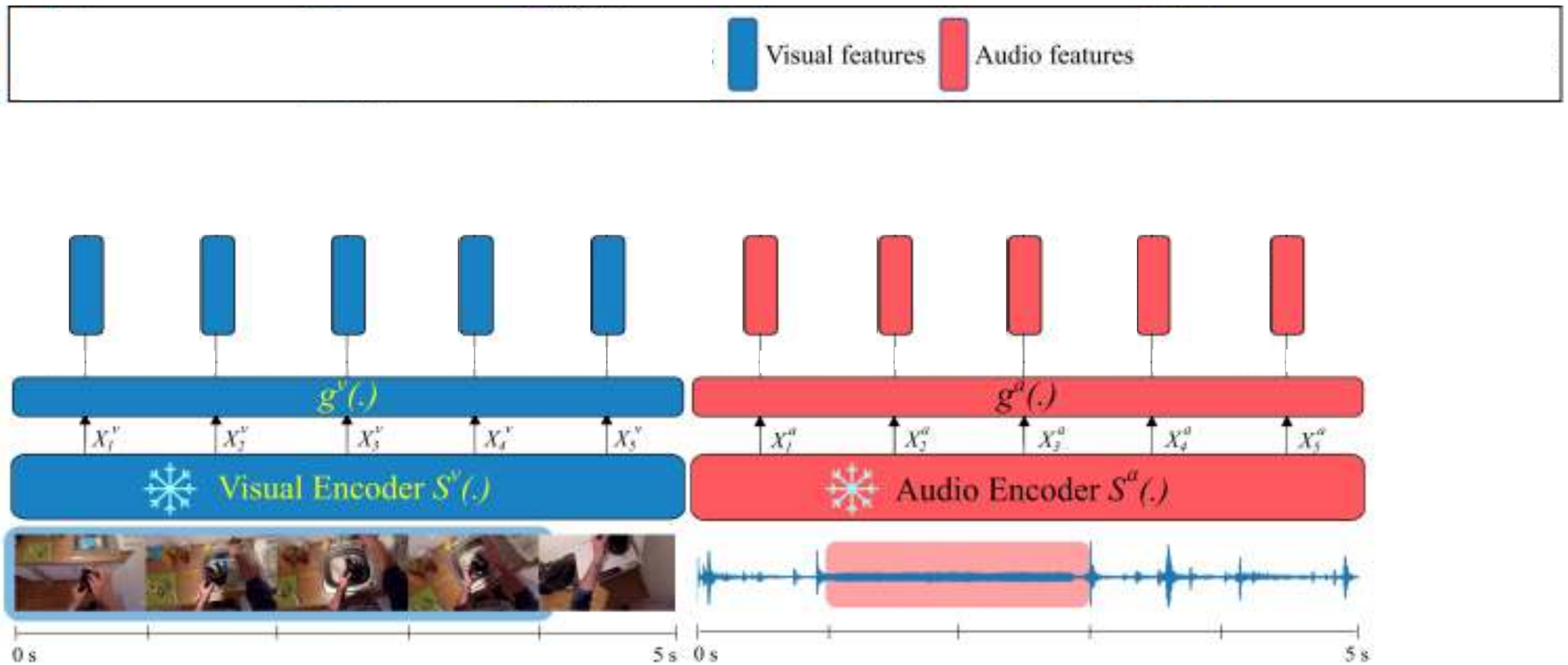
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



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

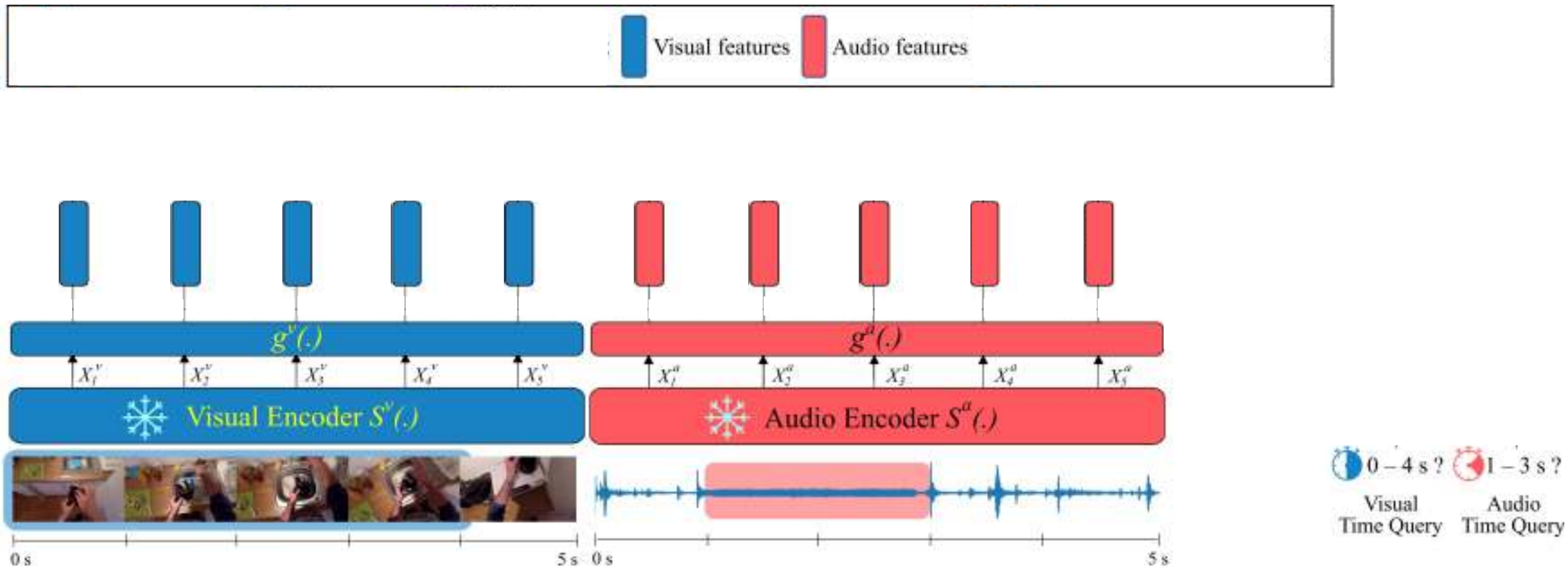
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



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

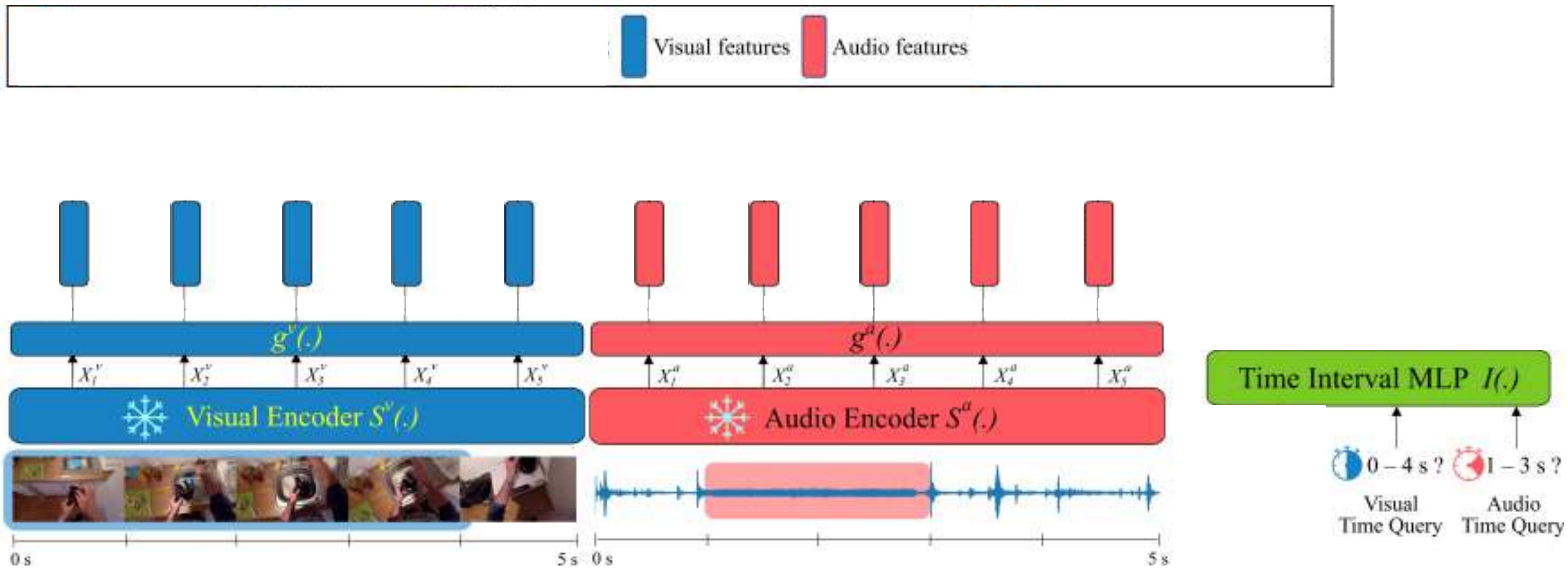
TIM: A Time-Interval Audio-Visual Machine

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



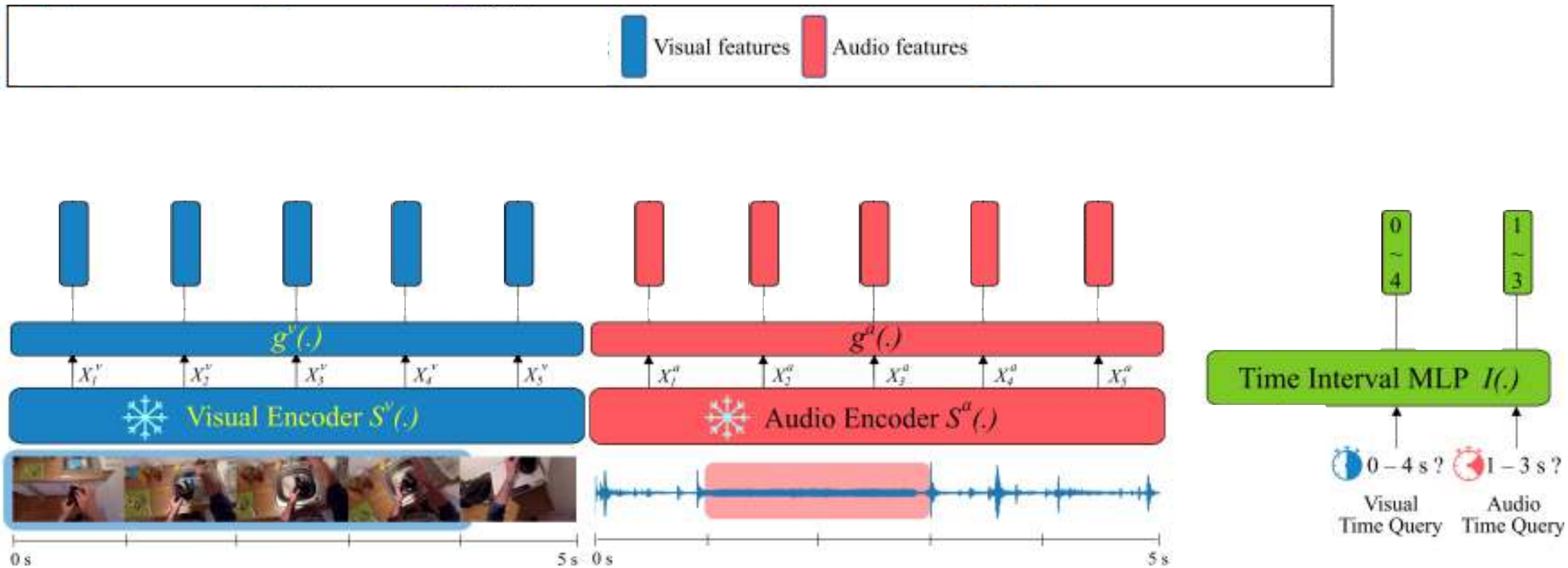
TIM: A Time-Interval Audio-Visual Machine

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



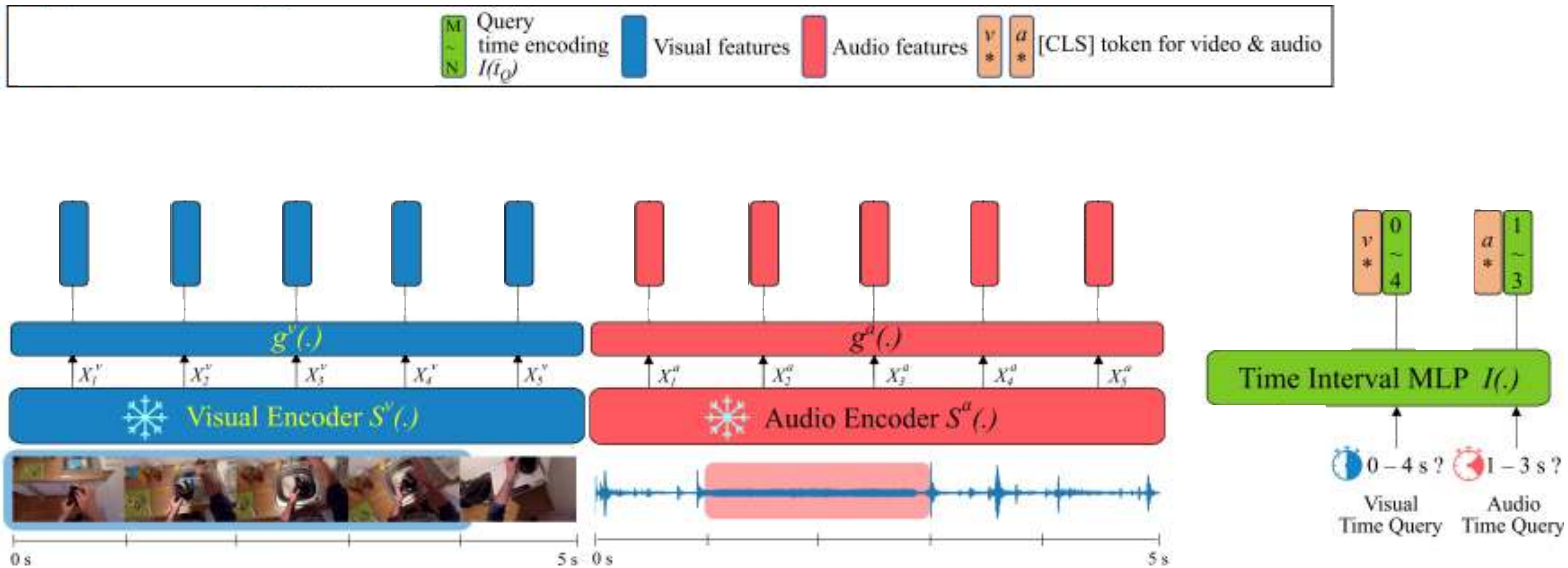
TIM: A Time-Interval Audio-Visual Machine

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



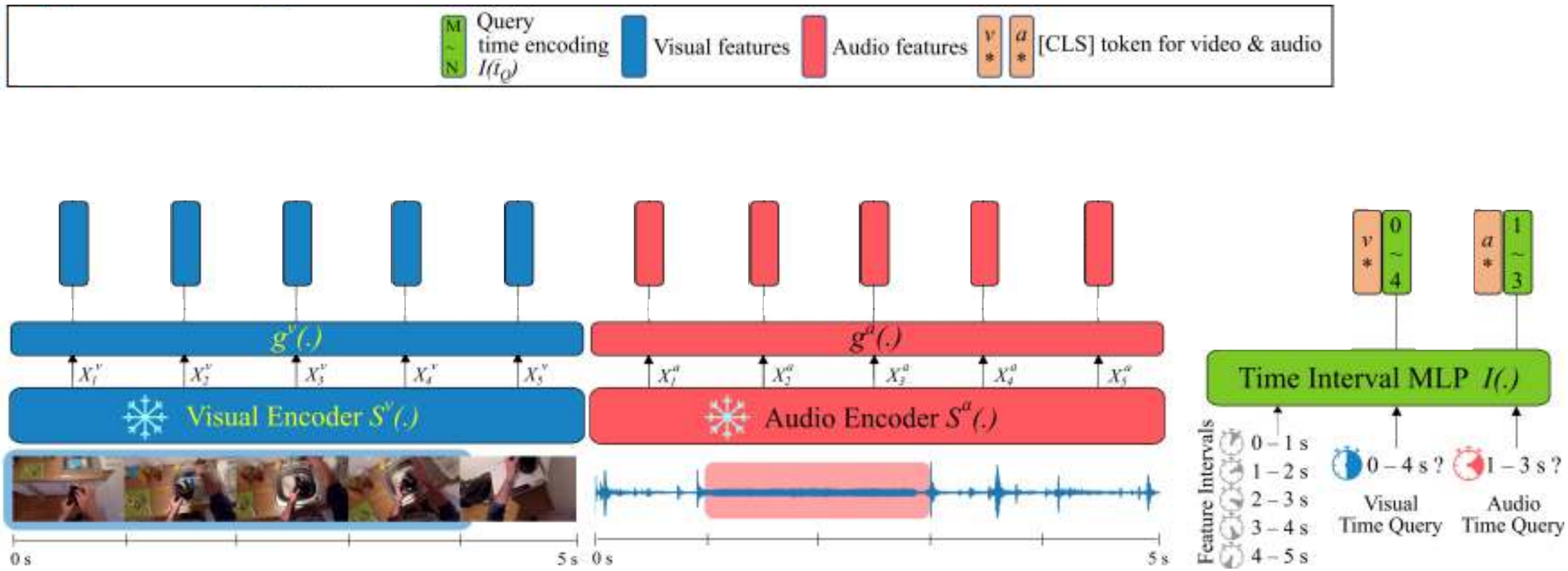
TIM: A Time-Interval Audio-Visual Machine

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



TIM: A Time-Interval Audio-Visual Machine

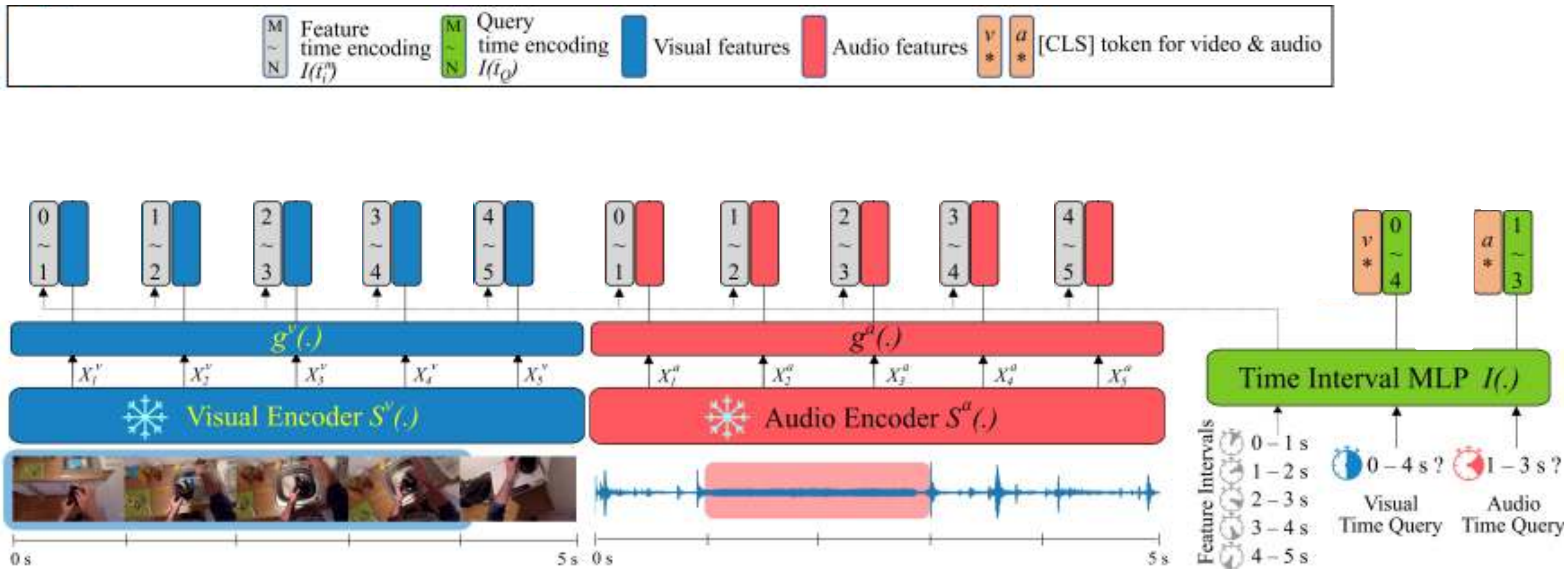
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



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

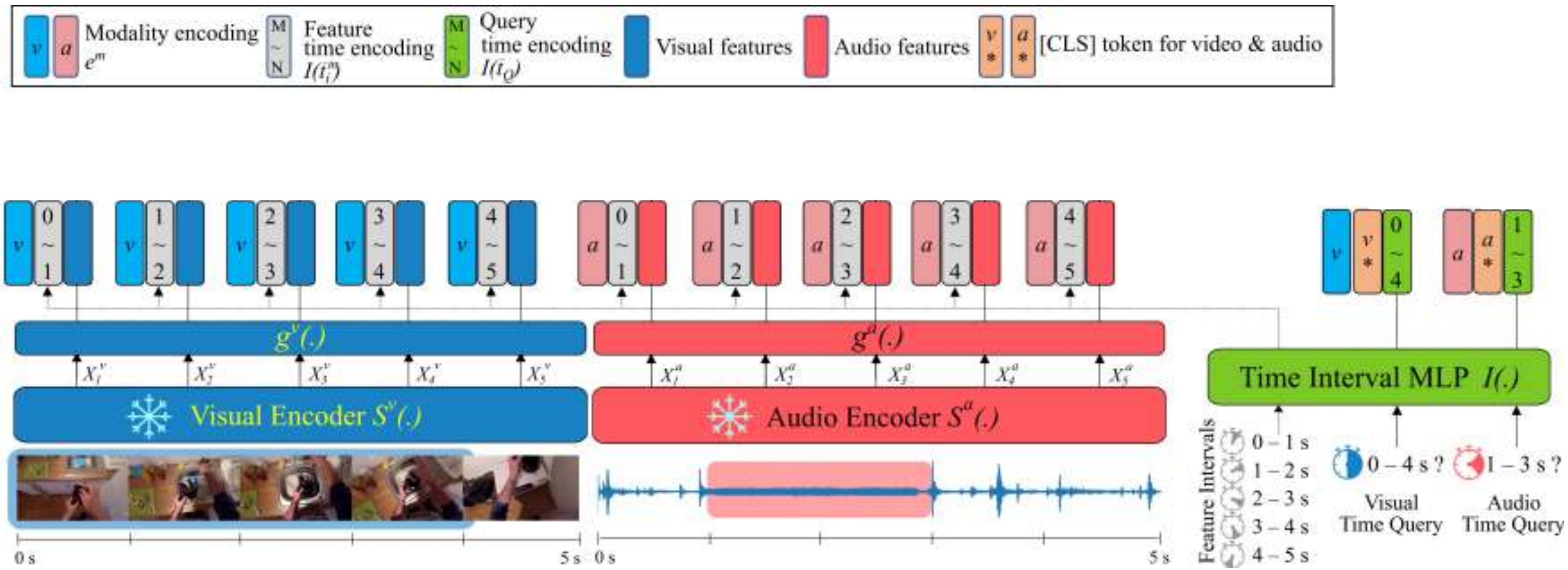
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



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

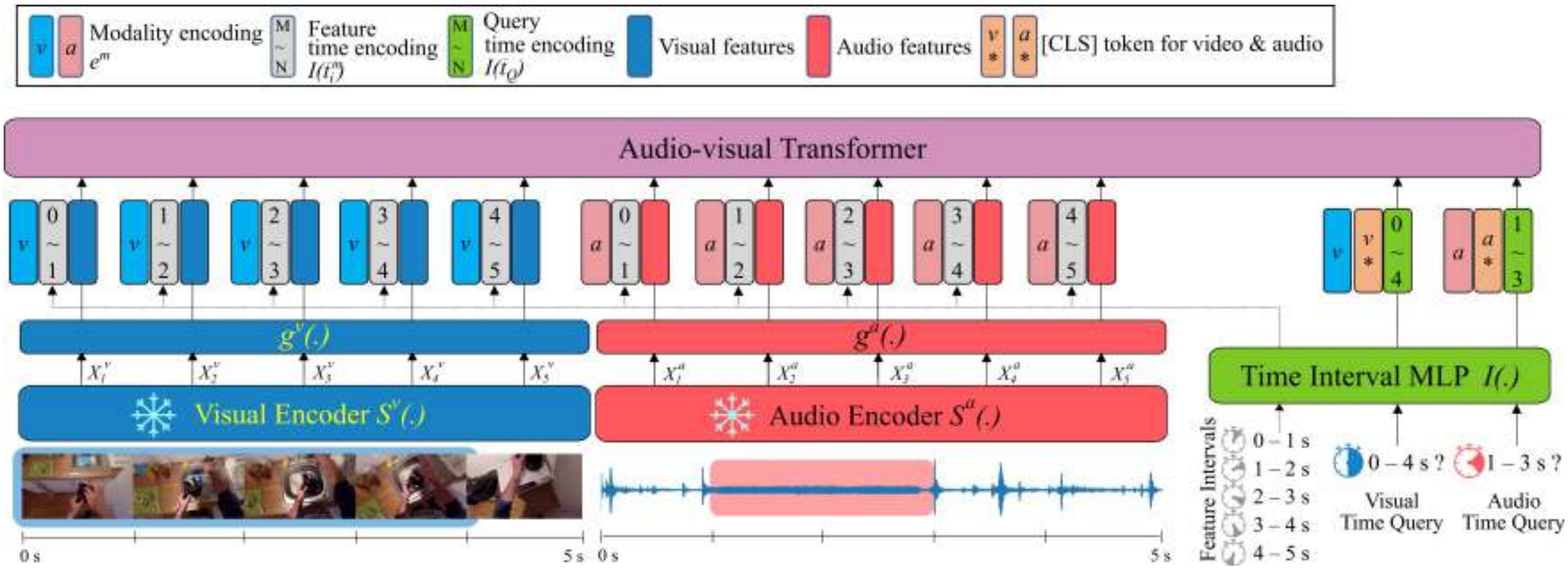
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



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

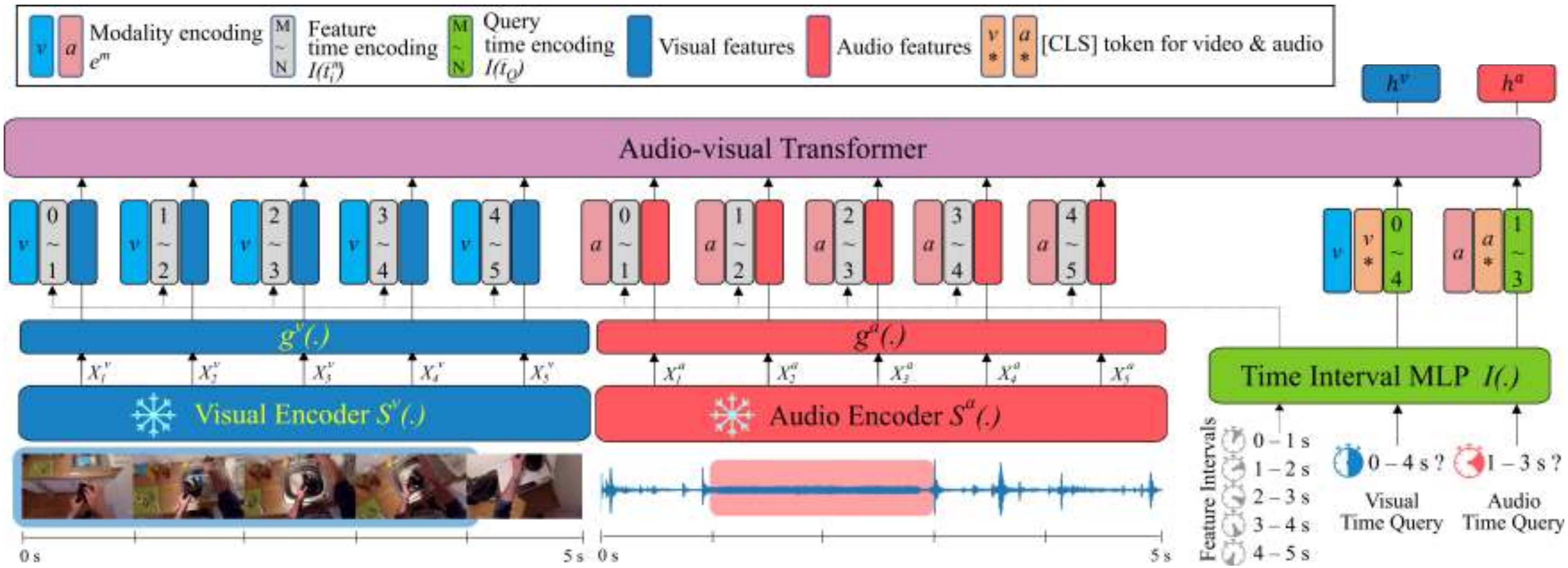


University of BRISTOL

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

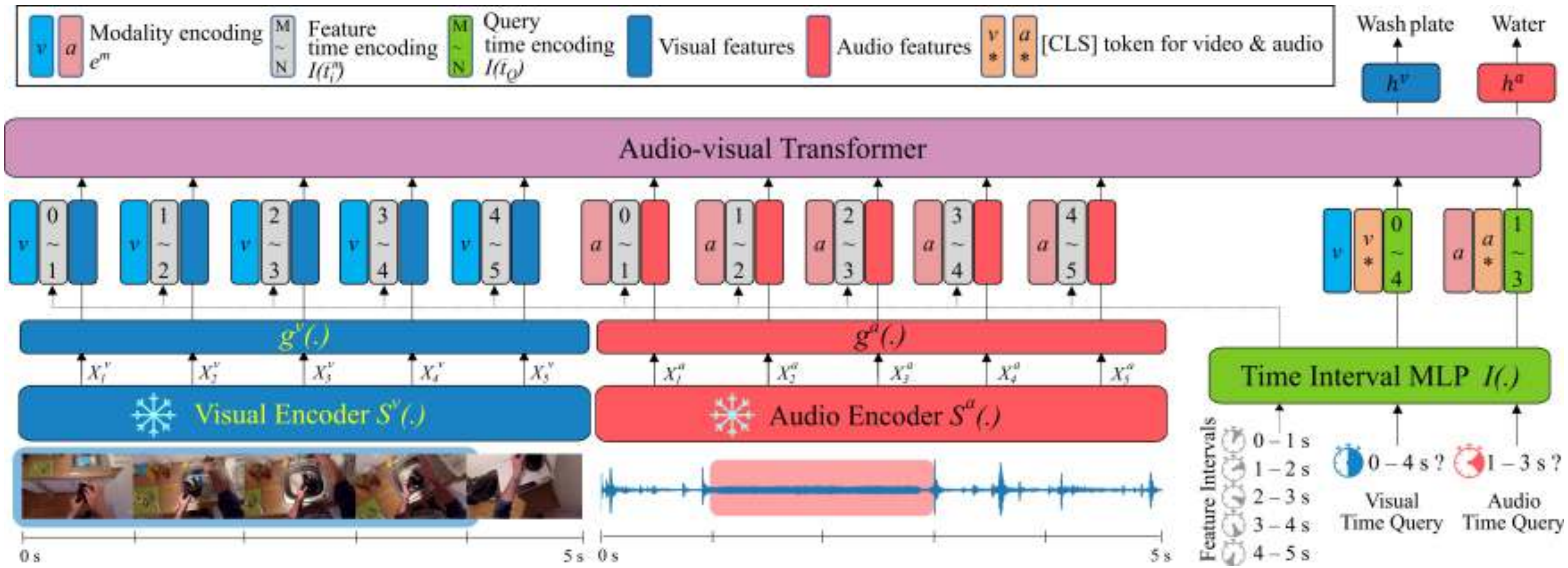
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



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

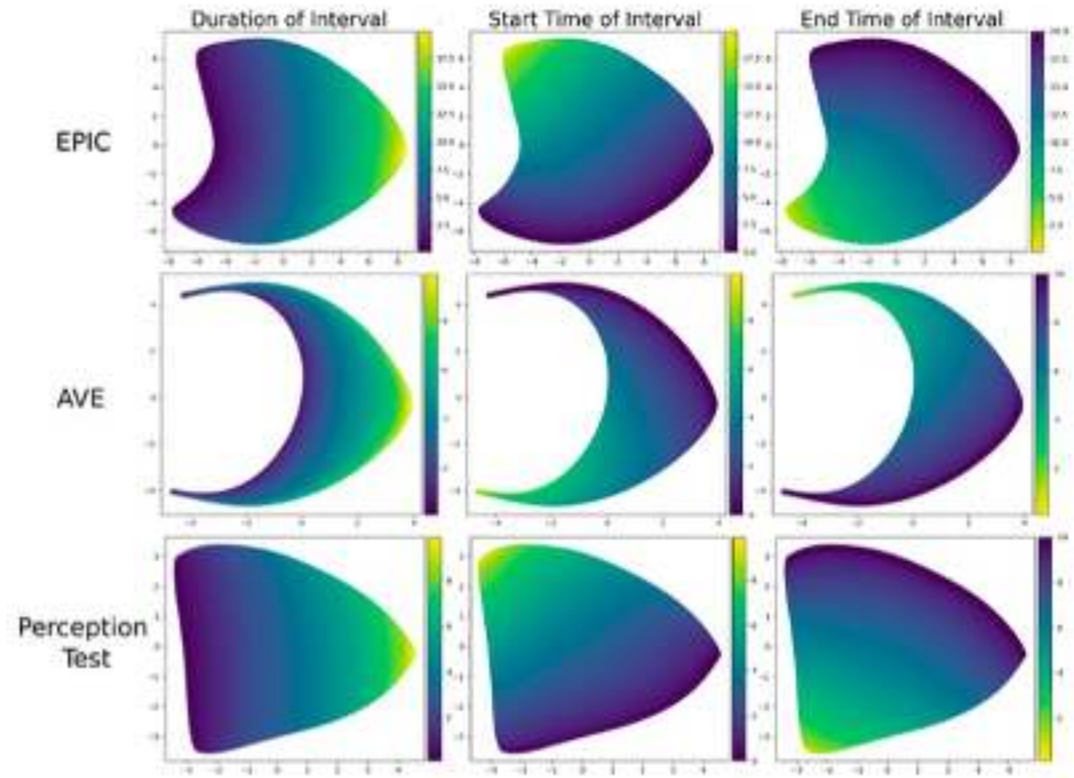
Model	xp	LLM	Verb	Noun	Action
Visual-only models					
MFormer-HR [37]	336p	✗	67.0	58.5	44.5
MoViNet-A6 [27]	320p	✗	72.2	57.3	47.7
MeMViT [55]	224p	✗	71.4	60.3	48.4
Omnivore [14]	224p	✗	69.5	61.7	49.9
MTV [59]	280p	✗	69.9	63.9	50.5
LaViLa (TSF-L) [63]	224p	✓	72.0	62.9	51.0
AVION (ViT-L) [62]	224p	✓	73.0	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

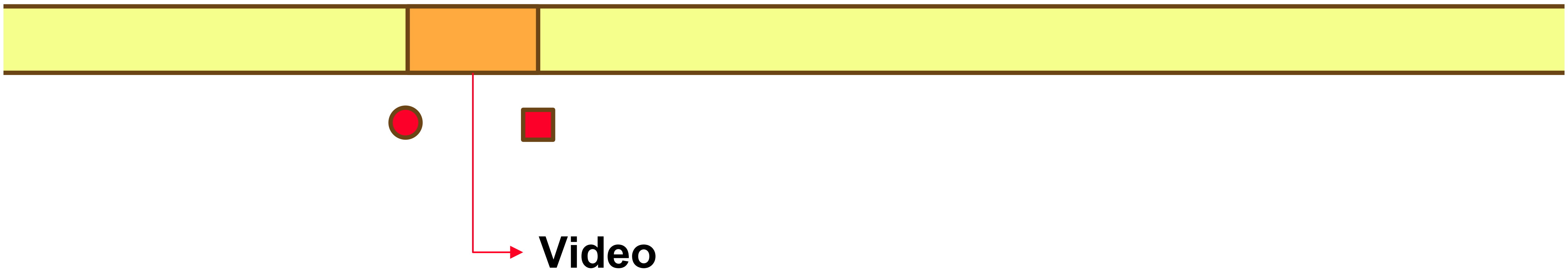
Video Understanding Out of the Frame

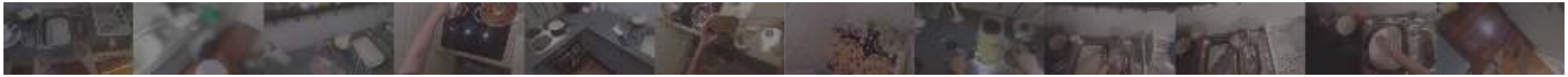
Teaser: The Wizard of Oz at the Sphere

Outlook into the Future of Egocentric Vision

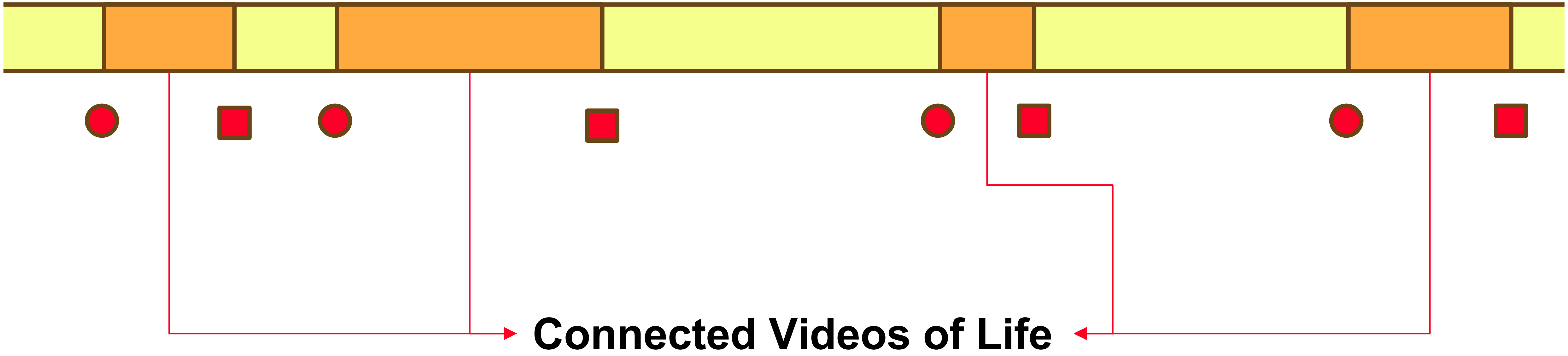
Conclusion

Current Video Understanding…

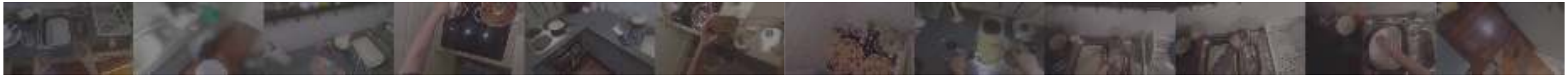




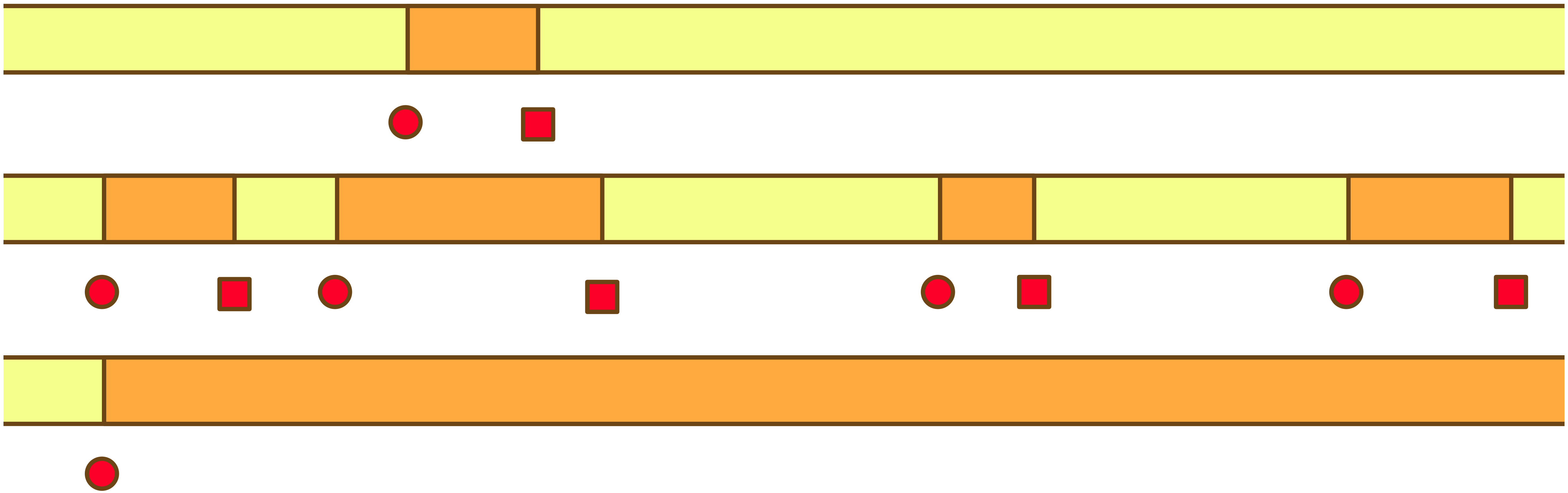
Upcoming Video Understanding…

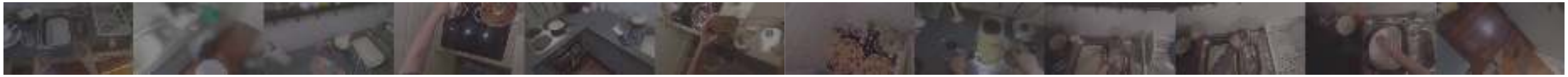


Eventually…

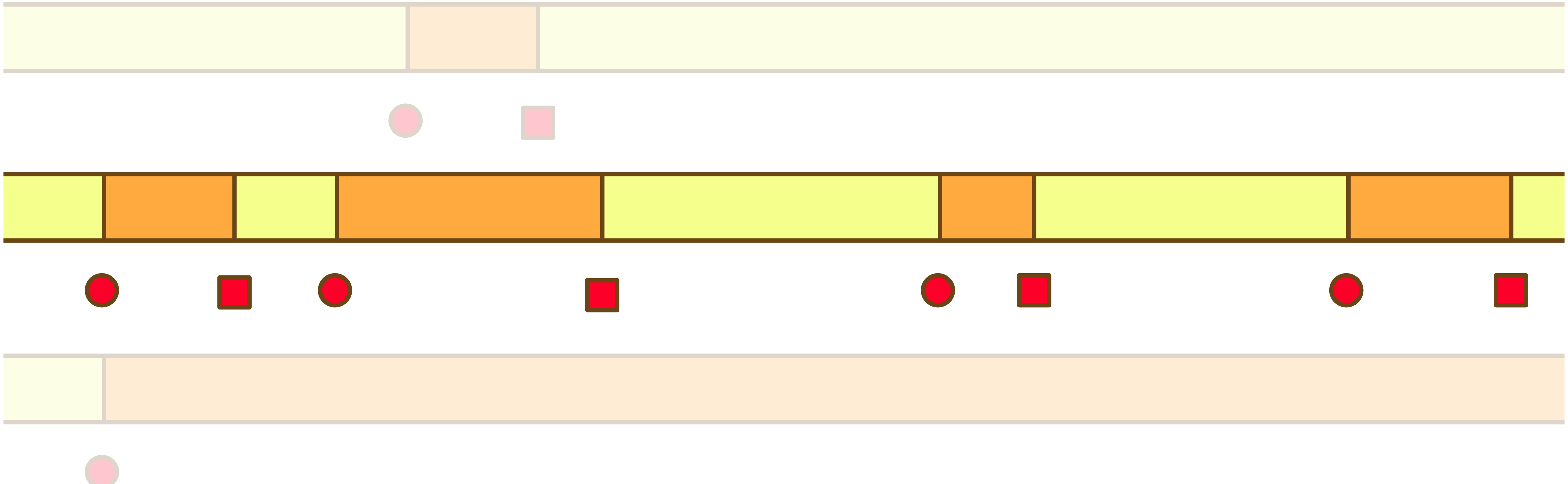


Egocentric Video Understanding





Egocentric Video Understanding



Dima Damen
ICVSS2025

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

Toby Perrett, Tengda Han, Dima Damen, Andrew Zisserman

Best Paper Award
ACCV HANOI VIETNAM 2024 DEC 8-12

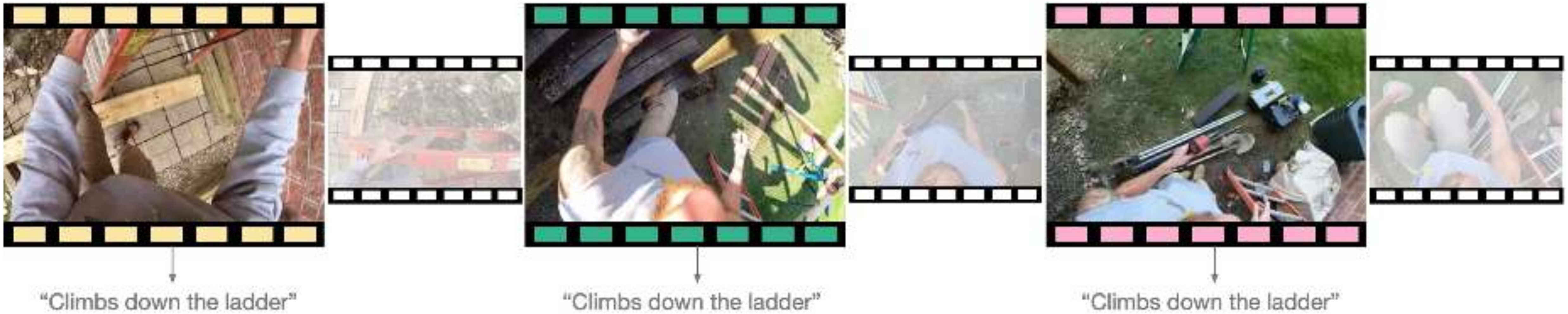
University of BRISTOL

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

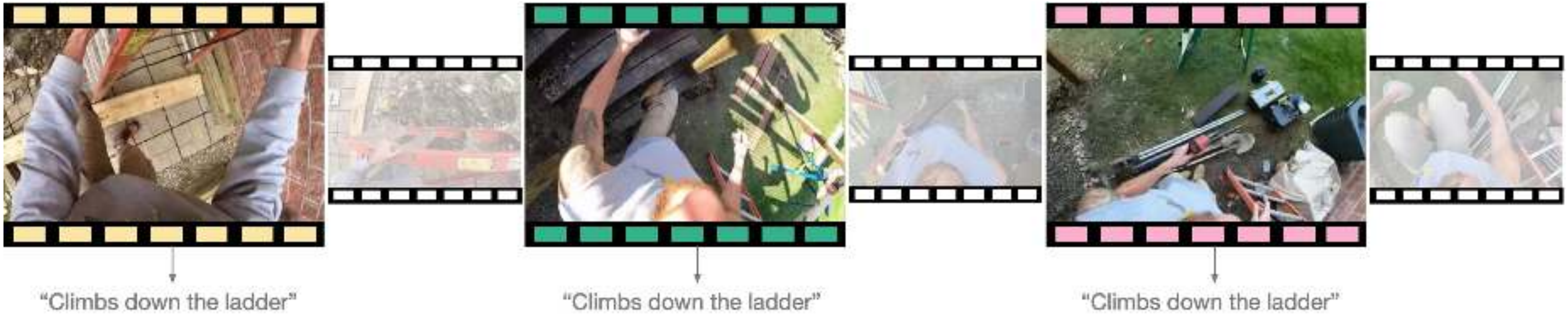
with: Toby Perrett
Tengda Han
Andrew Zisserman

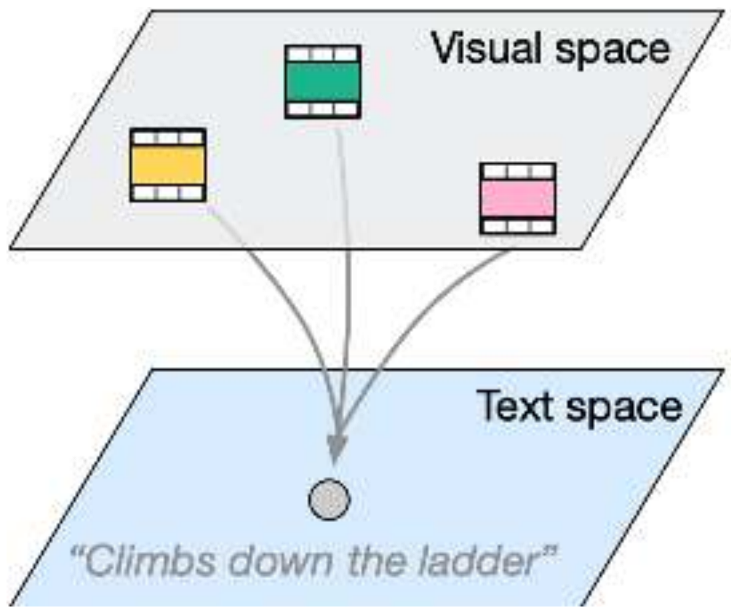
Goal:
Generate a unique caption for every clip in a set

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

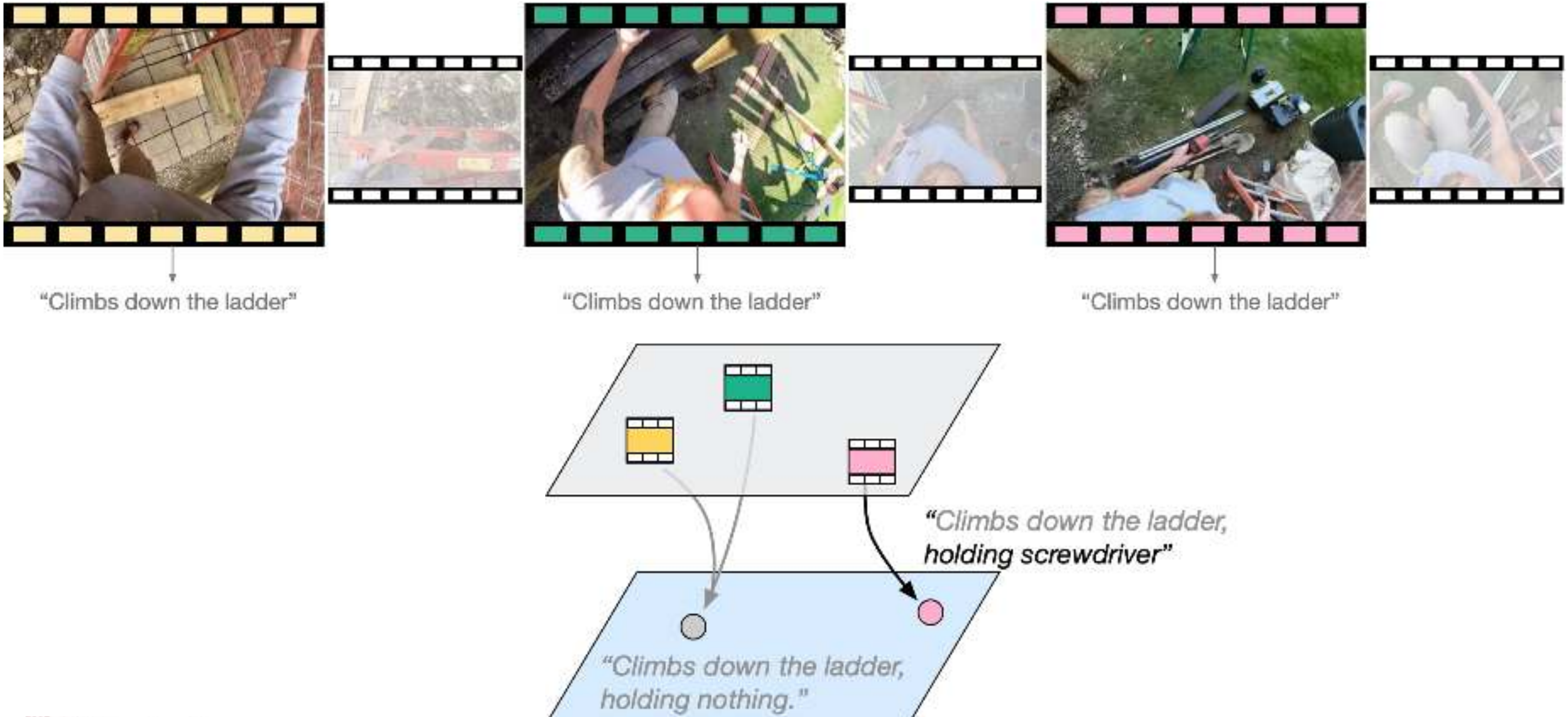




Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Unique Video Captioning

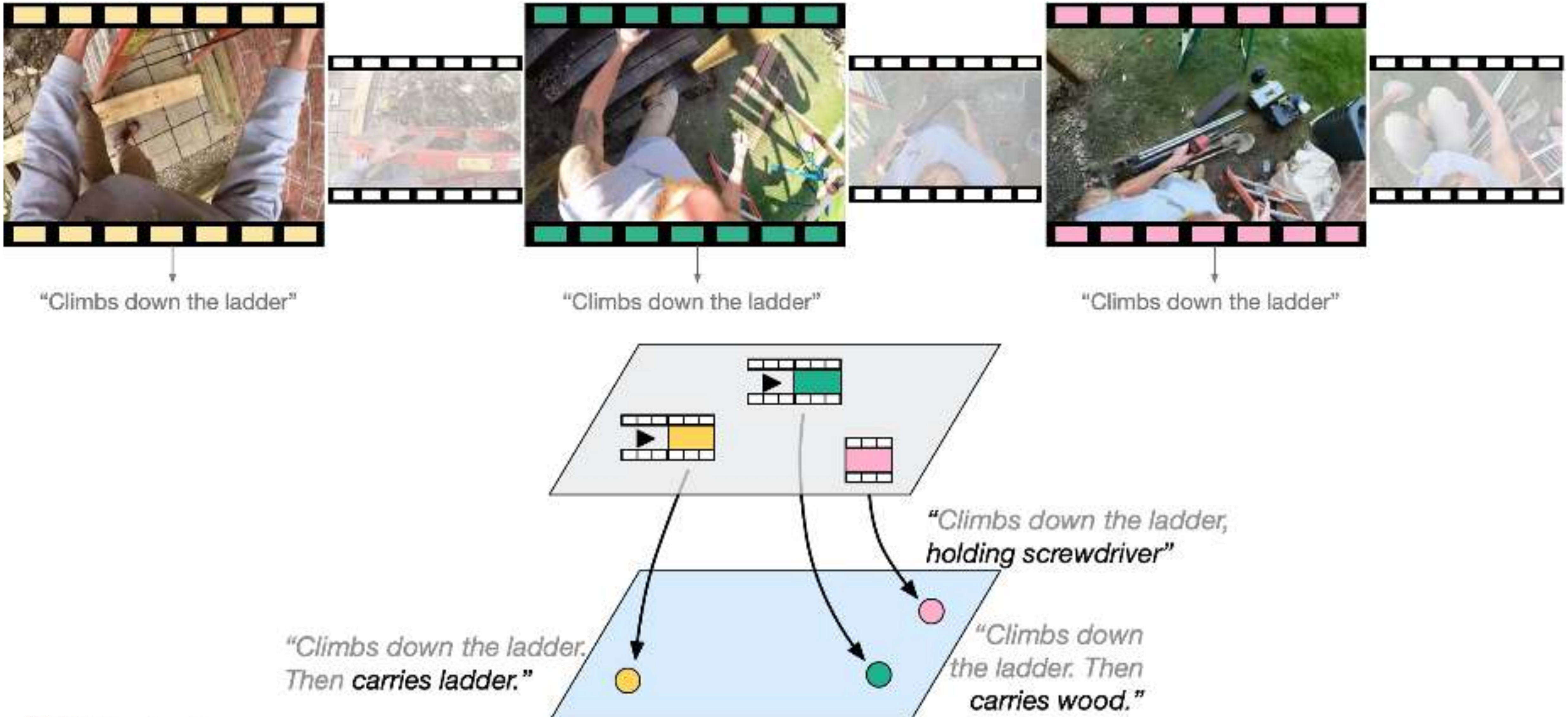
with: Toby Perrett
Tengda Han
Andrew Zisserman



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

Unique Video Captioning

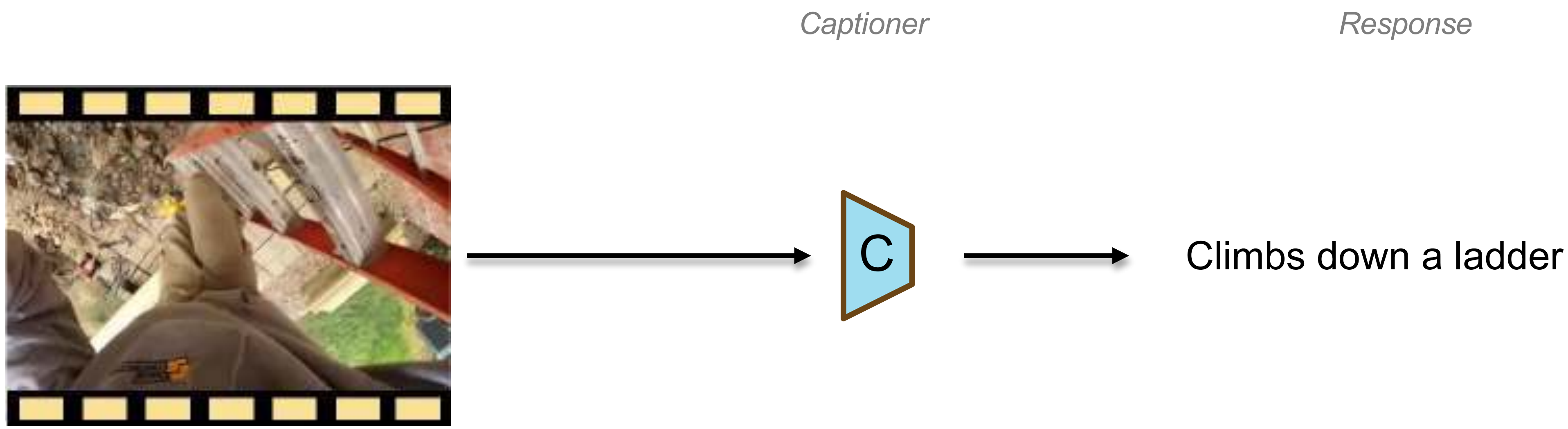
with: Toby Perrett
Tengda Han
Andrew Zisserman



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

Discriminative prompt *Captioner* *Response*

The person walks around

C

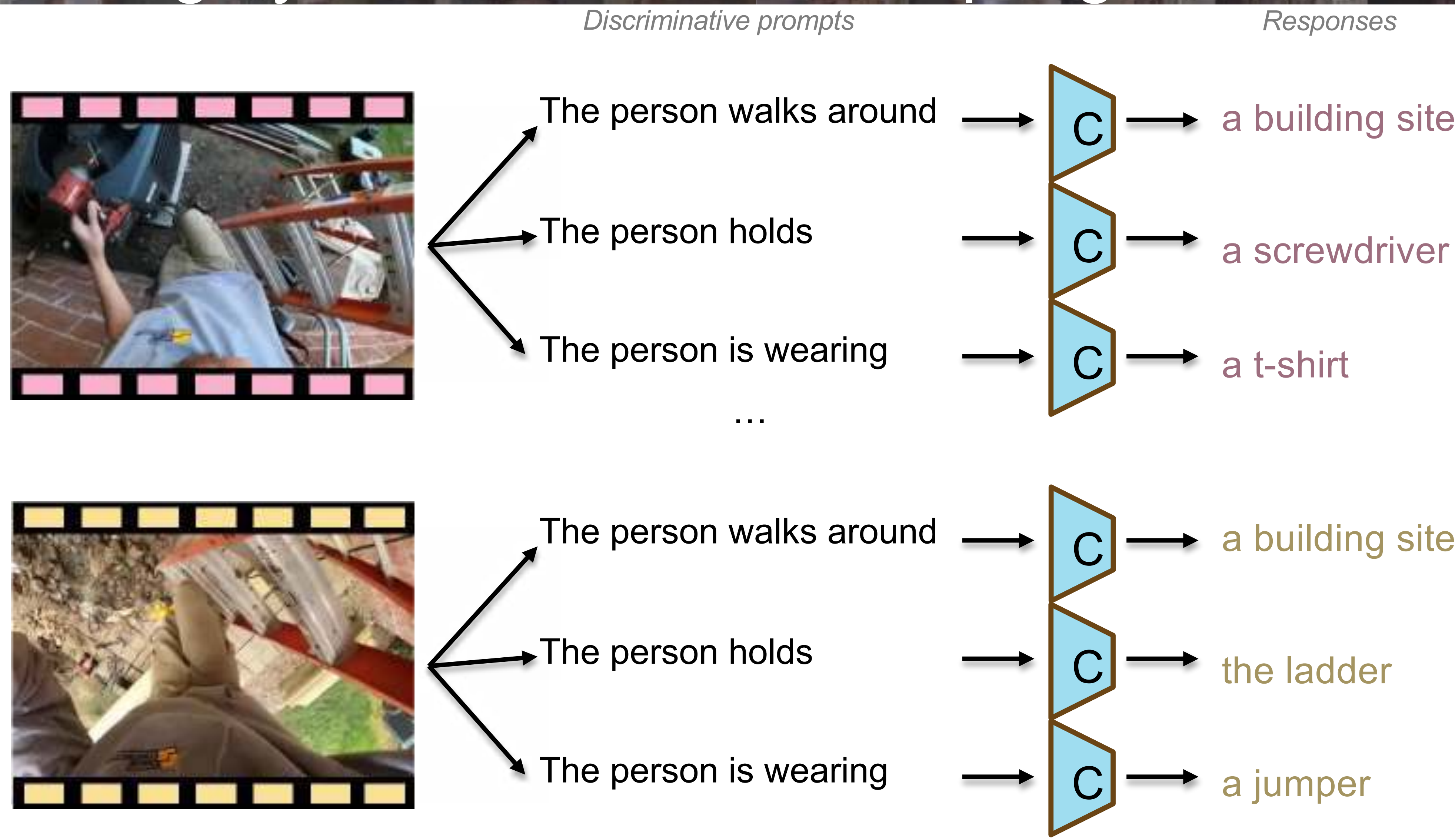
Climbs down a ladder and walks around

a building site.

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

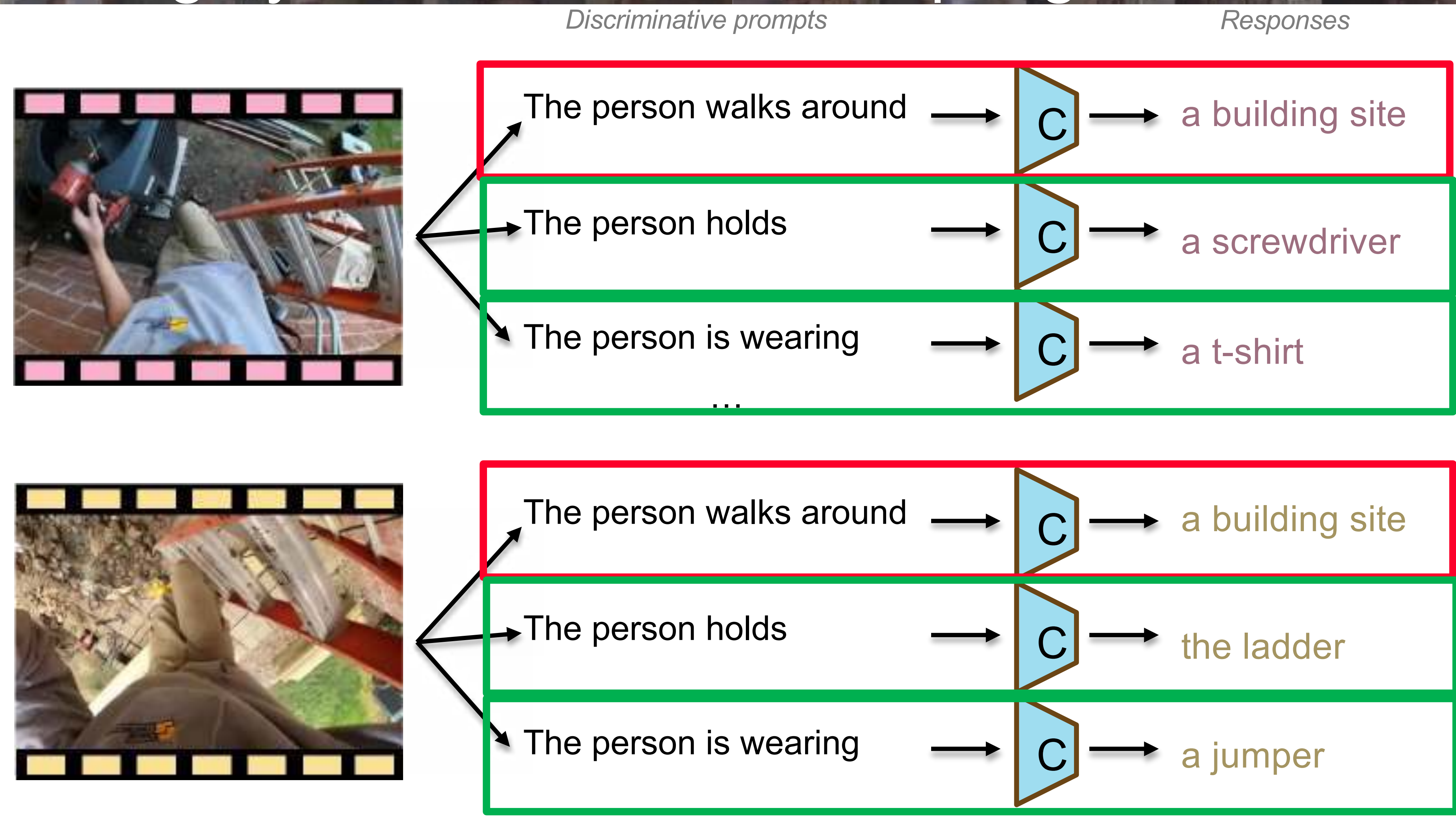
with: Toby Perrett
Tengda Han
Andrew Zisserman



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



University of Bristol

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

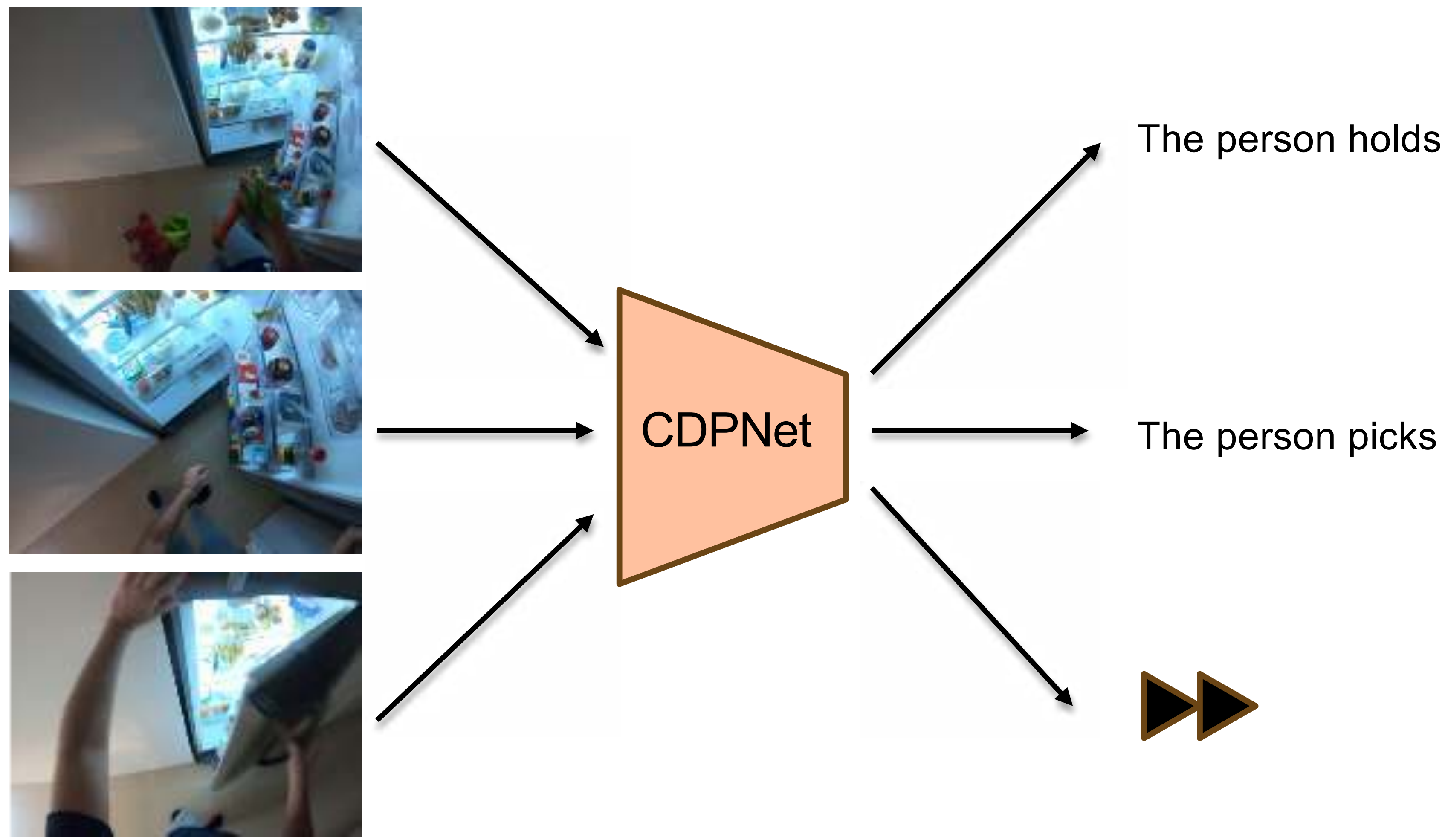
We propose to…
consider clips jointly
use a bank of discriminative prompts

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

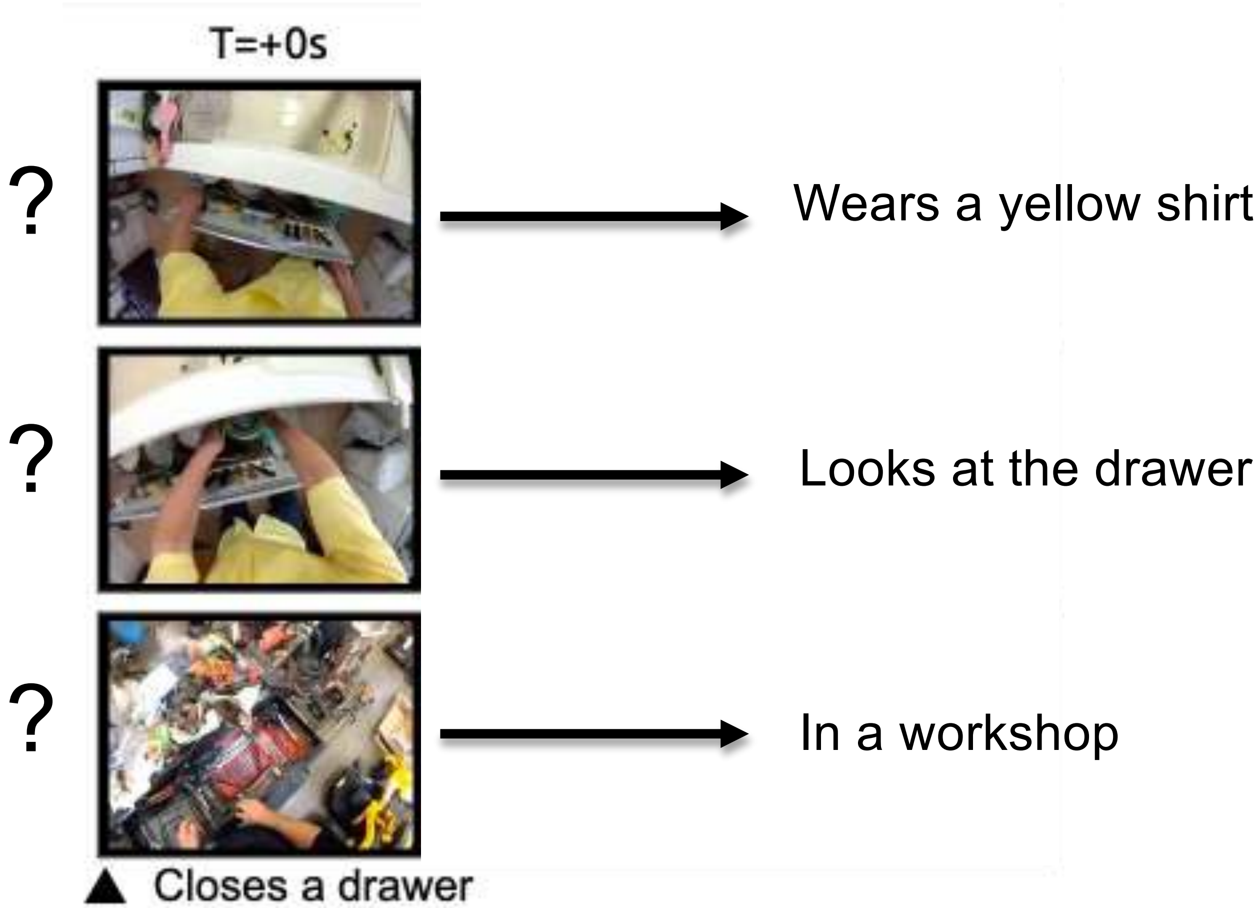


Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Benchmarks

with: Toby Perrett
Tengda Han
Andrew Zisserman

Task: Caption every clip, then evaluate retrieval.



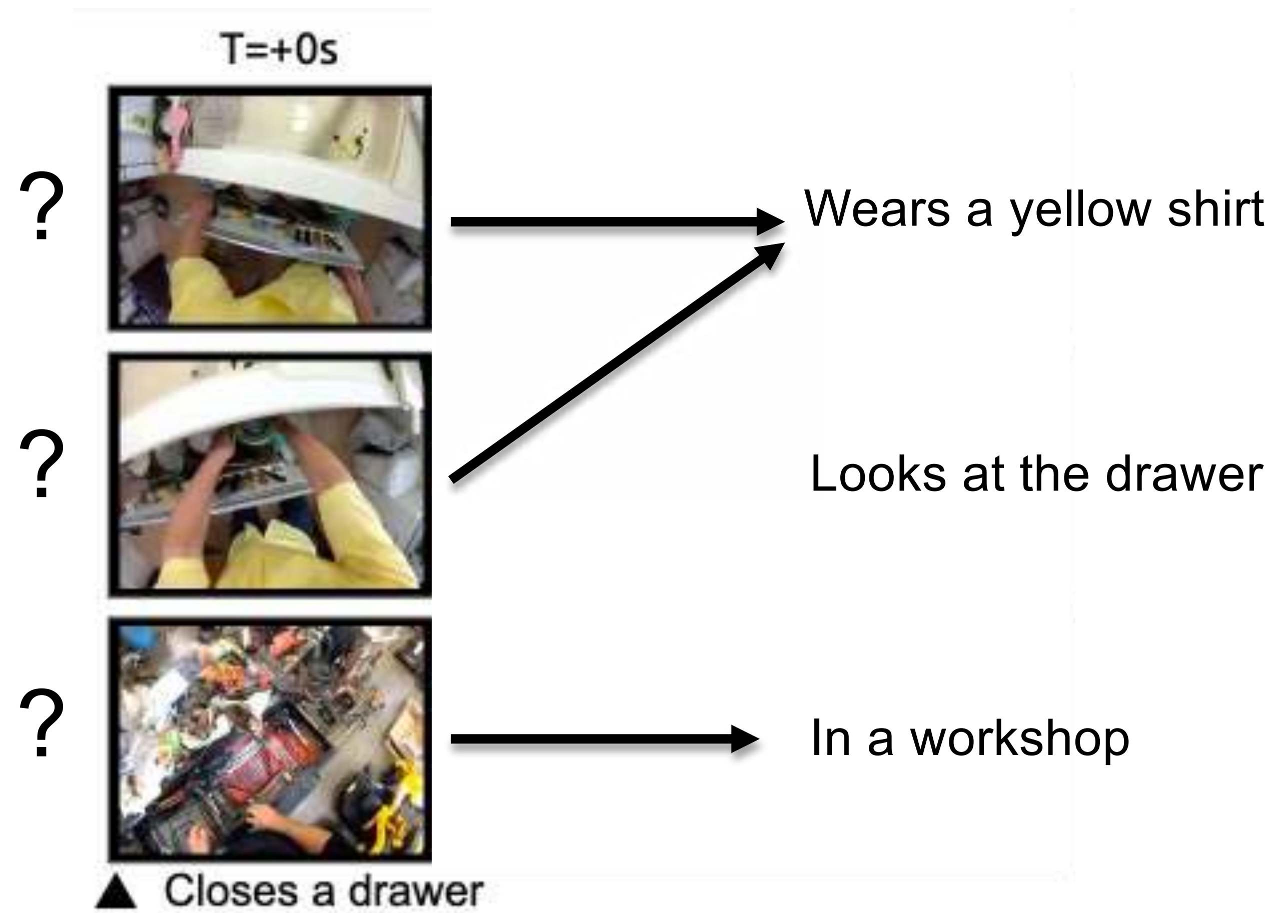
Video-text R @ 1

100%

Benchmarks

with: Toby Perrett
Tengda Han
Andrew Zisserman

Task: Caption every clip, then evaluate retrieval.



Video-text R @ 1

66%

with: Toby Perrett
Tengda Han
Andrew Zisserman

Wears a yellow shirt

Looks at the drawer

In a workshop

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

Text-video R @ 1
33%

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

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

Climbs the stairs and
holds the phone

Climbs the stairs and
picks up the drill

Climbs the stairs and
holds a tape measure

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

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

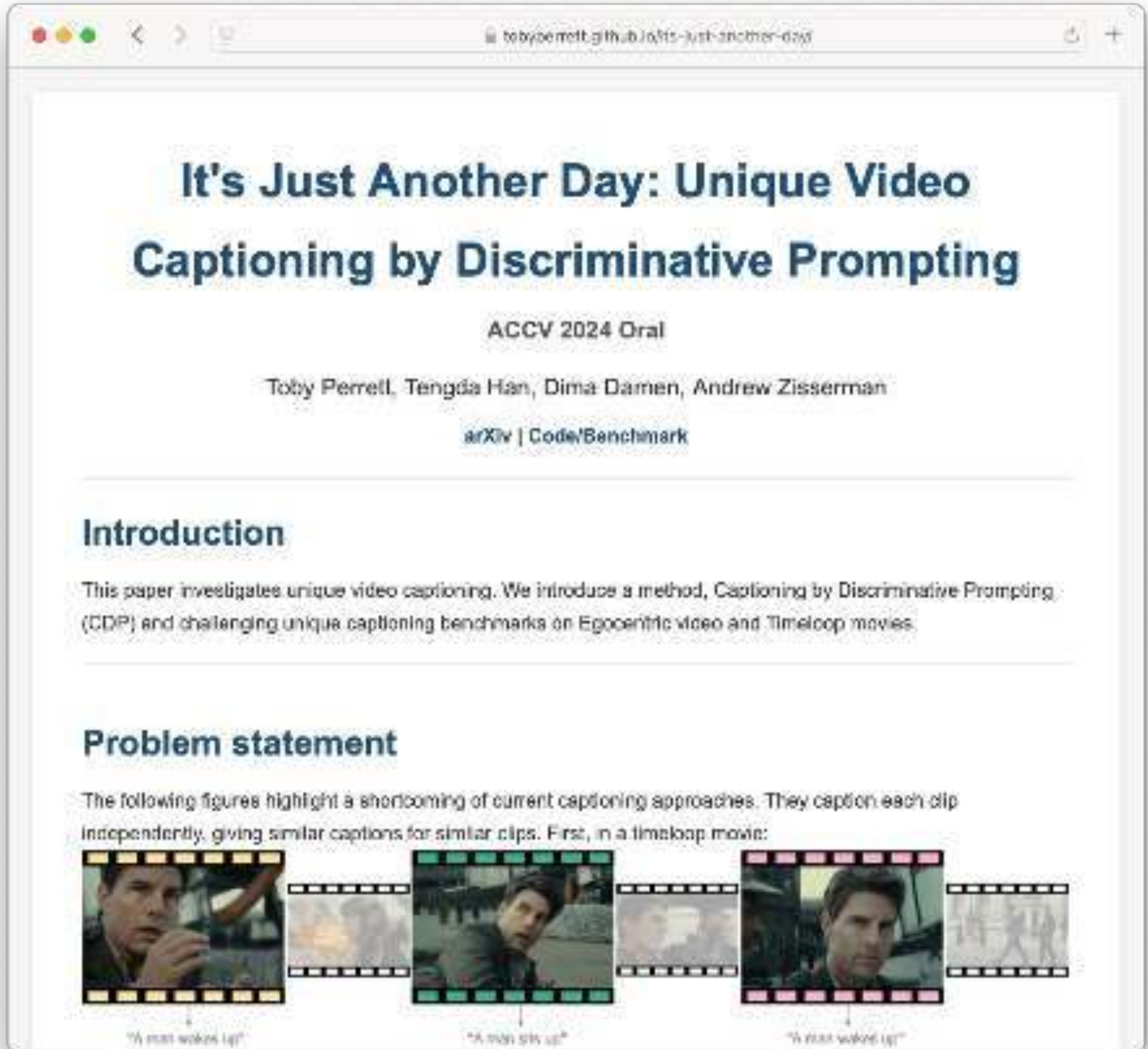
looks at the list

Looks around the shelves and then

picks up a packet of cough rubs

Unique Video Captioning

with: Toby Perrett
Tengda Han
Andrew Zisserman



tobyperrett.github.io/its-just-another-day/

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

ACCV 2024 Oral

Toby Perrett, Tengda Han, Dima Damen, Andrew Zisserman

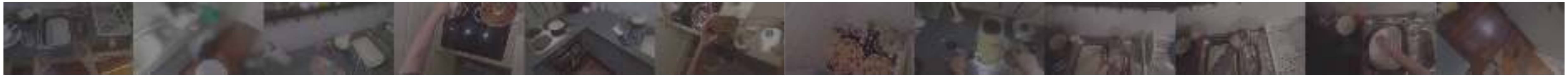
arXiv | Code/Benchmark

Introduction

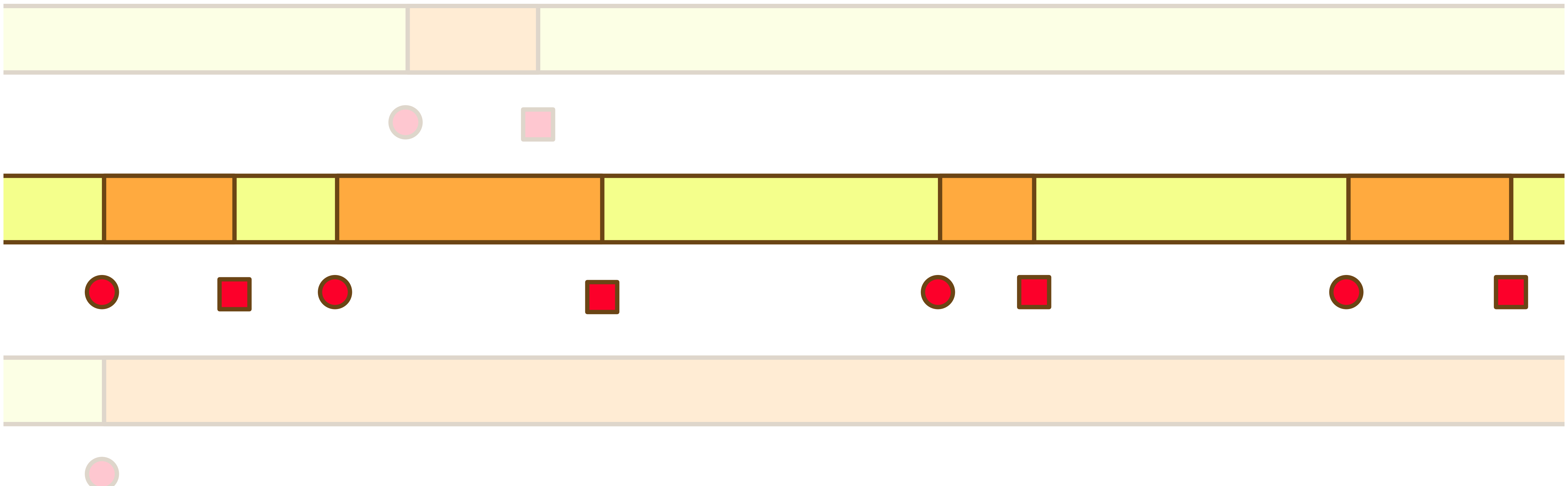
This paper investigates unique video captioning. We introduce a method, Captioning by Discriminative Prompting (CDP) and challenging unique captioning benchmarks on Egocentric video and Timeloop movies.

Problem statement

The following figures highlight a shortcoming of current captioning approaches. They caption each clip independently, giving similar captions for similar clips. First, in a timeloop movie:



Egocentric Video Understanding



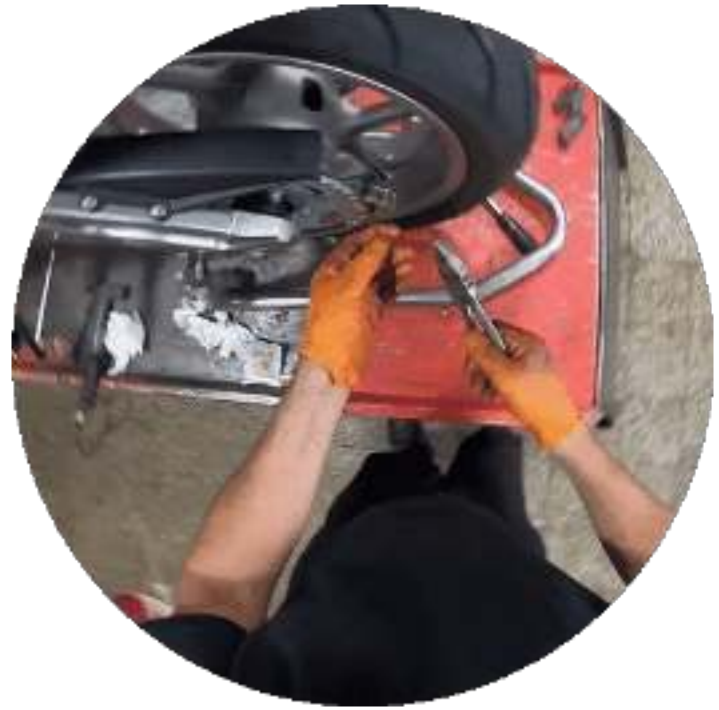
Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett

Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Generalisation across Scenarios and Locations

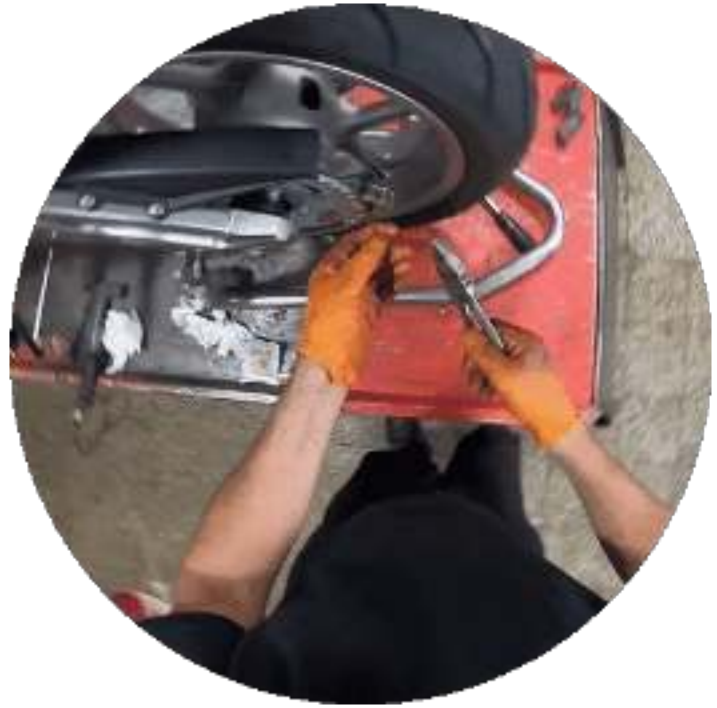
with: Chiara Plizzari
Toby Perrett



Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett

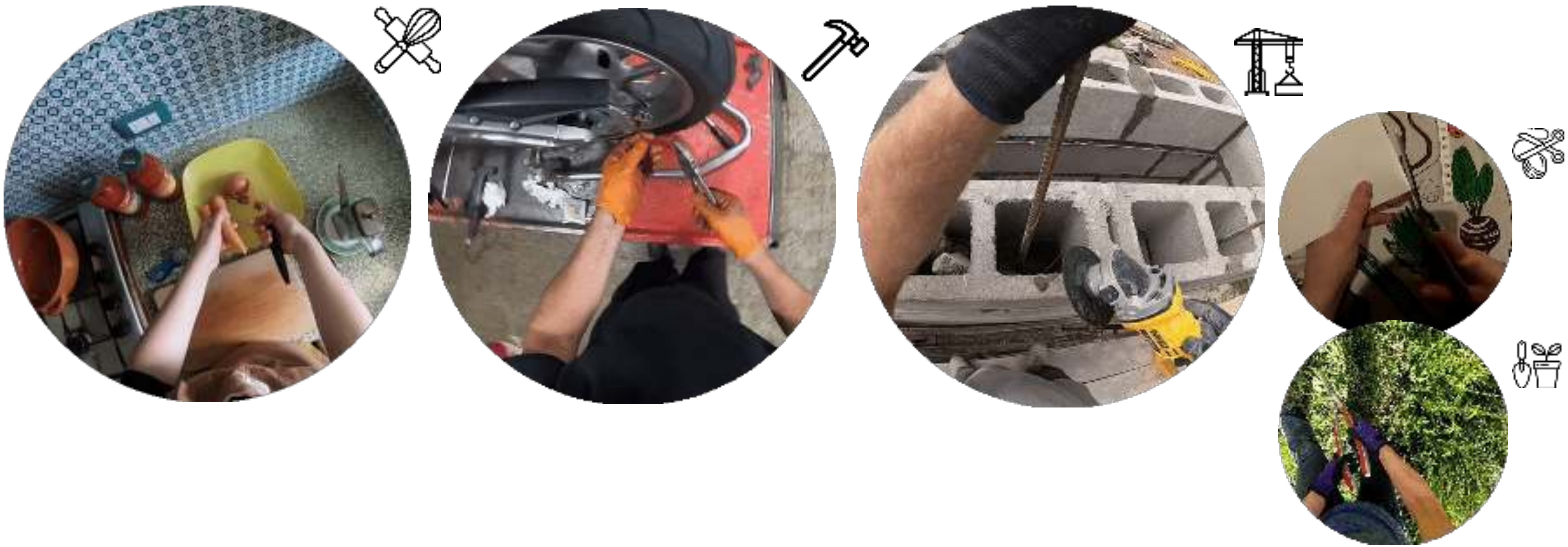






Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett



Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett

Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett

Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett

University of BRISTOL

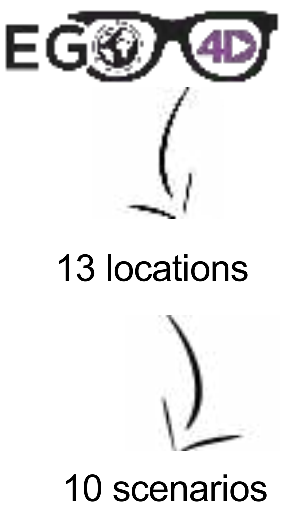
Plizzari et al (2023). What c... mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IE... ...rnational Conference on Computer Vision (ICCV).

Dima Damen
ICVSS2025

Dataset: ARGO1M

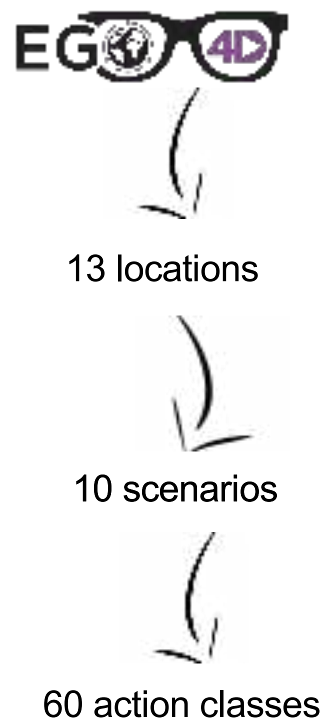
with: Chiara Plizzari
Toby Perrett

- We introduce **ARGO1M**, the first dataset to perform **Action Recognition Generalisation** Over Scenarios and Locations



with: Chiara Plizzari
Toby Perrett

- We introduce **ARGO1M**, the first dataset to perform **Action Recognition Generalisation** Over Scenarios and Locations



Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Dataset: ARGO1M

with: Chiara Plizzari
Toby Perrett

- We introduce **ARGO1M**, the first dataset to perform **Action Recognition Generalisation** Over Scenarios and Locations

NEW **1.1M samples**

EGO4D

13 locations

10 scenarios

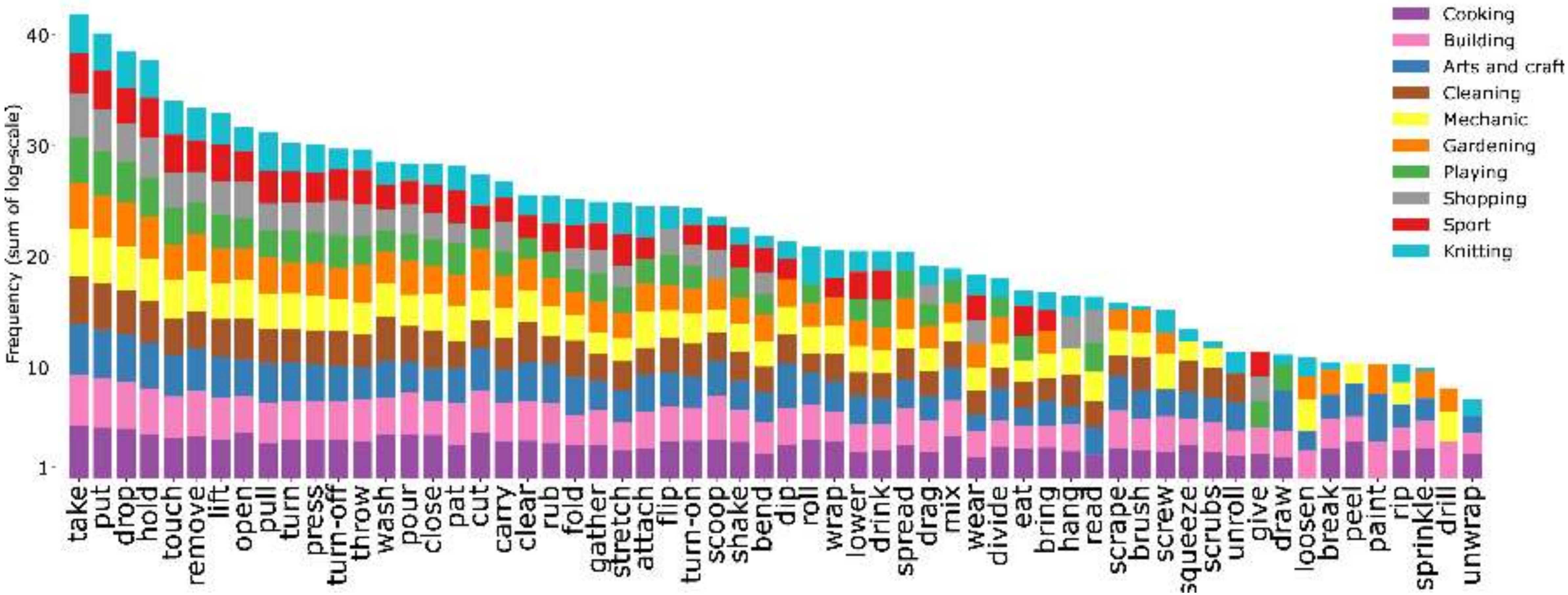
60 action classes

Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett

ARGO1M: 1.05M action clips from 60 action classes recorded in 13 locations within 10 scenarios



Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

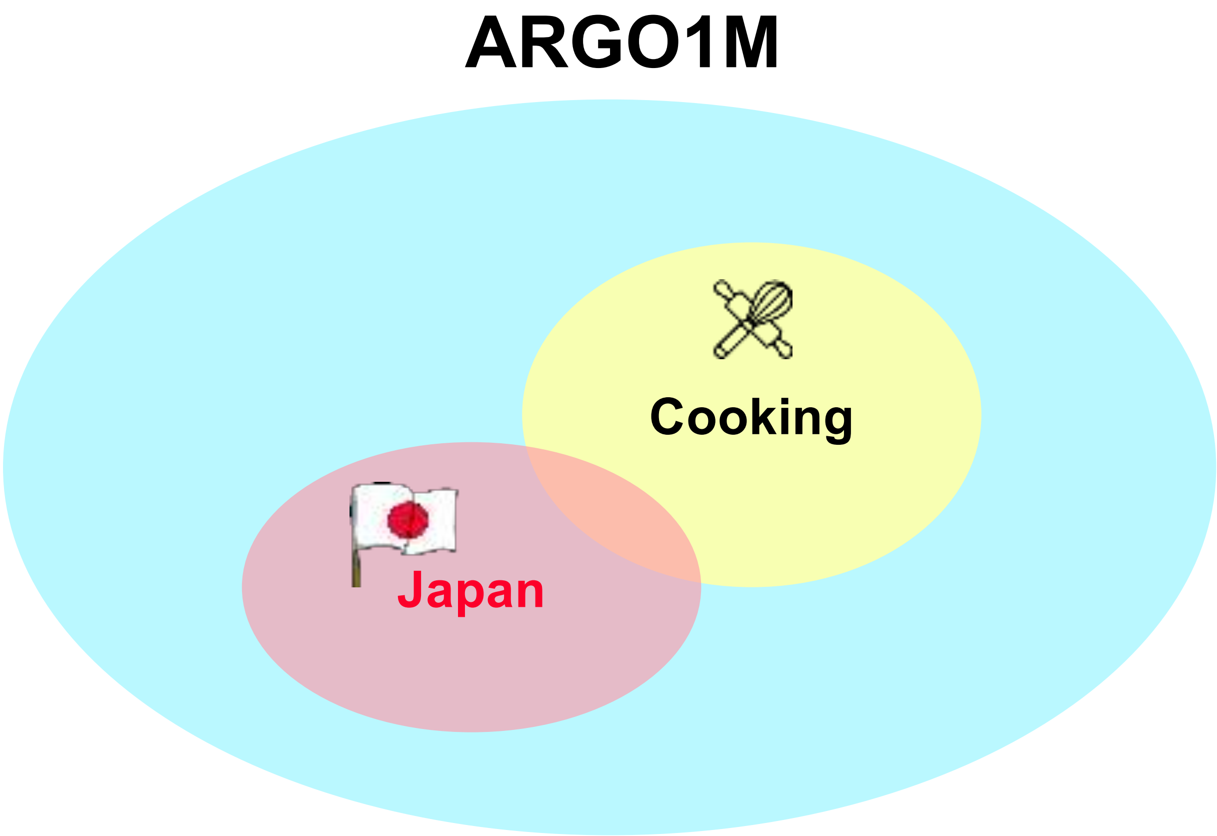
ARGO1M Splits

ARGO1M

Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

ARGO1M Splits

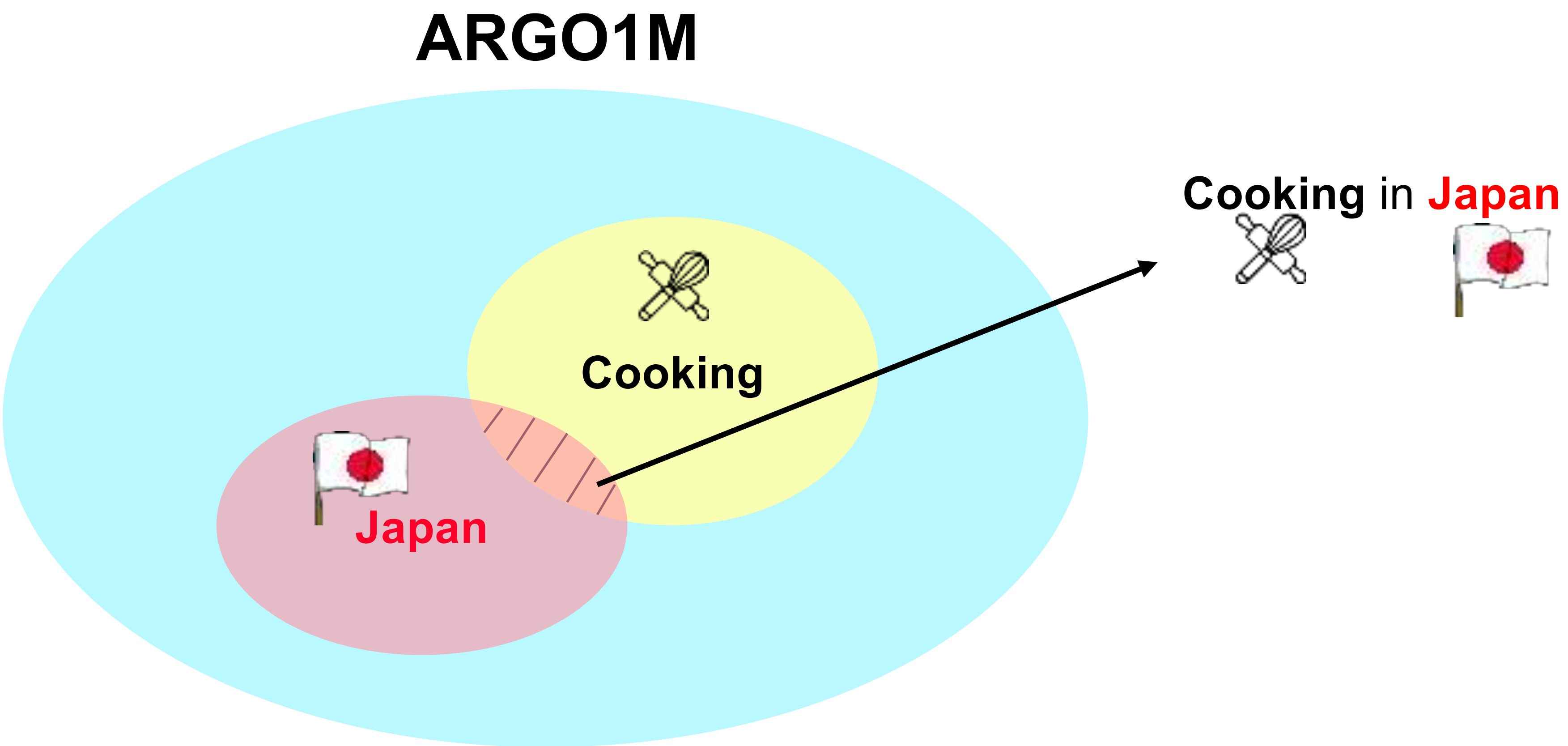
with: Chiara Plizzari
Toby Perrett



Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

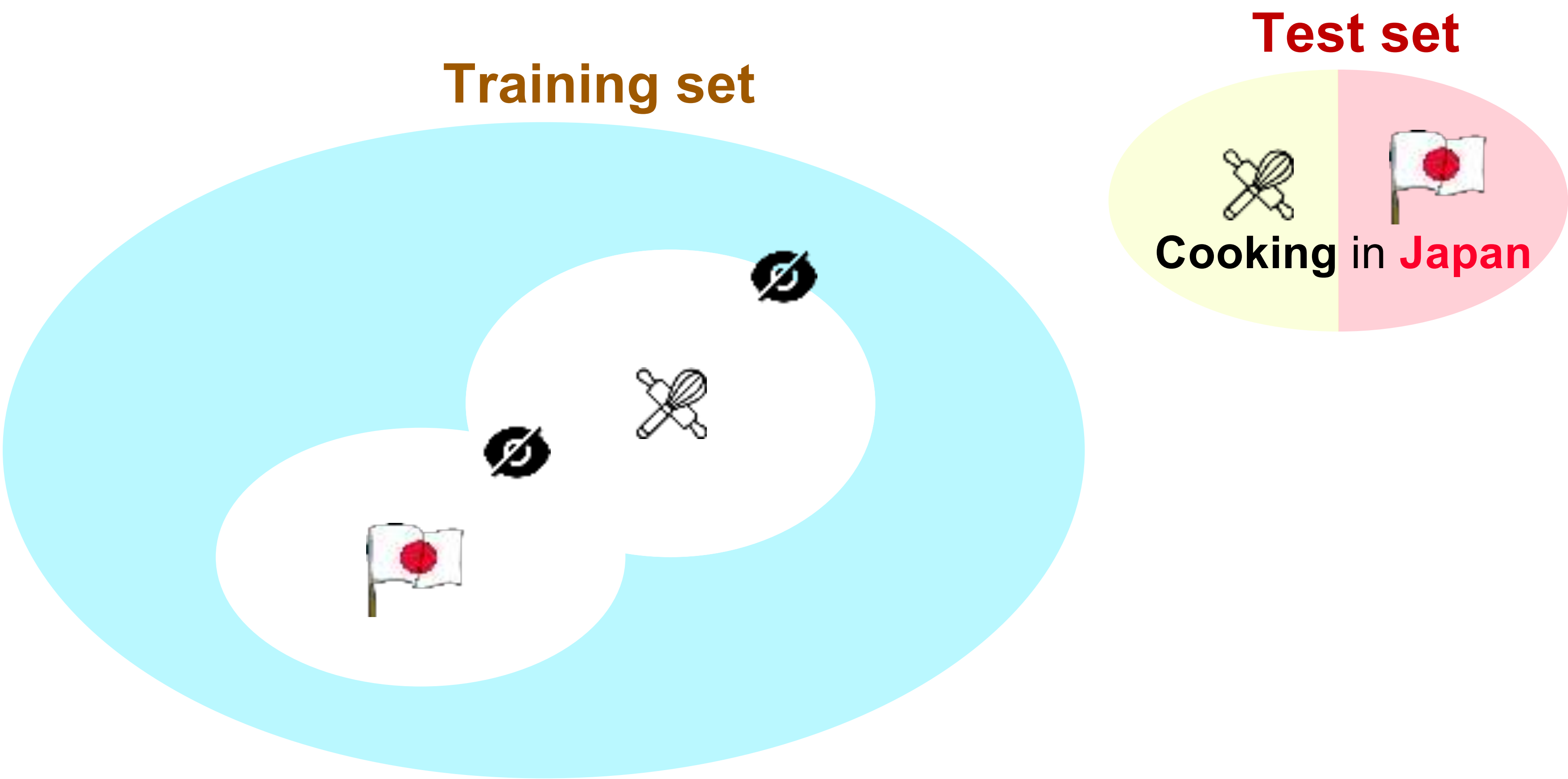
ARGO1M Splits

with: Chiara Plizzari
Toby Perrett



Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

ARGO1M Splits

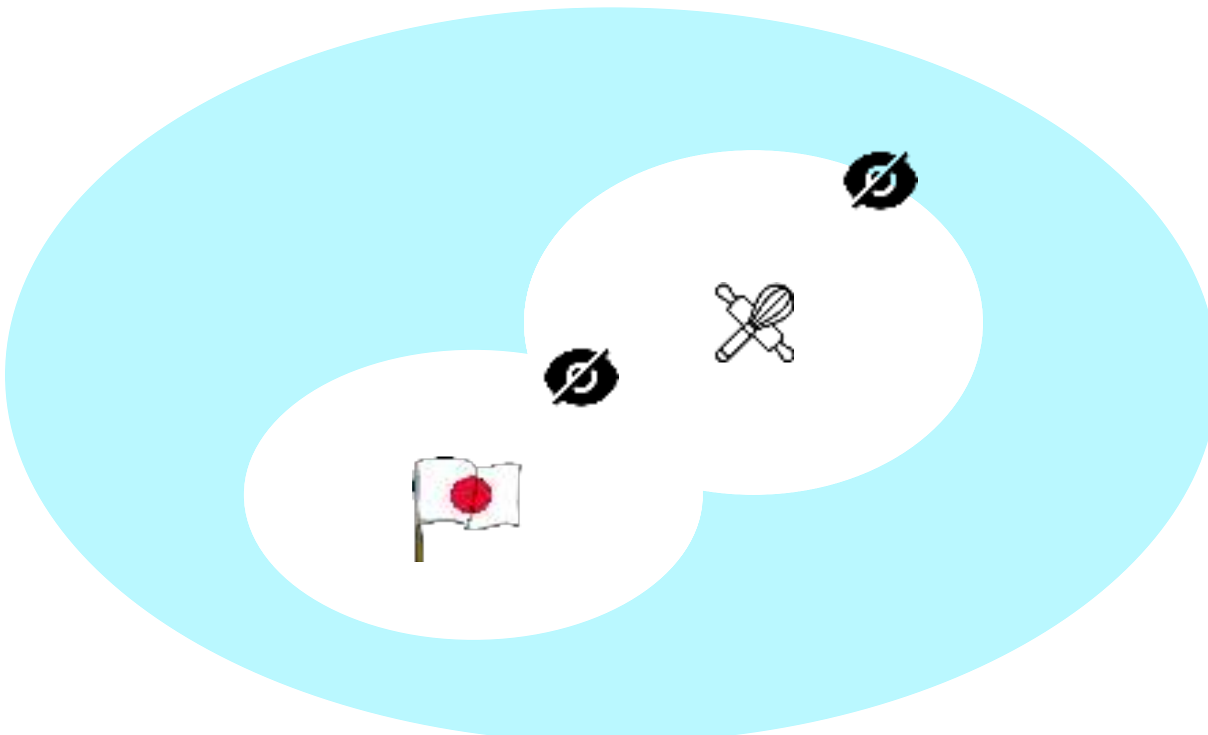


Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

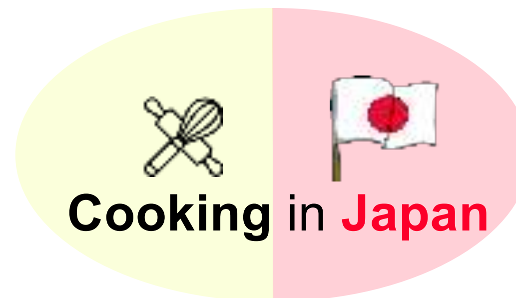
ARGO1M Splits

with: Chiara Plizzari
Toby Perrett

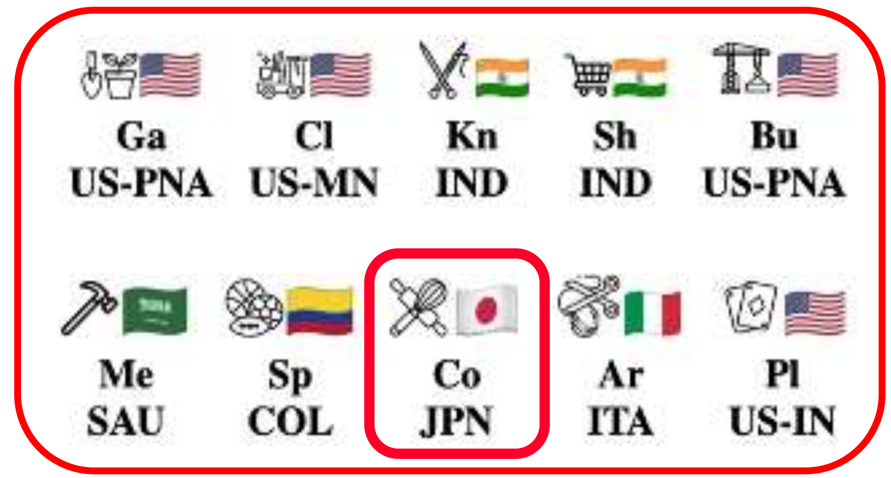
Training set



Test set



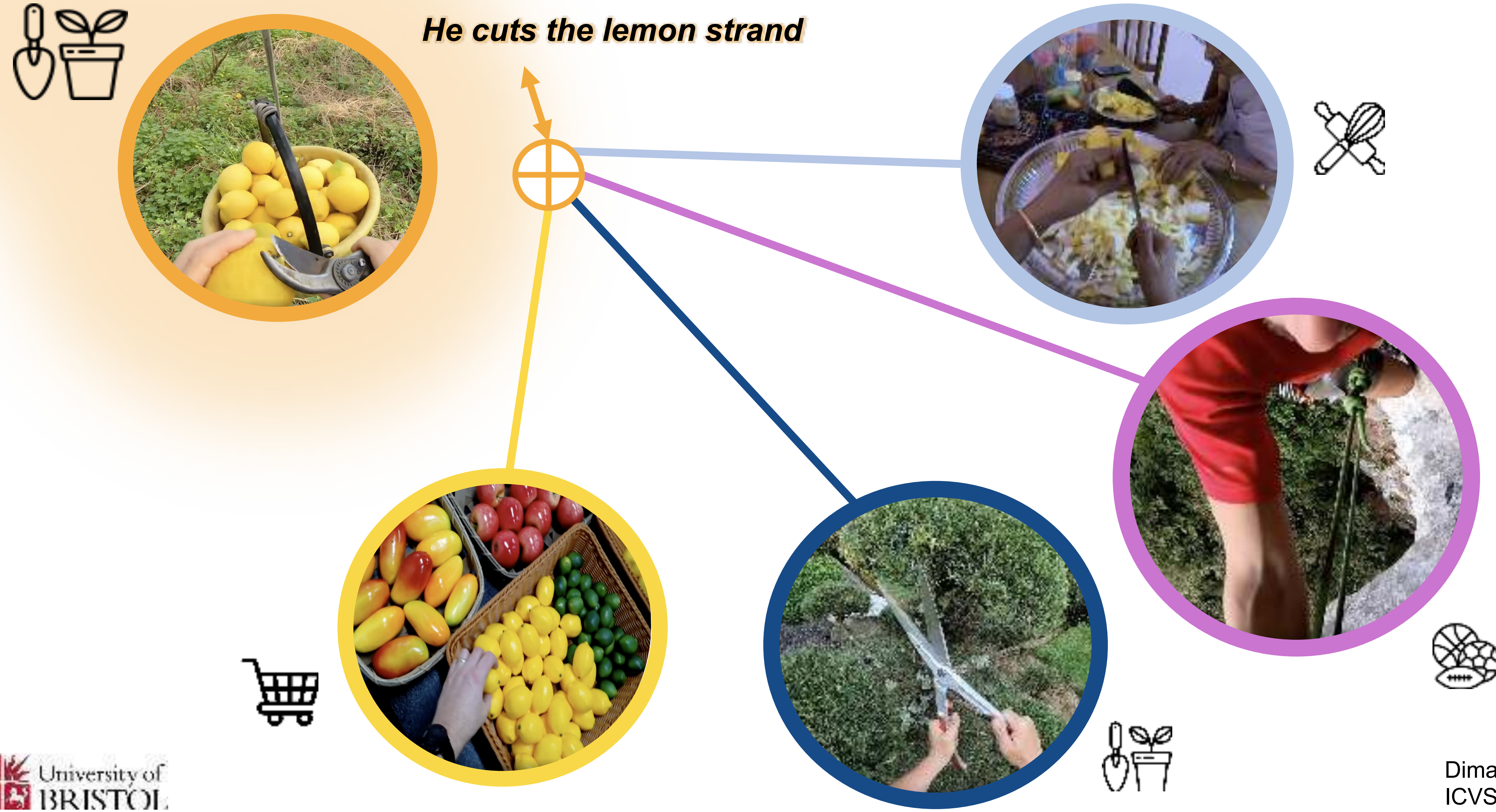
10 test sets



Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

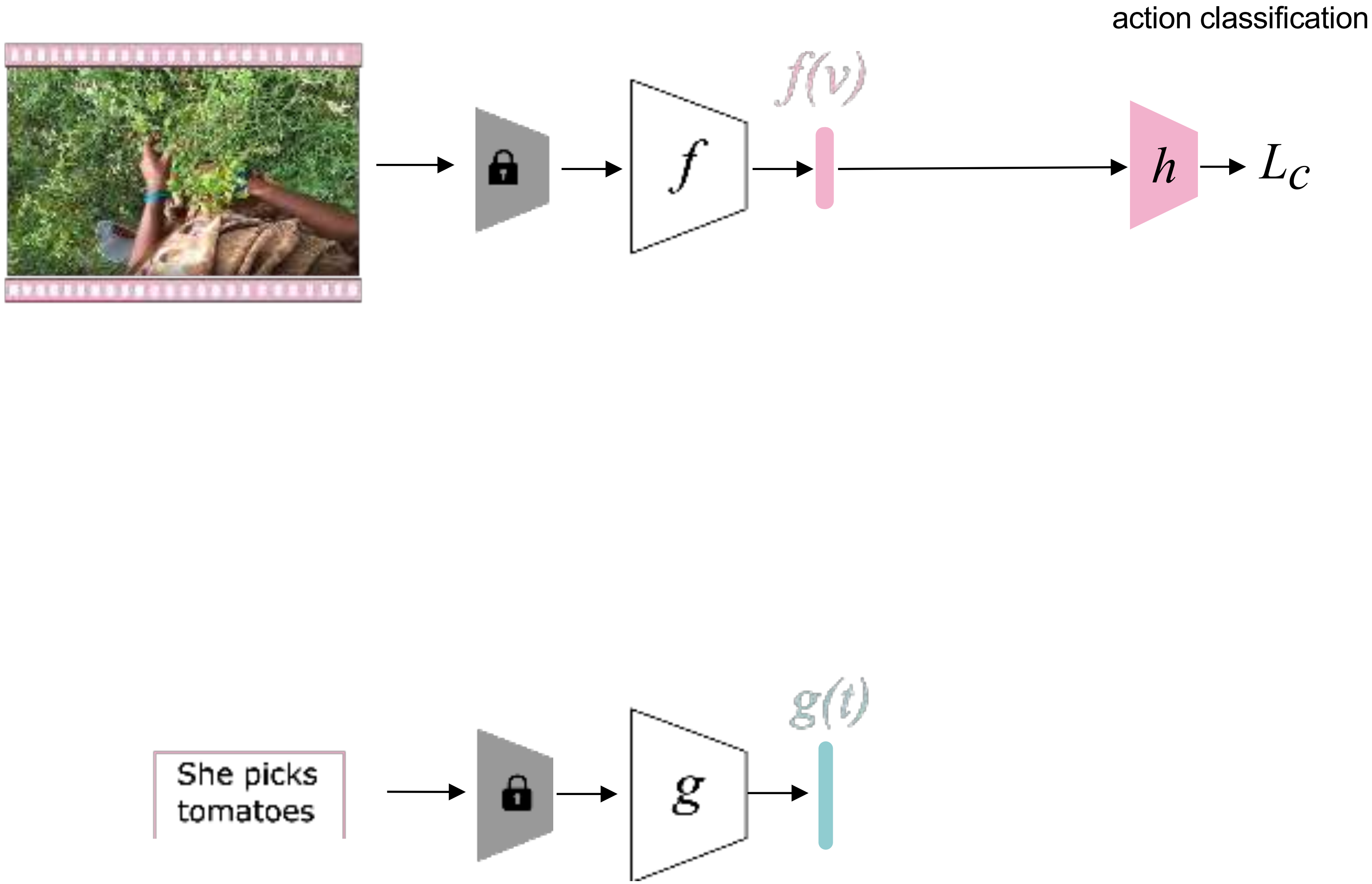
Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett



Proposed method: CIR

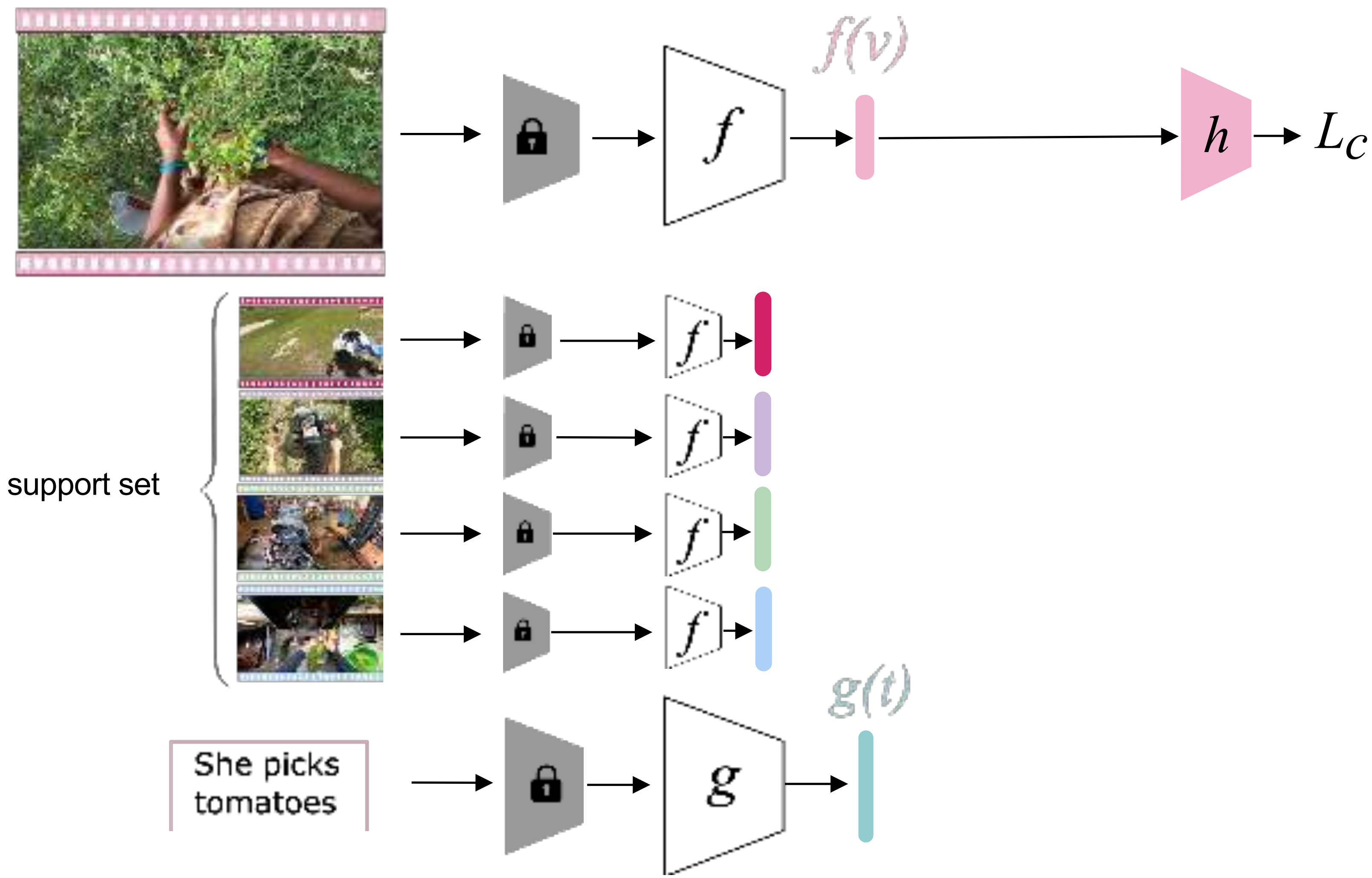
with: Chiara Plizzari
Toby Perrett



Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

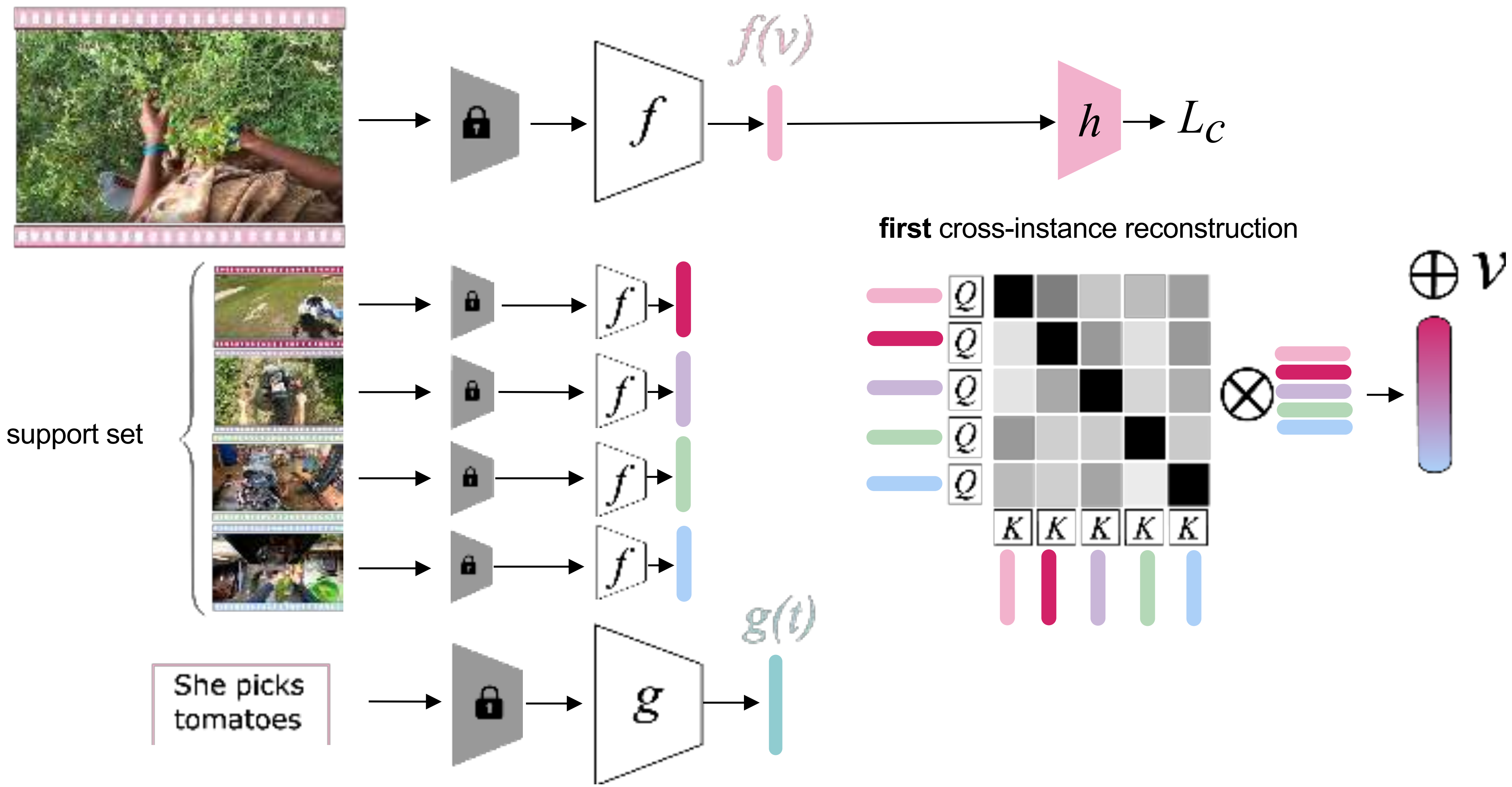
Proposed method: CIR

with: Chiara Plizzari
Toby Perrett



Proposed method: CIR

with: Chiara Plizzari
Toby Perrett

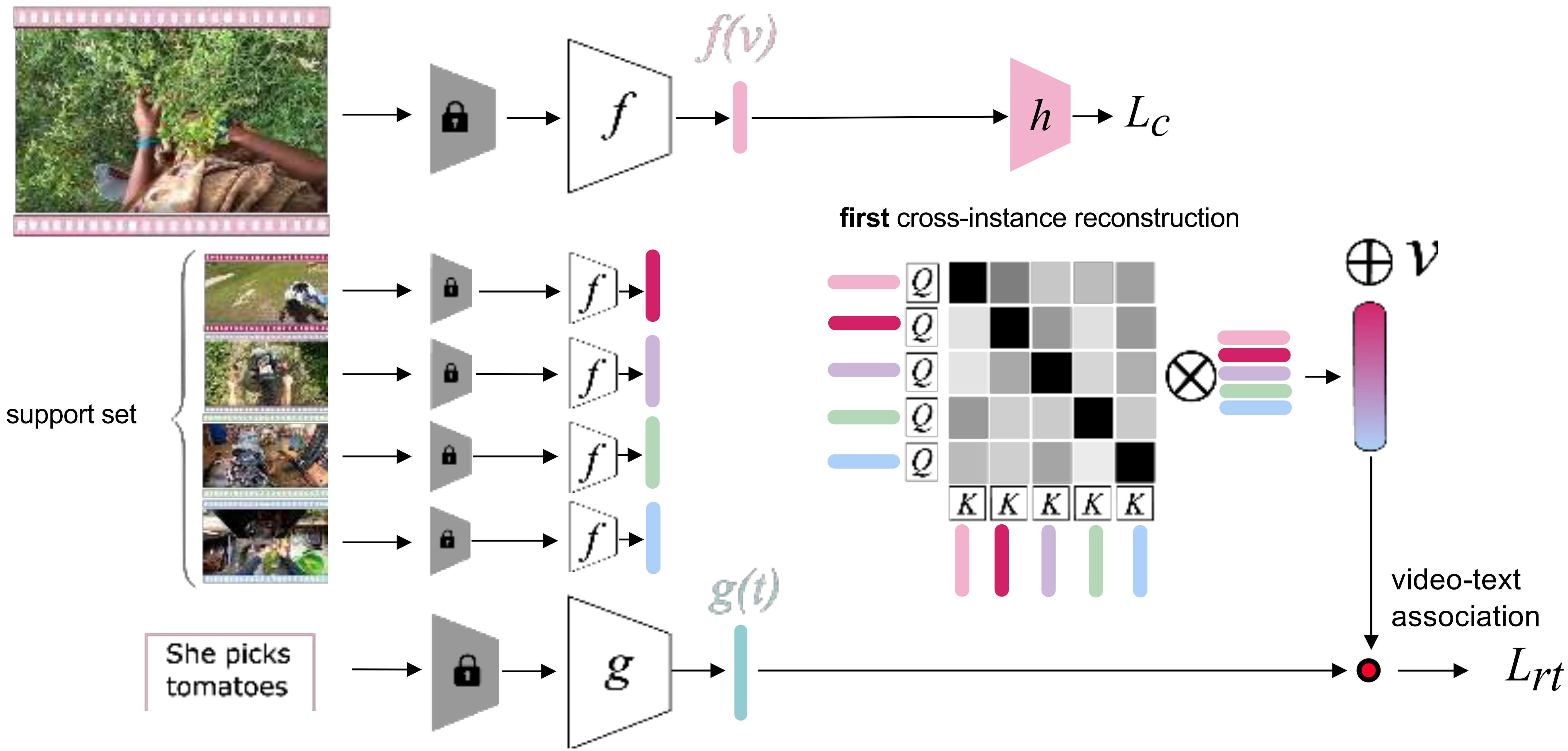


University of BRISTOL

Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Proposed method: CIR

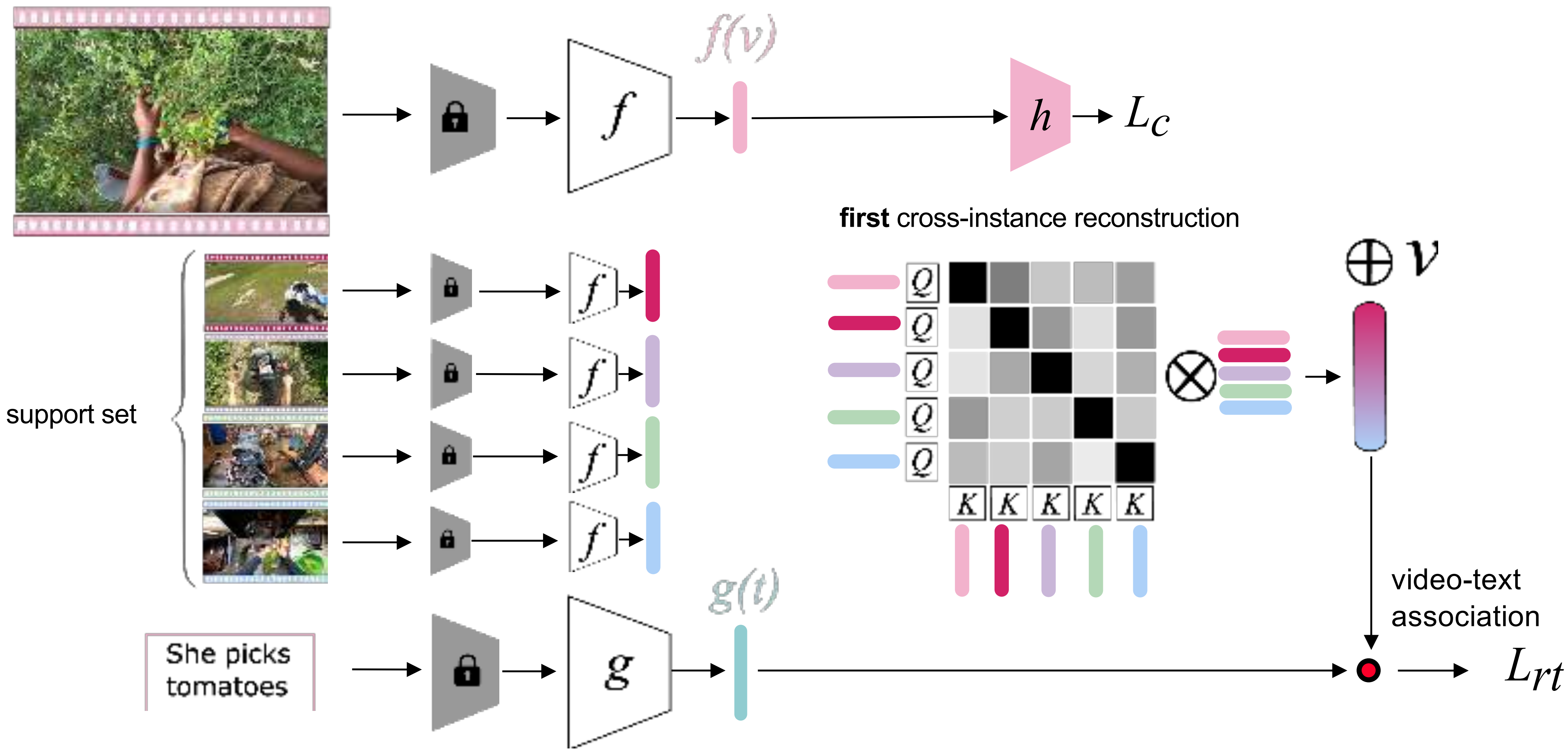
with: Chiara Plizzari
Toby Perrett



Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Proposed method: CIR

with: Chiara Plizzari
Toby Perrett

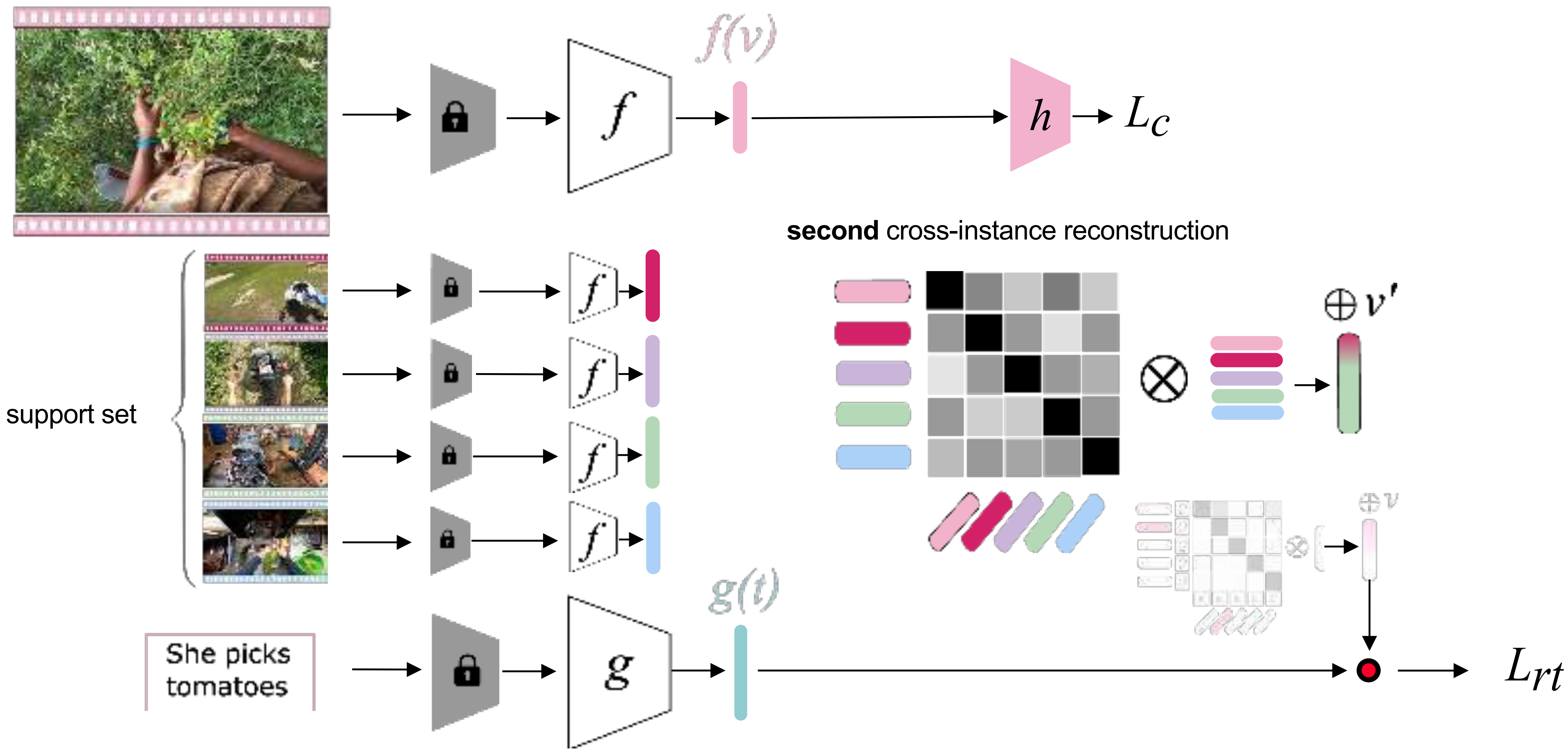


University of BRISTOL

Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Proposed method: CIR

with: Chiara Plizzari
Toby Perrett

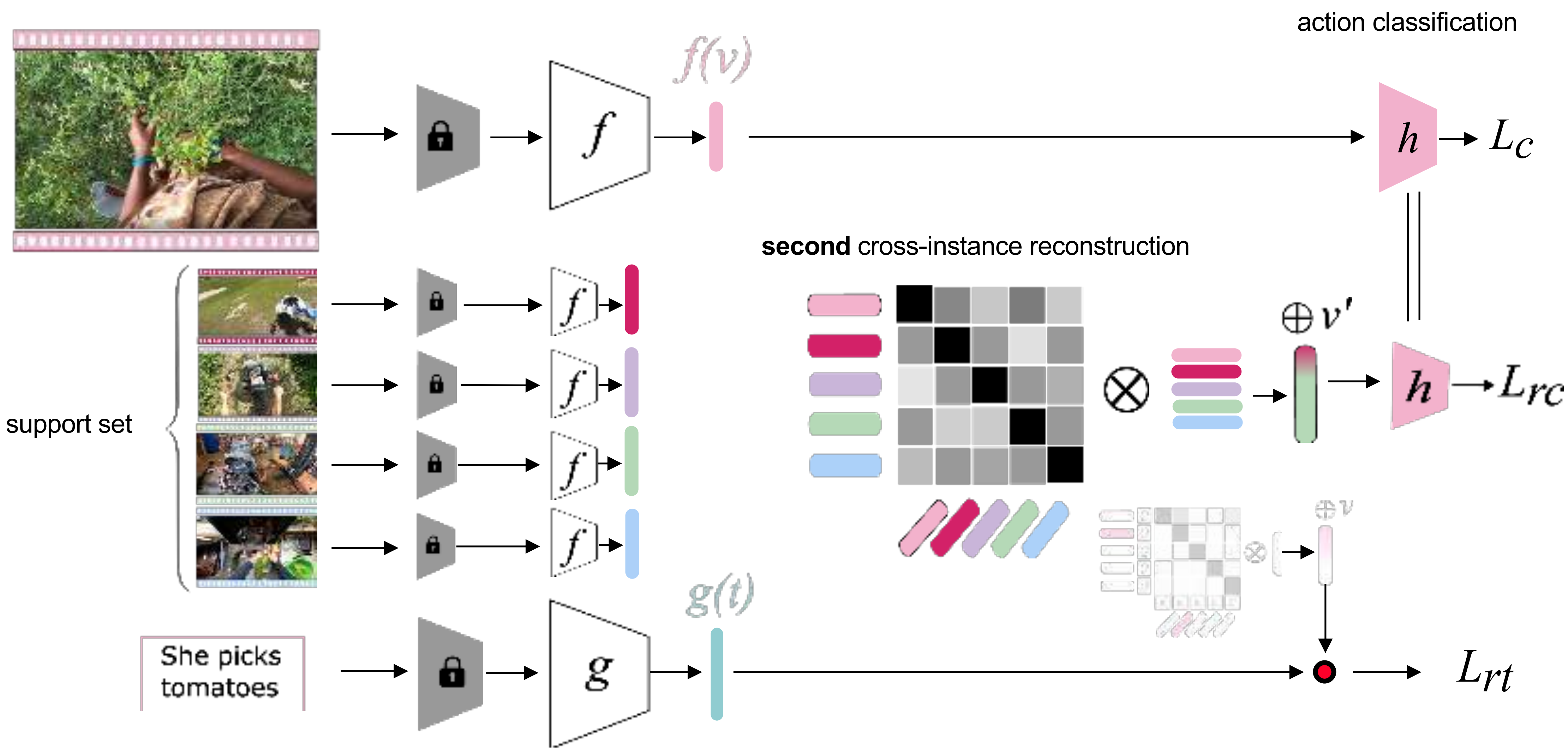


University of BRISTOL

Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

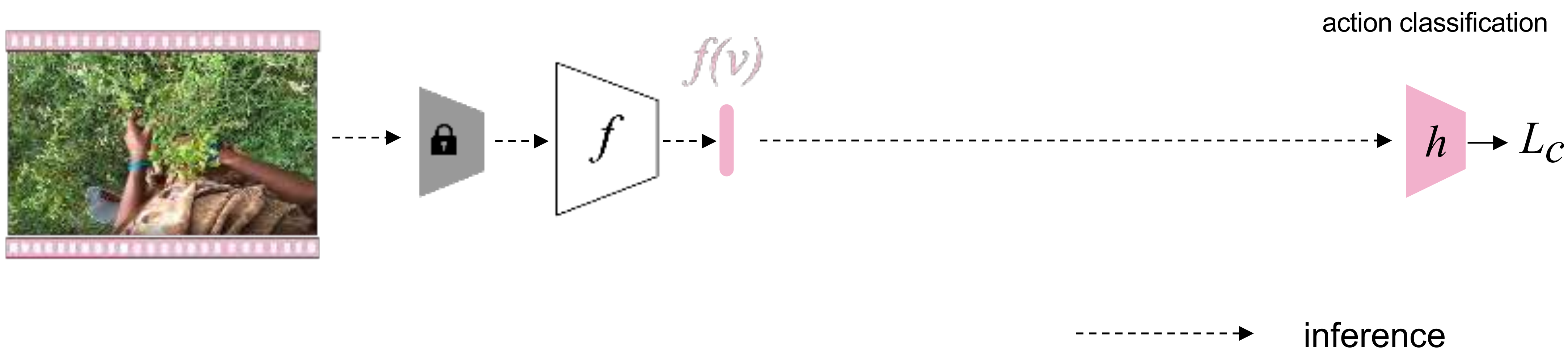
Proposed method: CIR

with: Chiara Plizzari
Toby Perrett



Proposed method: CIR

with: Chiara Plizzari
Toby Perrett



Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Examples

Chiara Plizzari
Toby Perrett
Dima Damen

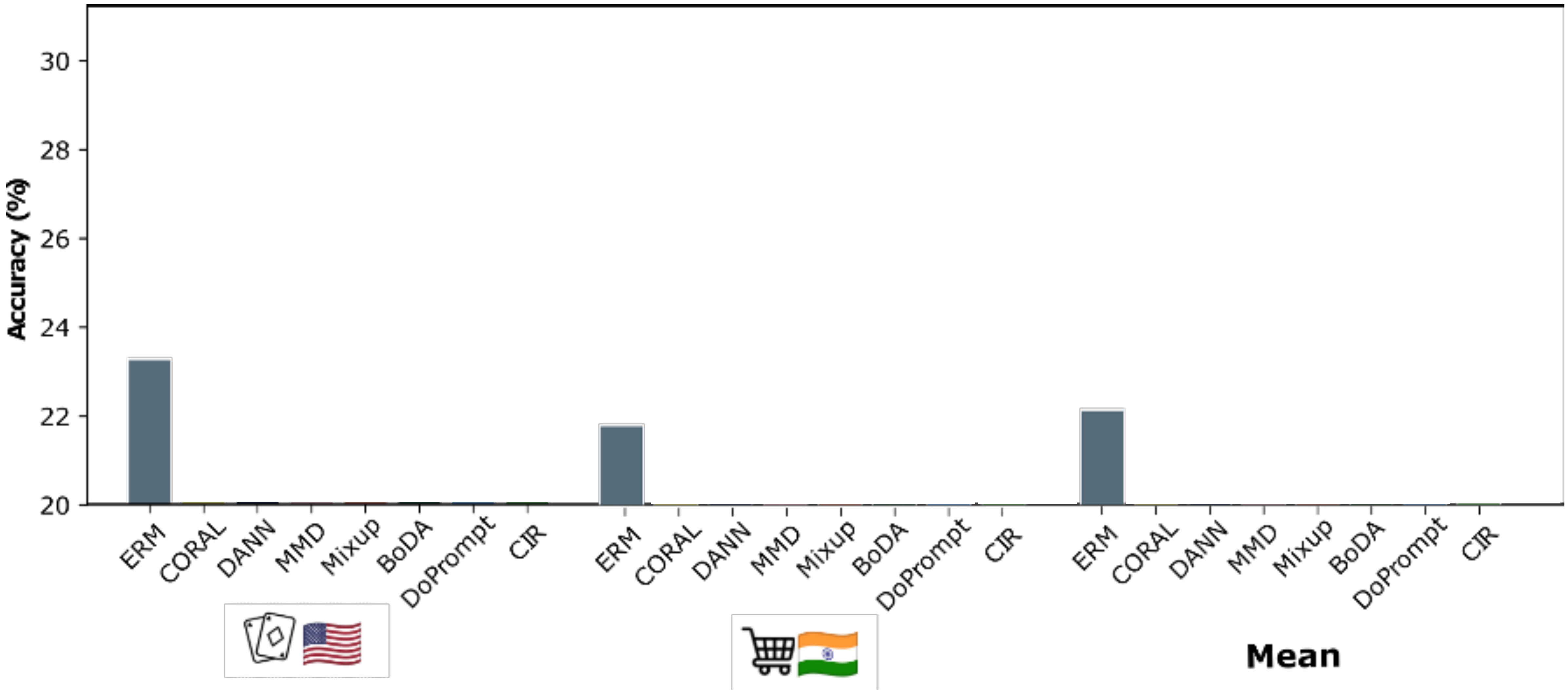
#C C drops the cut vegetables

query

support 1 support 2 support 3 support 4 support 5

Proposed method: CIR

with: Chiara Plizzari
Toby Perrett



Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

What can a cook in Italy teach a mechanic in India?

with: Chiara Plizzari
Toby Perrett


arXiv:2306.08713v1 [cs.CV] 14 Jun 2023

**What can a cook in Italy teach a mechanic in India?
Action Recognition Generalisation Over Scenarios and Locations**

Chiara Plizzari*† Toby Perrett* Barbara Caputo* Dima Damen*

* Politecnico di Torino, Italy * University of Bristol, United Kingdom

Abstract

1. Introduction

ARGO1M Dataset
CIR Method
Code and Models

RELEASED

University of BRISTOL

Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Dima Damen
ICVSS2025

HD-EPIC: A Highly-Detailed Egocentric Video Dataset

HD-EPIC

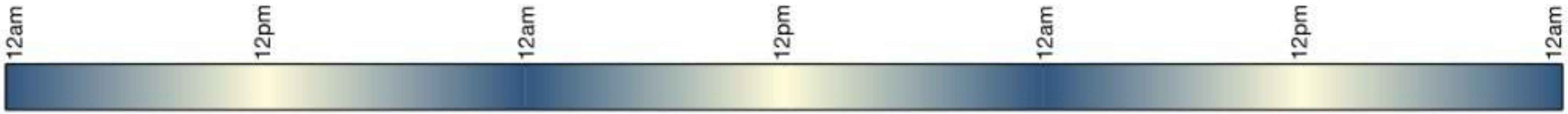
University of Bristol
Universiteit Leiden
Singapore Management University
University of Bath

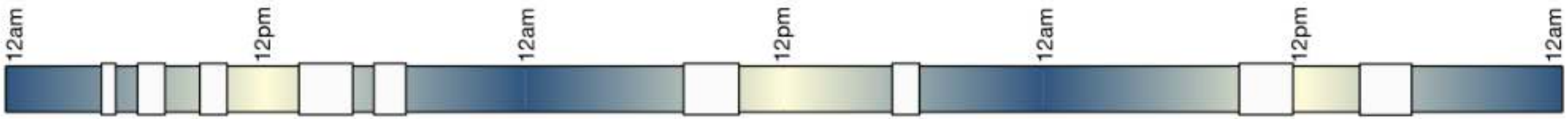
Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha
Omar Emara, Sam Pollard, Kranti Parida
Kaiting Liu, Prajwal Gatti, Siddhant Bansal
Kevin Flanagan, Jacob Chalk, Zhifan Zhu
Rhodri Guerrier, Fahd Abdelazim, Bin Zhu
Davide Moltisanti, Michael Wray, Hazel Doughty
Dima Damen

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

HD-EPIC

HD-EPIC





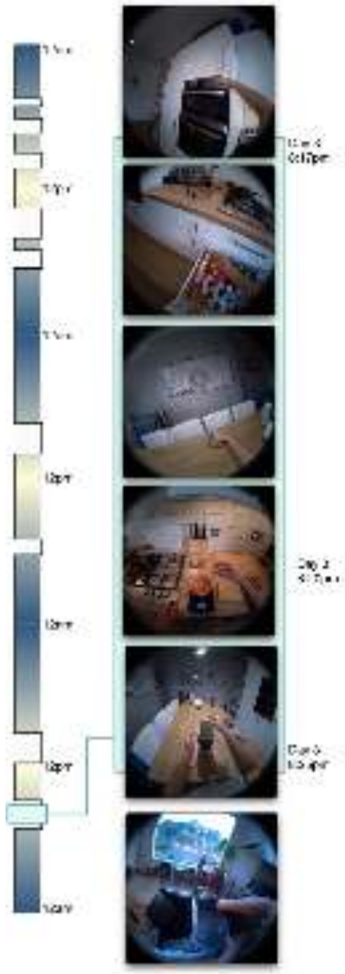
Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

HD-EPIC

HD-EPIC

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

HD-EPIC



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

HD-EPIC

Day 3
:17pm

Recipe: Southwestern Salad

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.

Recipe and nutrition

HD-EPIC

Cacio e Pepe (modified)

Ingredients:

200 g → penne

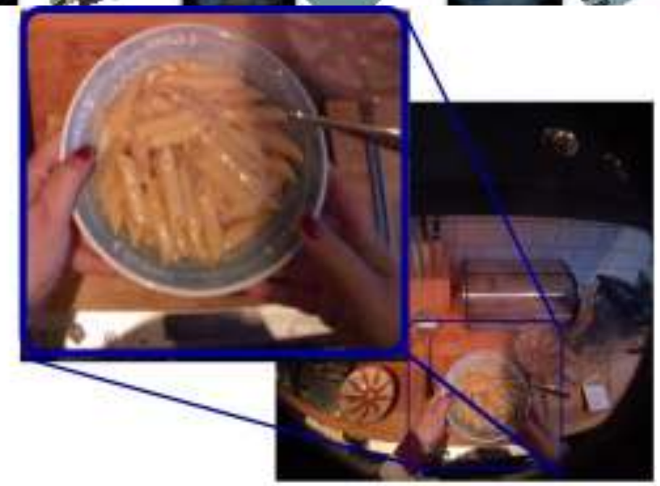
~~400g~~ of pasta of your choice (we recommend bucatini)

1 ~~2~~ 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

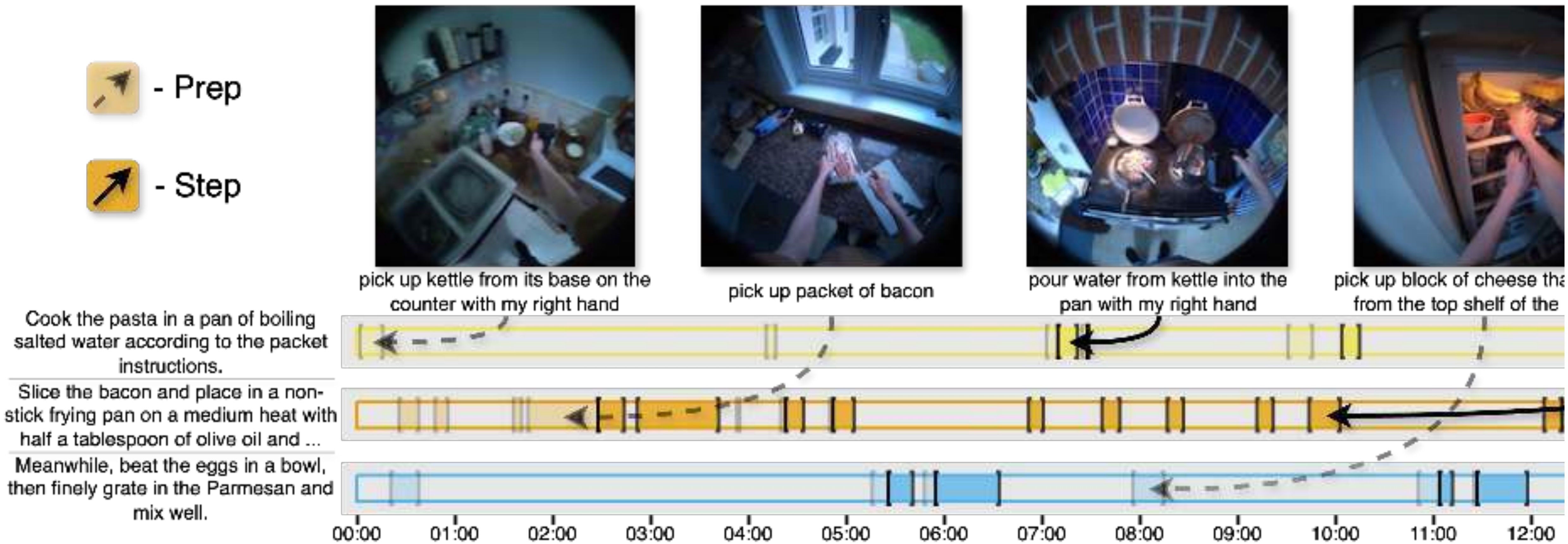
HD-EPIC

- 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.**

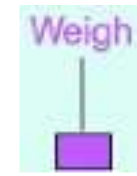
HD-EPIC



- 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.
00:00
01:00
02:00
03:00
04:00
05:00
06:00
07:00
08:00
09:00
10:00
11:00
12:00

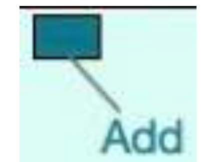
HD-EPIC

"P01_R03_I01": {
"name": "penne pasta",
"amount": 125,
"amount_unit": "g",
"calories": 445,
"fat": 1.9,
"carbs": 90,
"protein": 15,

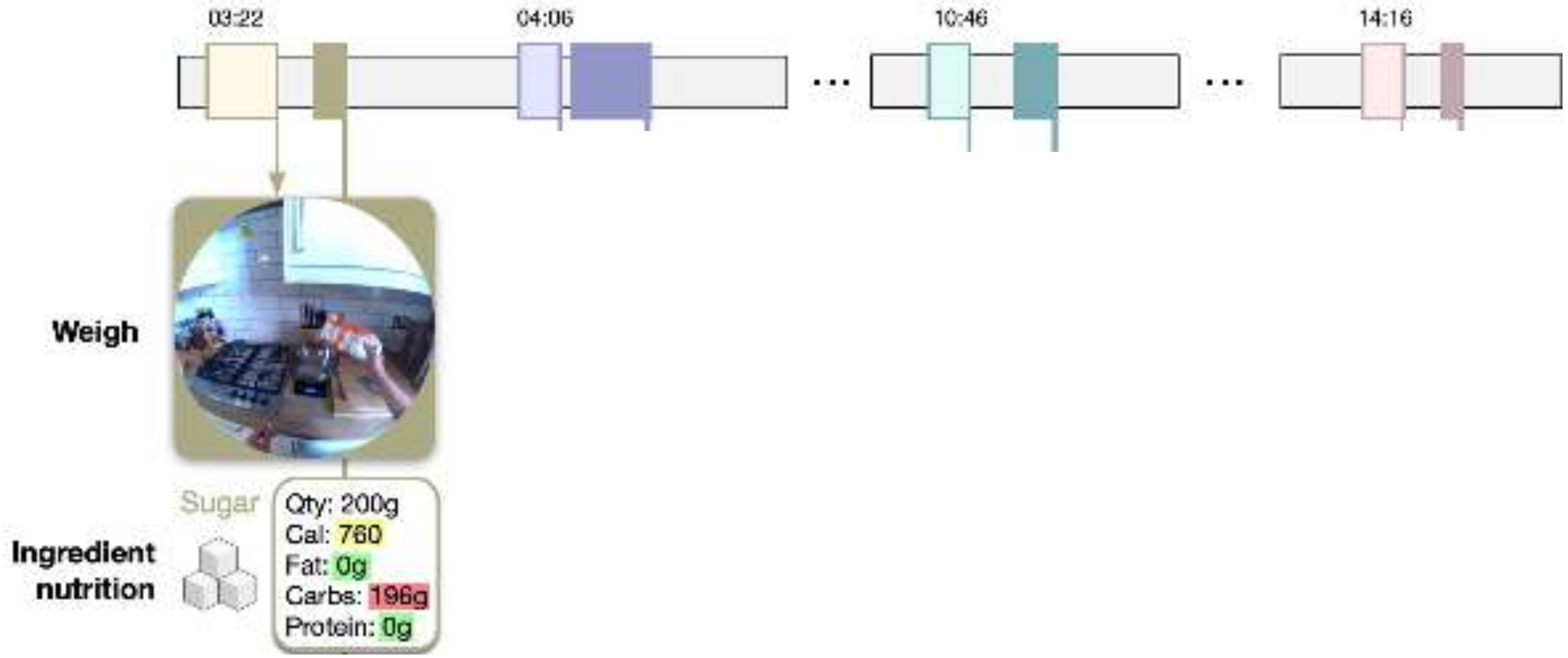


HD-EPIC

"P01_R03_I01": {
"name": "penne pasta",
"amount": 125,
"amount_unit": "g",
"calories": 445,
"fat": 1.9,
"carbs": 90,
"protein": 15,



HD-EPIC

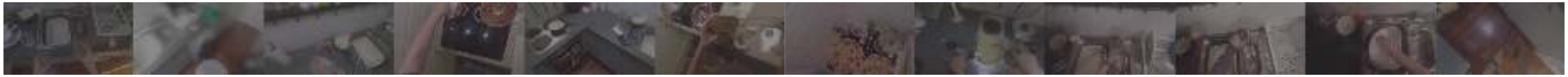


03:22
04:06
10:46
14:16
Weigh
Sugar
Ingredient nutrition
Qty: 200g
Cal: 760
Fat: 0g
Carbs: 196g
Protein: 0g

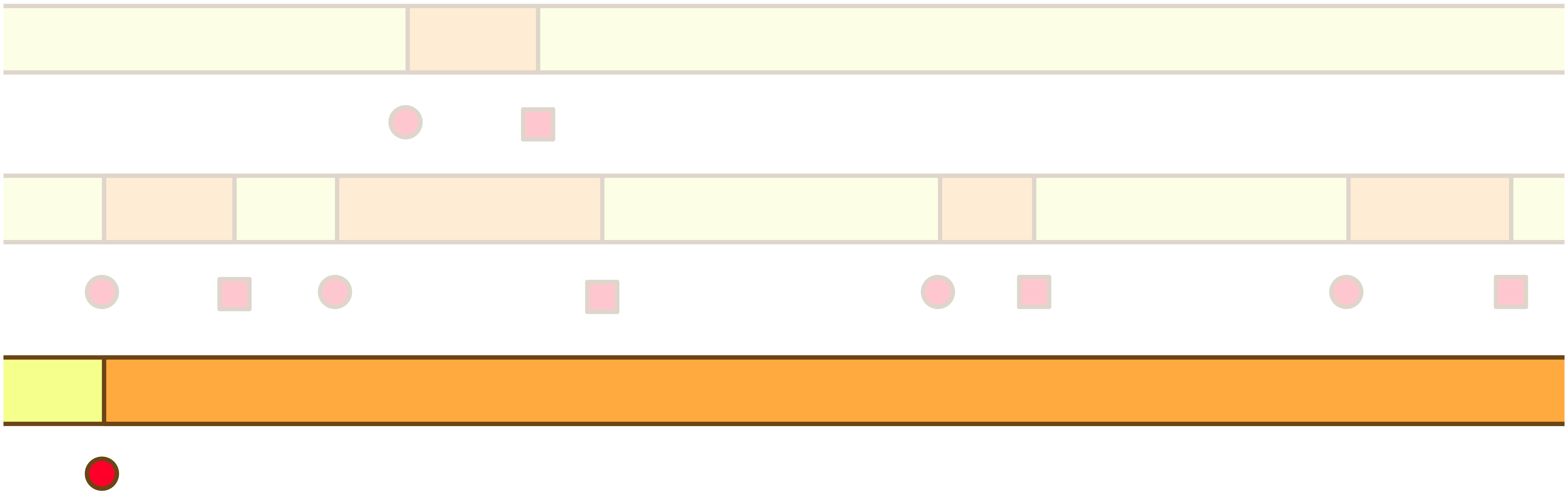
HD-EPIC

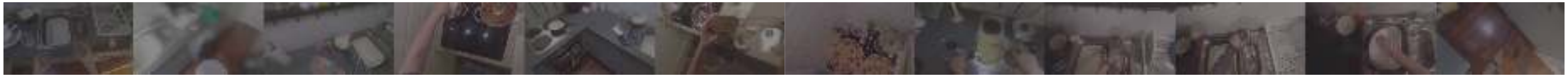
- 69 recipes. Avg: 6.6 steps, 8.1 ingredients, 4.8 hours, 2.1 videos.
- Our longest recipe took 2 days and 6 hours to complete.

HD-EPIC Videos	Recipe Videos
Multiple recipes / video, Multiple videos / recipe	One video == one recipe
All the action including preparatory and cleaning/clearing	Only the steps (prep edited out)
Faster (no explanations)	Slower (with descriptions)
Ingredients weighed on camera	Ingredients pre-weighed



Egocentric Video Understanding





Eventually…

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

But….

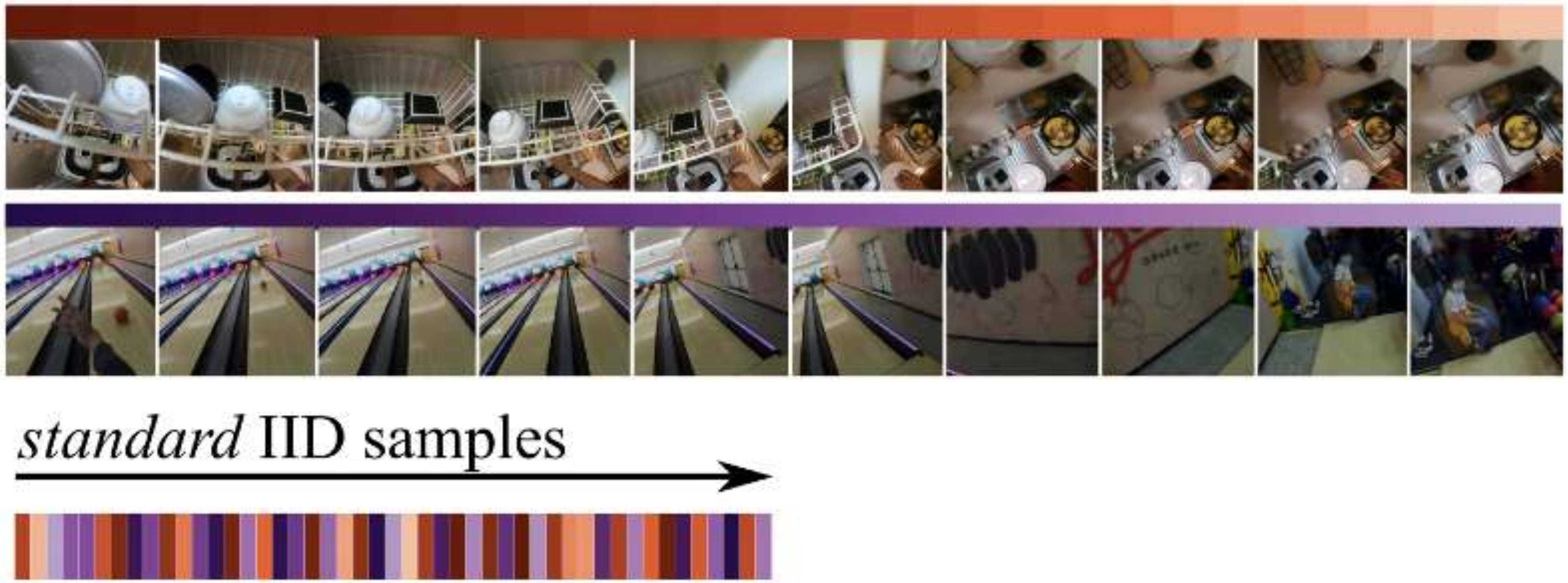
- Immense potential …

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



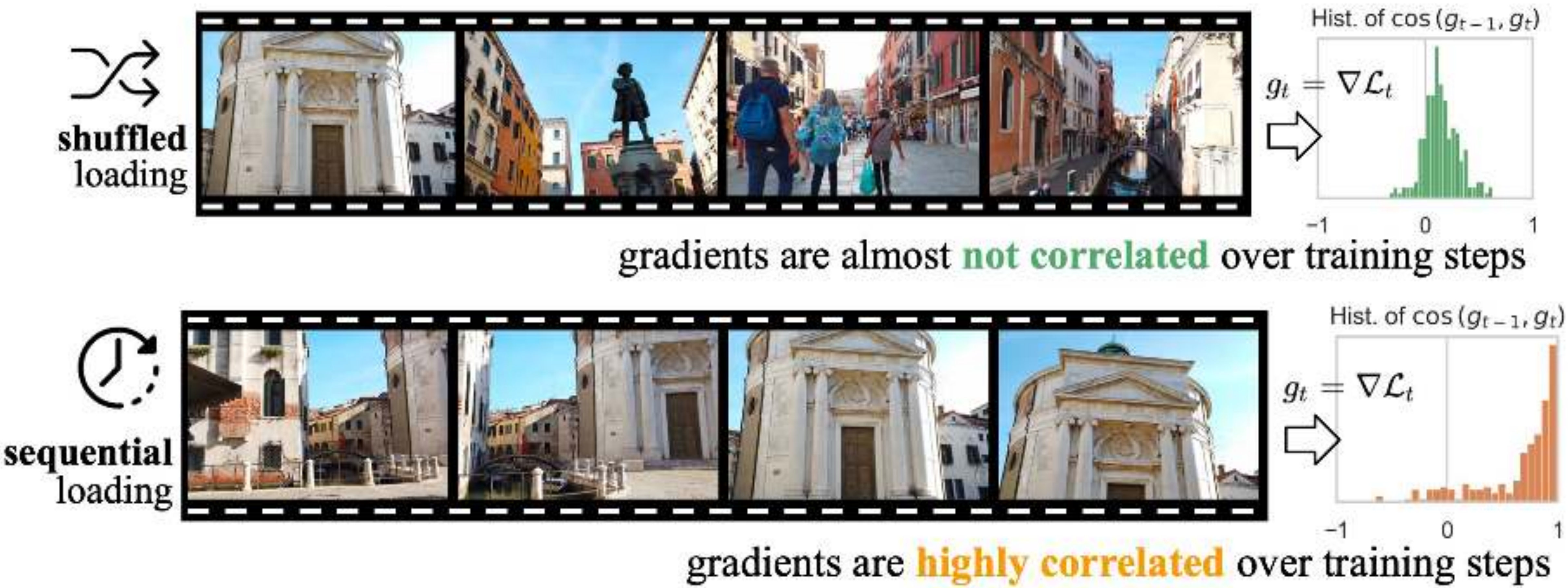
Carriera et al (2024). Learning from One Continuous Video Stream. CVPR

Dima Damen
ICVSS2025

Learning from Streaming Videos with Orthogonal Gradients

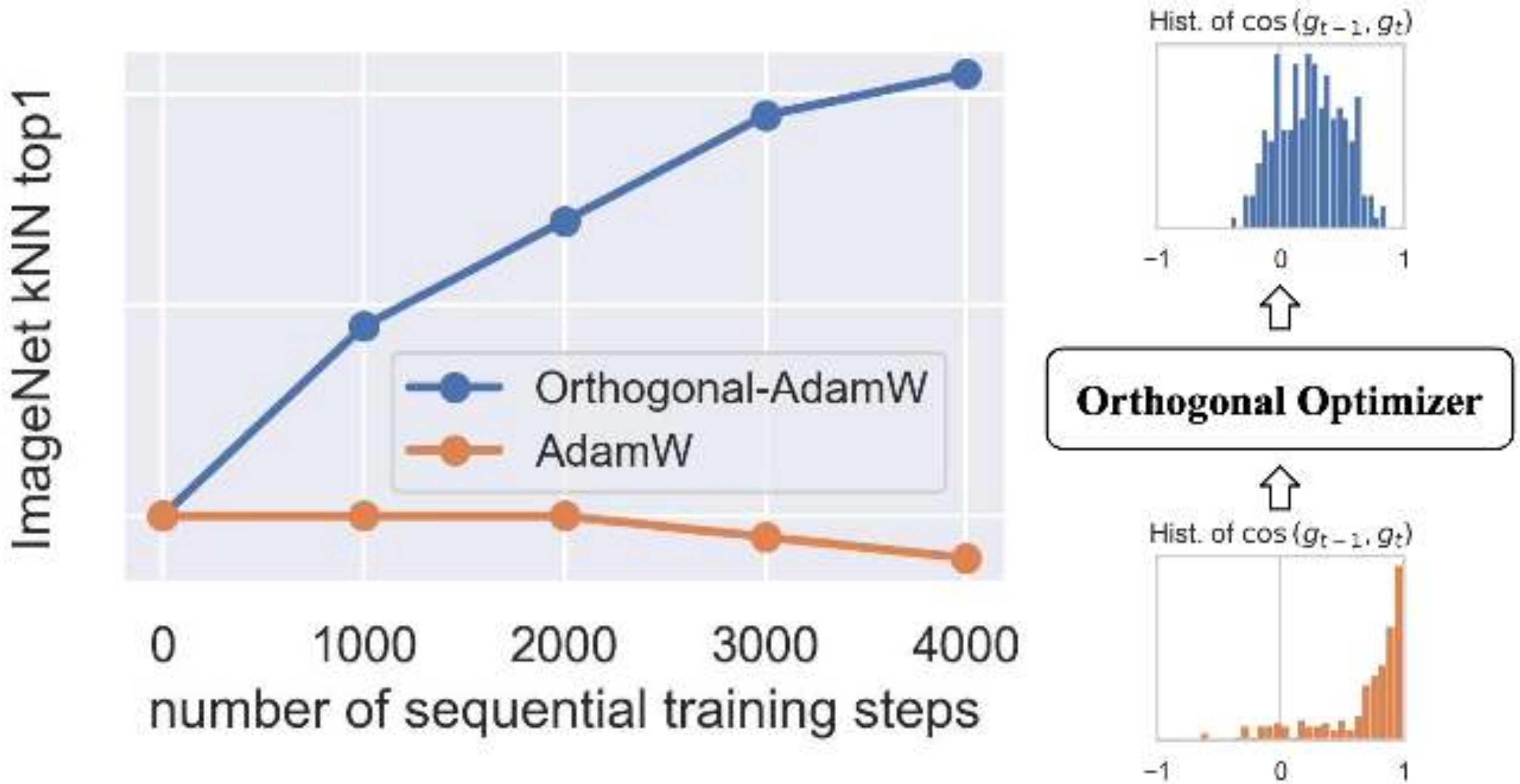
Han et al (2025). Learning from Streaming Video with Orthogonal Gradients. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

Learning from Streaming Videos with Orthogonal Gradients



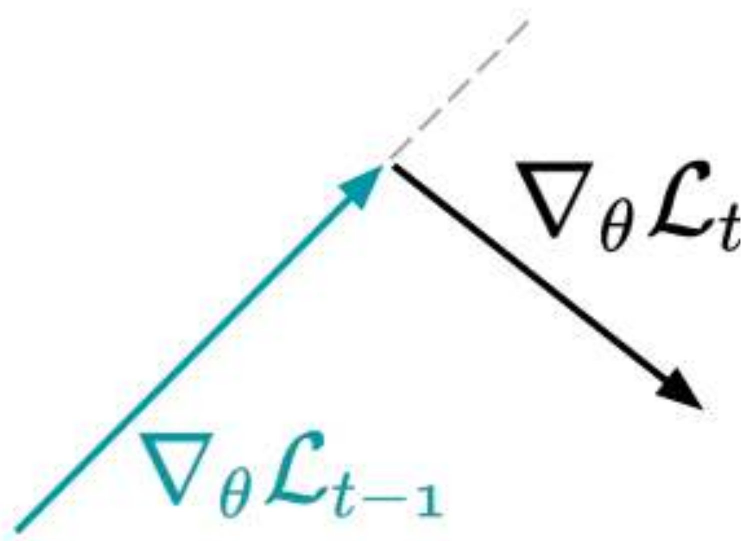
Han et al (2025). Learning from Streaming Video with Orthogonal Gradients. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

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

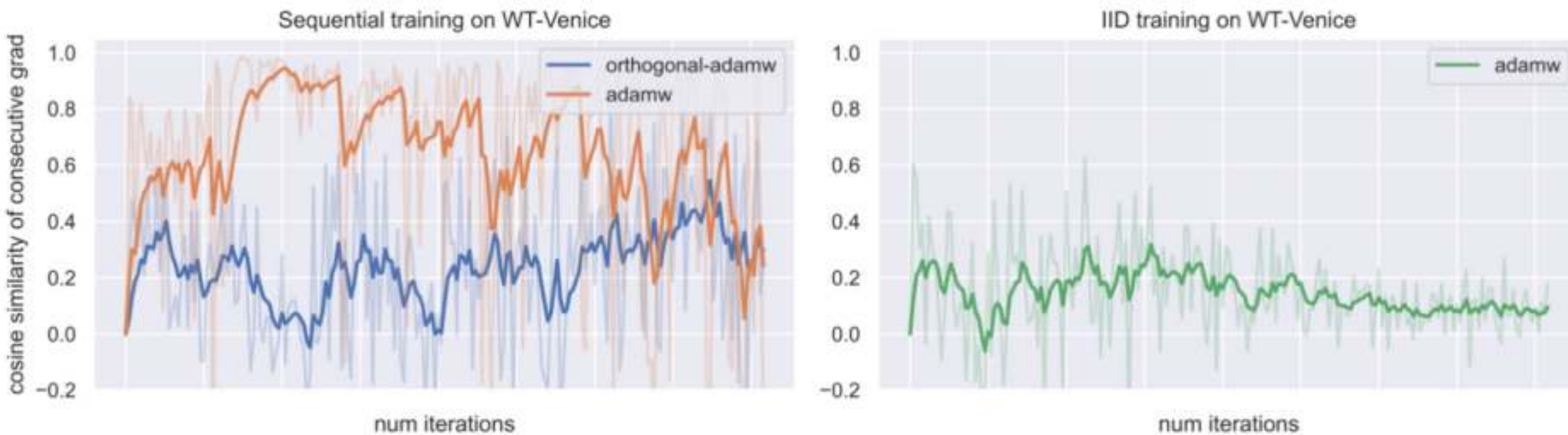
Algorithm 2 AdamW

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

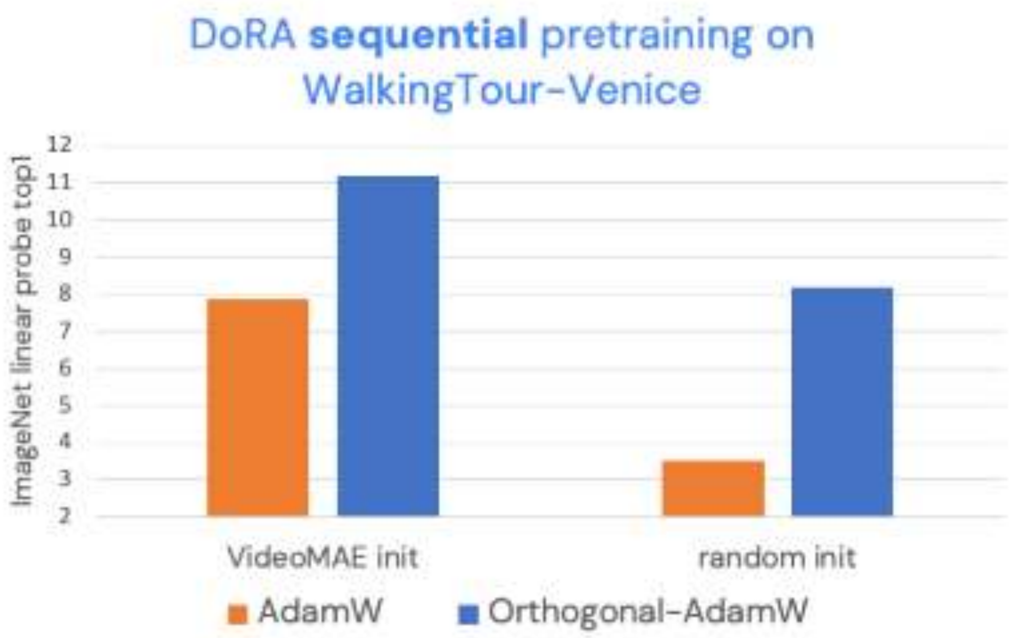
Han et al (2025). Learning from Streaming Video with Orthogonal Gradients. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

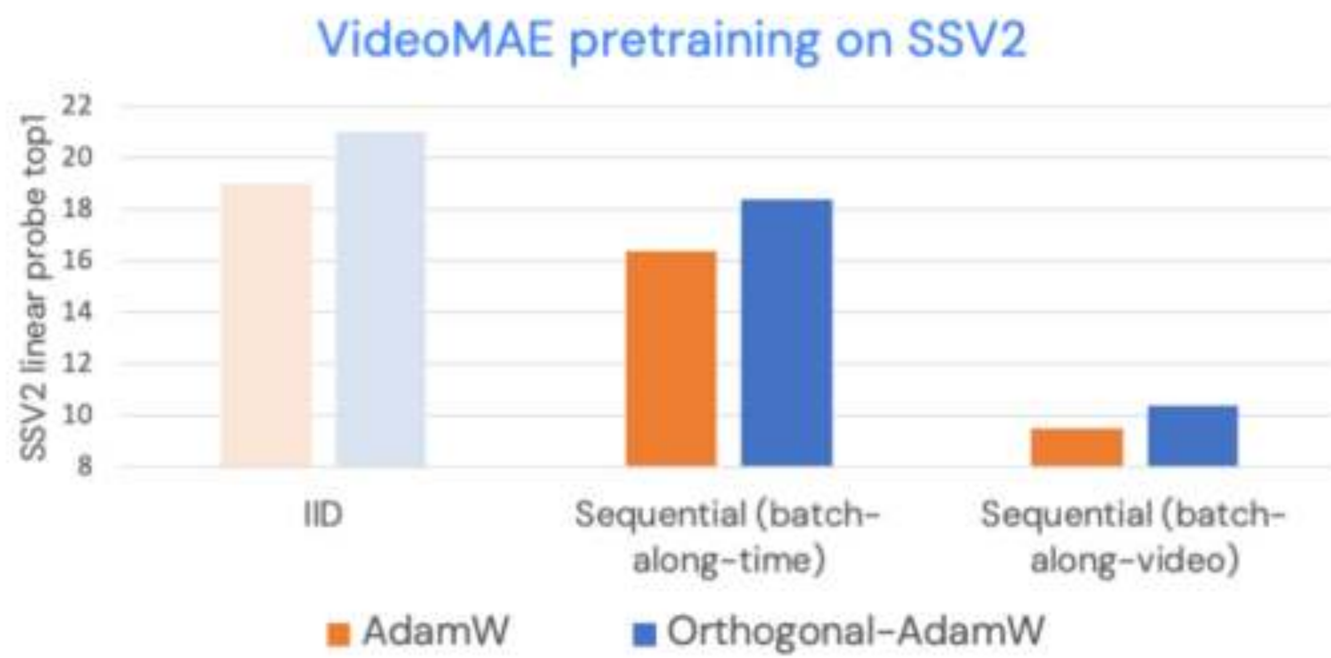
Learning from Streaming Videos with Orthogonal Gradients



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)

Dima Damen
ICVSS2025

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 at the Sphere
- Outlook into the Future of Egocentric Vision
- 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

EPIC-KITCHENS

EPIC Fields

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

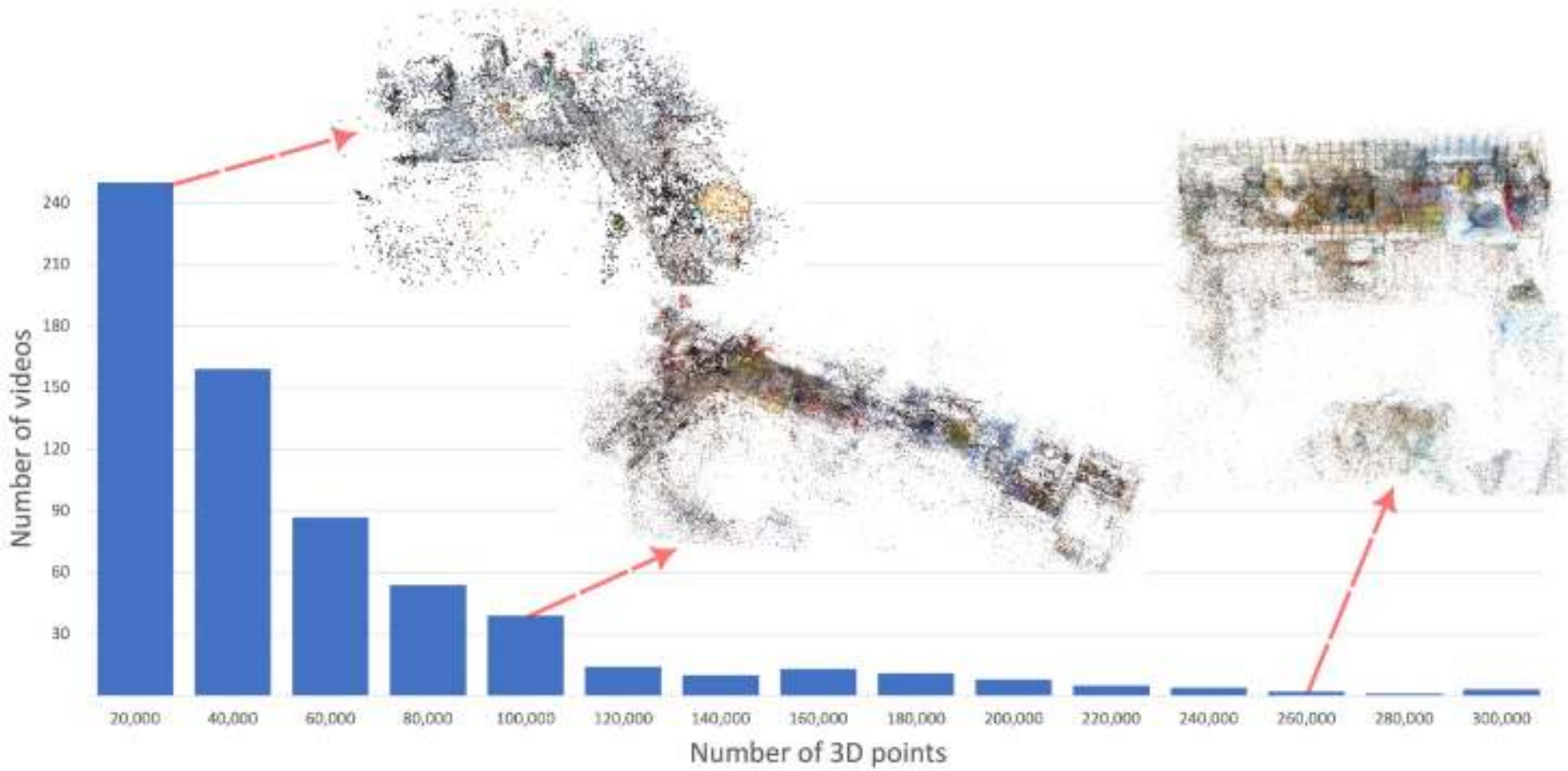


Figure 4: **Number of 3D points histogram.** The majority of our reconstructions generate less than 40,000 points that are enough to represent the kitchen. However, some reconstructions have more than 100,000, we include the point clouds for each points range showing the fine details covered by having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

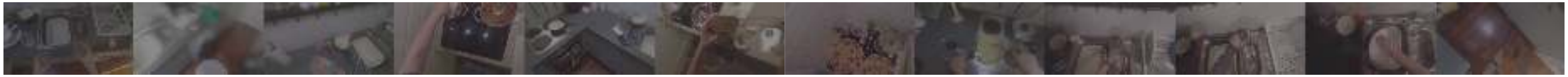
Table 1: Comparison of datasets commonly used in dynamic new-view synthesis.

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	✓

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]

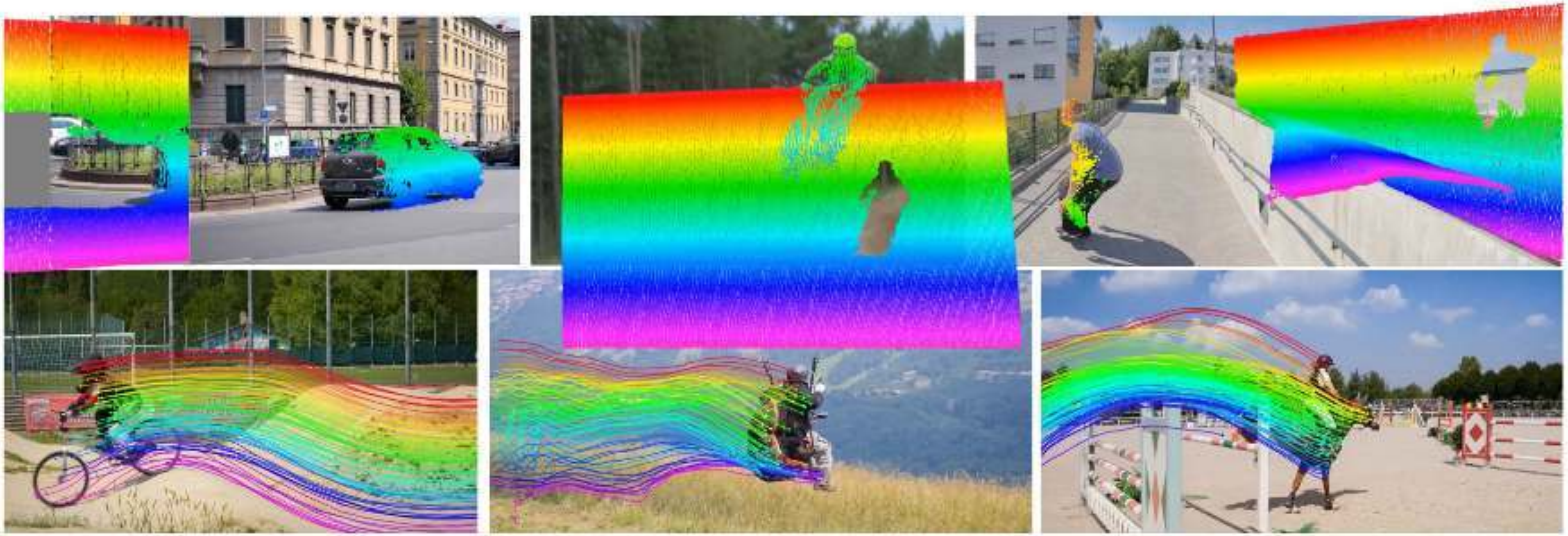
[1]University of Bristol [2]Stanford University

WACV 2025
TUCSON, ARIZONA • FEB 28 - MAR 4

What is point tracking?

with: Ahmad Darkhalil
Rhodri Guerrier
Adam W Harley

- Given: Query points in one frame
- Track these points throughout the video

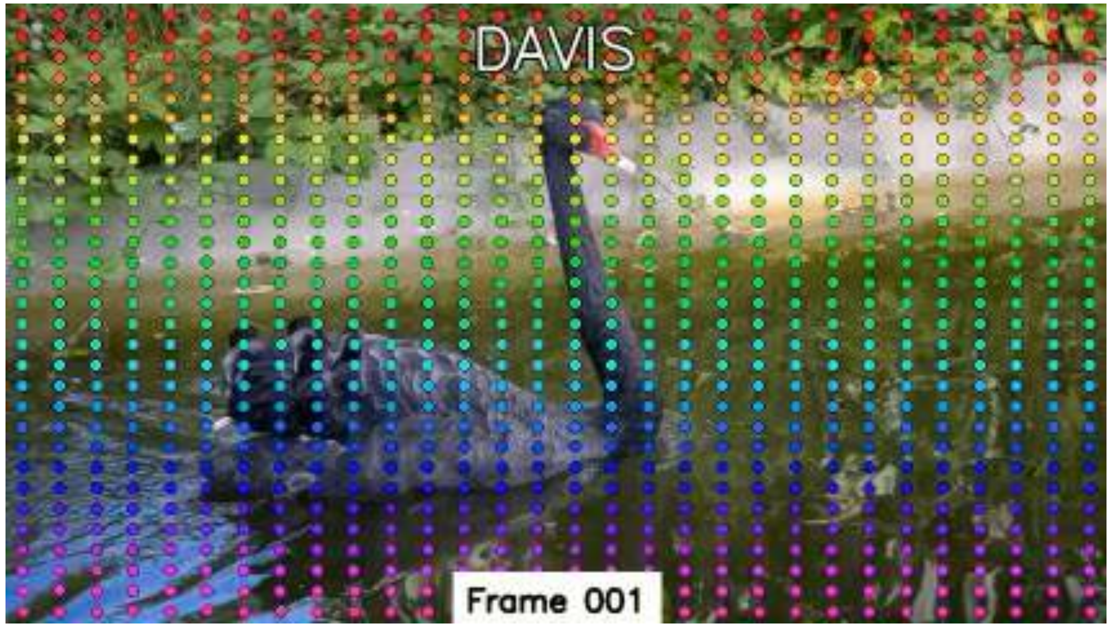


Darkhalil et al (2025). EgoPoints: Advancing Point Tracking for Egocentric Videos. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)

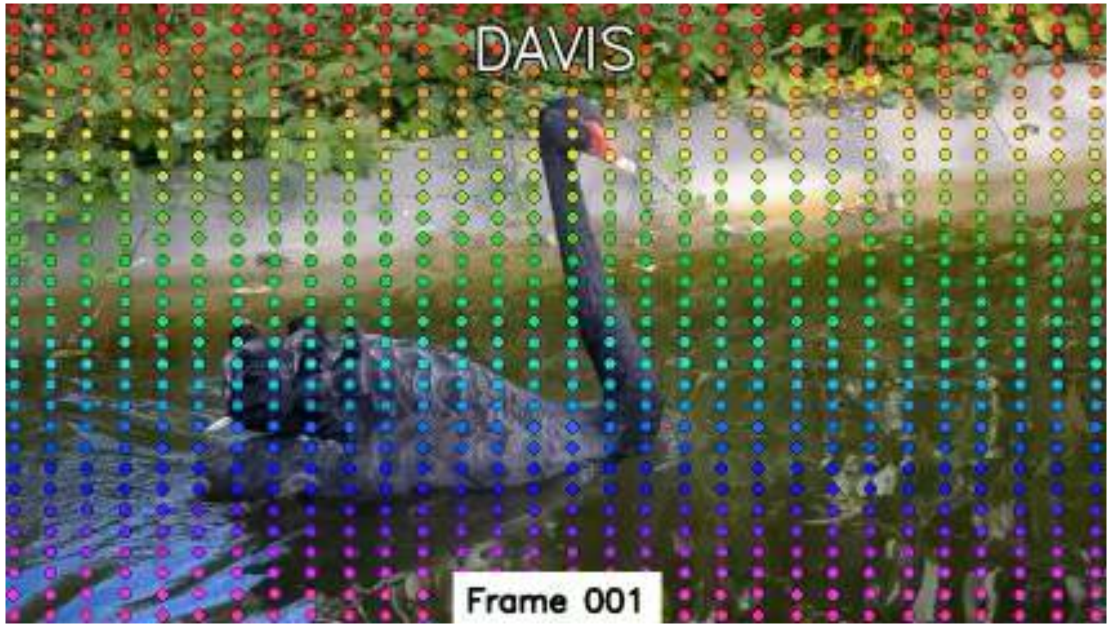
SOTA on current benchmarks

with: Ahmad Darkhalil
Rhodri Guerrier
Adam W Harley

LocoTrack



CoTracker3



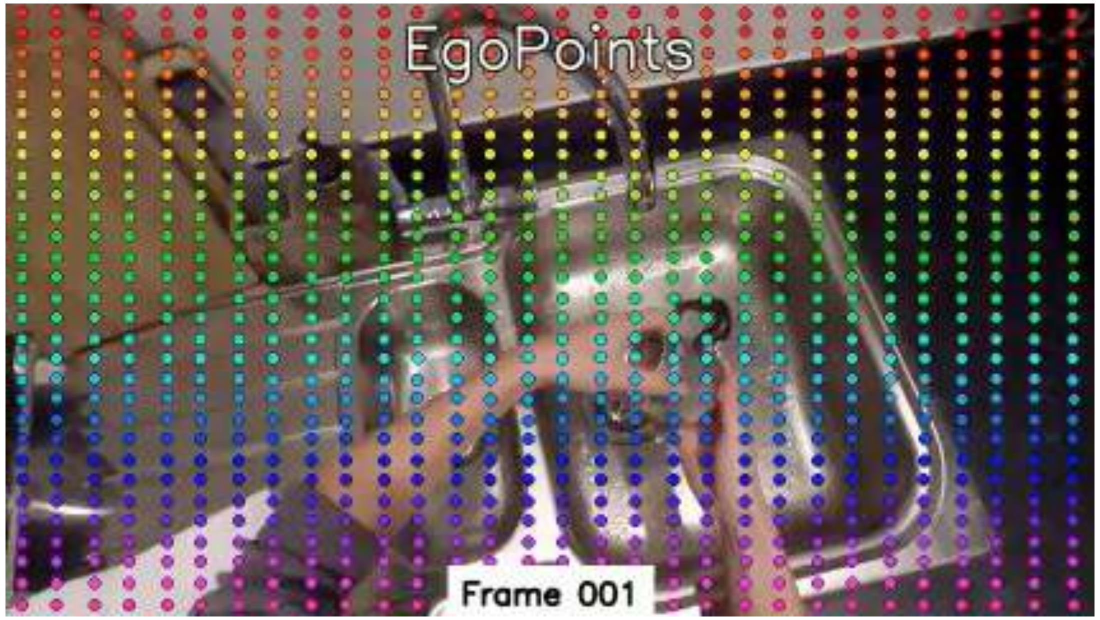
Darkhalil et al (2025). EgoPoints: Advancing Point Tracking for Egocentric Videos. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)

Current Models Struggle with Egocentric Videos

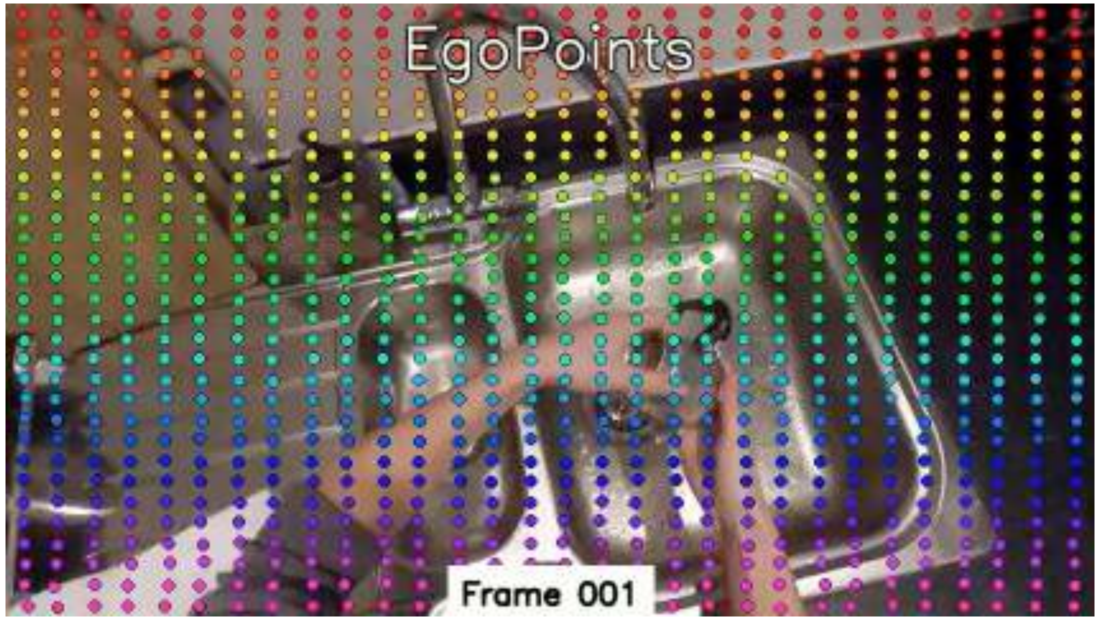
with: Ahmad Darkhalil
Rhodri Guerrier
Adam W Harley

- Head motion and motion blur
- Frequent re-identification

LocoTrack



CoTracker3



Darkhalil et al (2025). EgoPoints: Advancing Point Tracking for Egocentric Videos. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)

Main Contributions

with: Ahmad Darkhalil
Rhodri Guerrier
Adam W Harley

- Identify challenges that point trackers face in egocentric videos.
- Propose a new benchmark (EgoPoints) and new metrics to showcase these challenges
- Propose K-EPIC, a pipeline to generate semi-real training data

Darkhalil et al (2025). EgoPoints: Advancing Point Tracking for Egocentric Videos. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)

EgoPoints Annotation interface

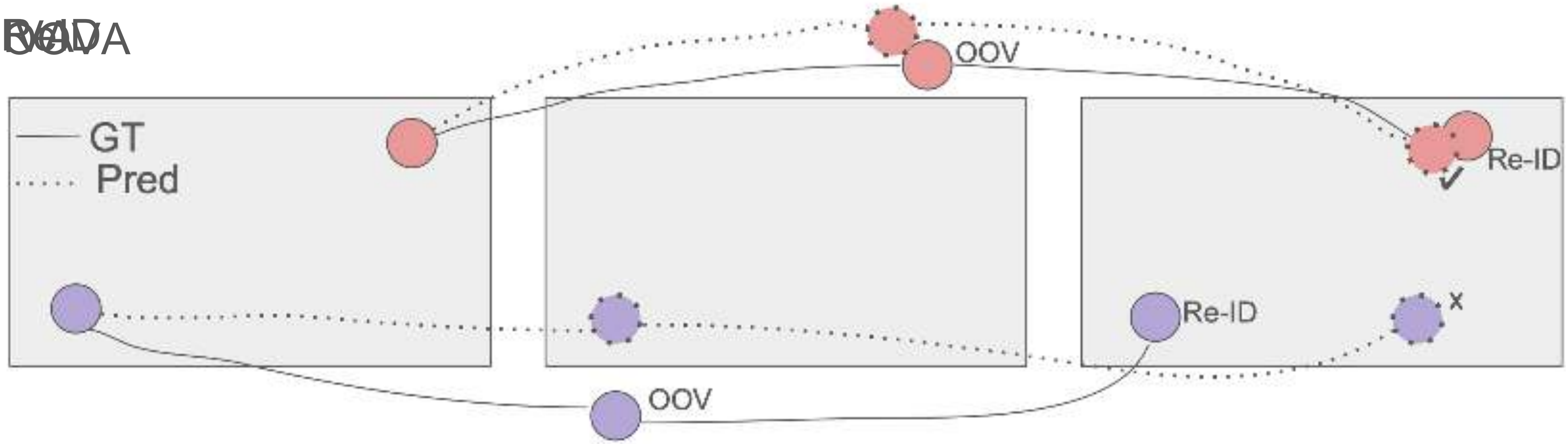
with: Ahmad Darkhalil
Rhodri Guerrier
Adam W Harley

Darkhalil et al (2025). EgoPoints: Advancing Point Tracking for Egocentric Videos. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)

EgoPoints Benchmark

with: Ahmad Darkhalil
Rhodri Guerrier
Adam W Harley

with: Ahmad Darkhalil
Rhodri Guerrier
Adam W Harley



Darkhalil et al (2025). EgoPoints: Advancing Point Tracking for Egocentric Videos. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)

EgoPoints Benchmark

with: Ahmad Darkhalil
Rhodri Guerrier
Adam W Harley

Dataset	Total Tracks	OOV Tracks	ReID Tracks	Avg. Video Length	Avg. Points/Frame
TAP-Vid-DAVIS	650	94	10	66.6	**21.7**
EgoPoints	**4703**	**875**	**593**	**511.0**	8.5

Comparisons of our annotated sequences, EgoPoints, and the commonly used TAP-Vid-DAVIS [7] point tracking benchmarks

Darkhalil et al (2025). EgoPoints: Advancing Point Tracking for Egocentric Videos. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)

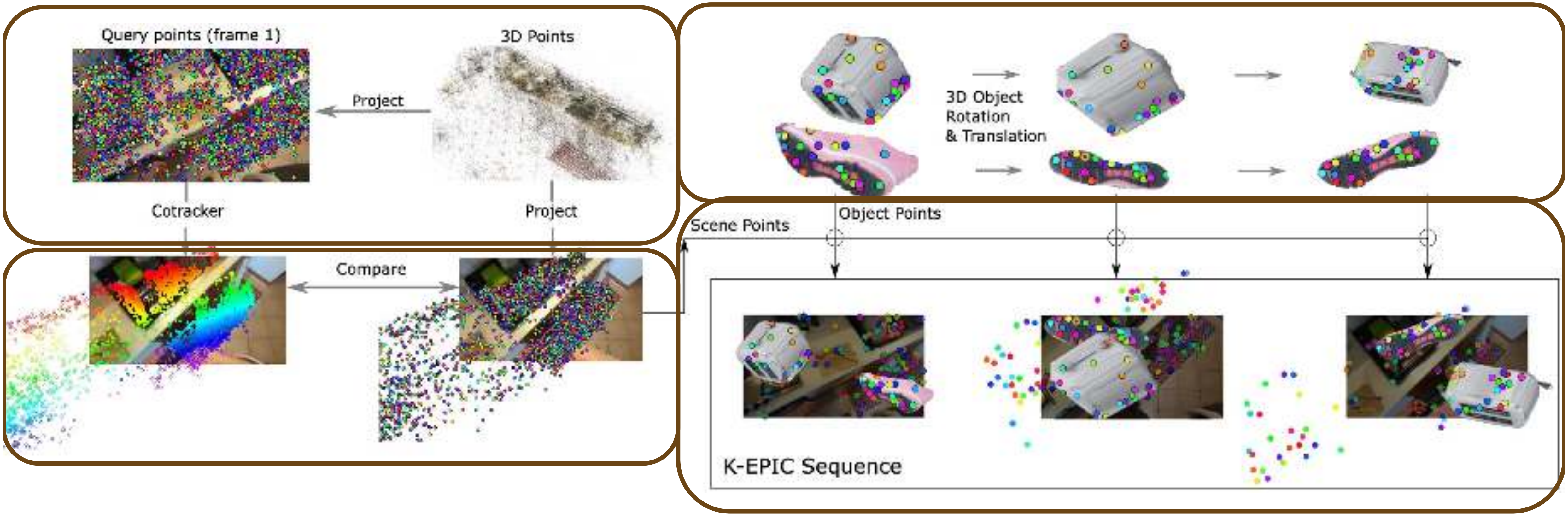
SOTA Models Struggle on EgoPoints

with: Ahmad Darkhalil
Rhodri Guerrier
Adam W Harley

	TAP-Vid-DAVIS	EgoPoints			
Model	$\delta_{avg}\uparrow$	$\delta_{avg}\uparrow$	ReID$\delta_{avg}\uparrow$	OOVA↑	IVA↑
PIPs++ [42]	64.0	36.9	14.6	50.4	89.2
CoTracker [22]	74.7	38.5	4.8	**81.4**	73.4
BootsTAPIR Online [8]	65.2	39.6	0.0	0.0	**100.0**
LocoTrack [4]	75.3	**59.4**	0.1	0.2	99.9
CoTracker v3 [21]	**77.2**	50.0	**15.0**	31.8	99.3

Pipeline of K-EPIC

with: Ahmad Darkhalil
Rhodri Guerrier
Adam W Harley



Examples from K-EPIC

with: Ahmad Darkhalil
Rhodri Guerrier
Adam W Harley

University of BRISTOL

Darkhalil et al (2025). EgoPoints: Advancing Point Tracking for Egocentric Videos. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)

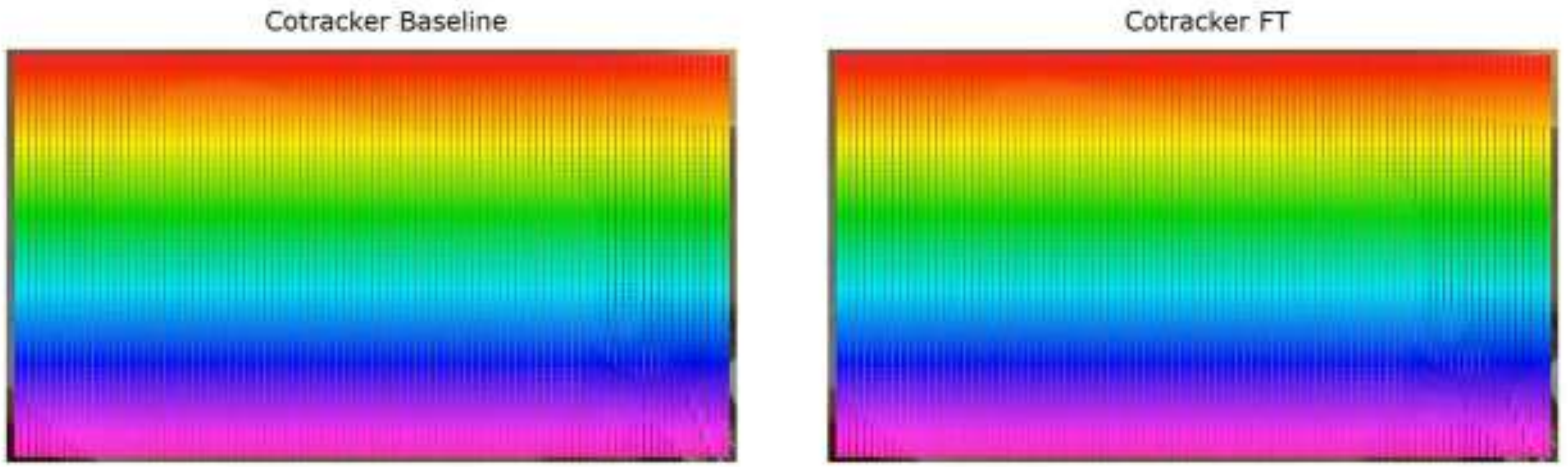
Improvements After Fine-Tuning on K-EPIC

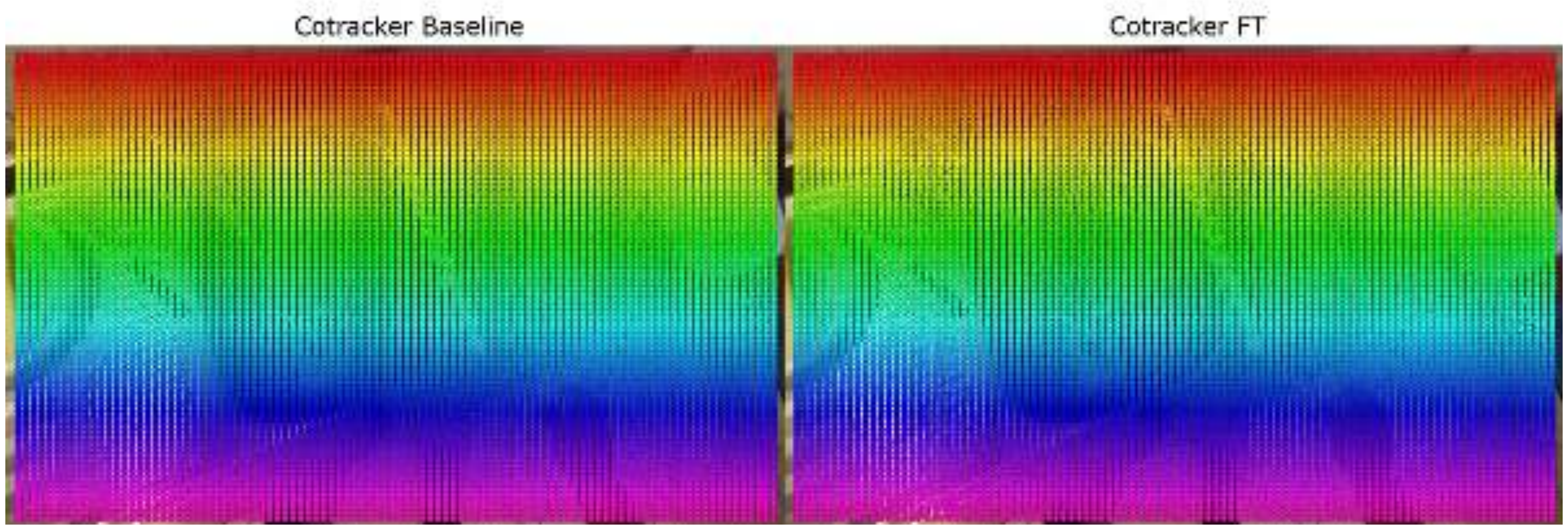
with: Ahmad Darkhalil
Rhodri Guerrier
Adam W Harley

Model	δ Metrics			Accuracy Metrics			Error
	$\delta_{avg}\uparrow$	$\delta^{*}_{avg}\uparrow$	$\text{ReID}\delta_{avg}\uparrow$	IVA↑	OOVA↑	OA↑	MTE↓
PIPs++ [42]	**36.9**	57.8	14.0	89.2	50.4	–	22.9
PIPs++ w. K-EPIC FT (scene points only)	36.3	57.8	13.0	**90.1**	**53.0**	–	22.9
PIPs++ w. K-EPIC FT (scene and object points)	36.6	**58.1**	**16.8**	89.9	52.0	–	**22.2**
CoTracker [22]	38.5	54.8	4.8	73.4	81.4	81.0	52.1
CoTracker w. K-EPIC FT (scene points only)	38.9	56.0	6.3	74.8	**85.4**	80.7	51.3
CoTracker w. K-EPIC FT (scene and object points)	**39.6**	**57.5**	**7.2**	**78.1**	82.0	**81.8**	**40.5**

Qualitative Examples of CoTracker

with: Ahmad Darkhalil
Rhodri Guerrier
Adam W Harley

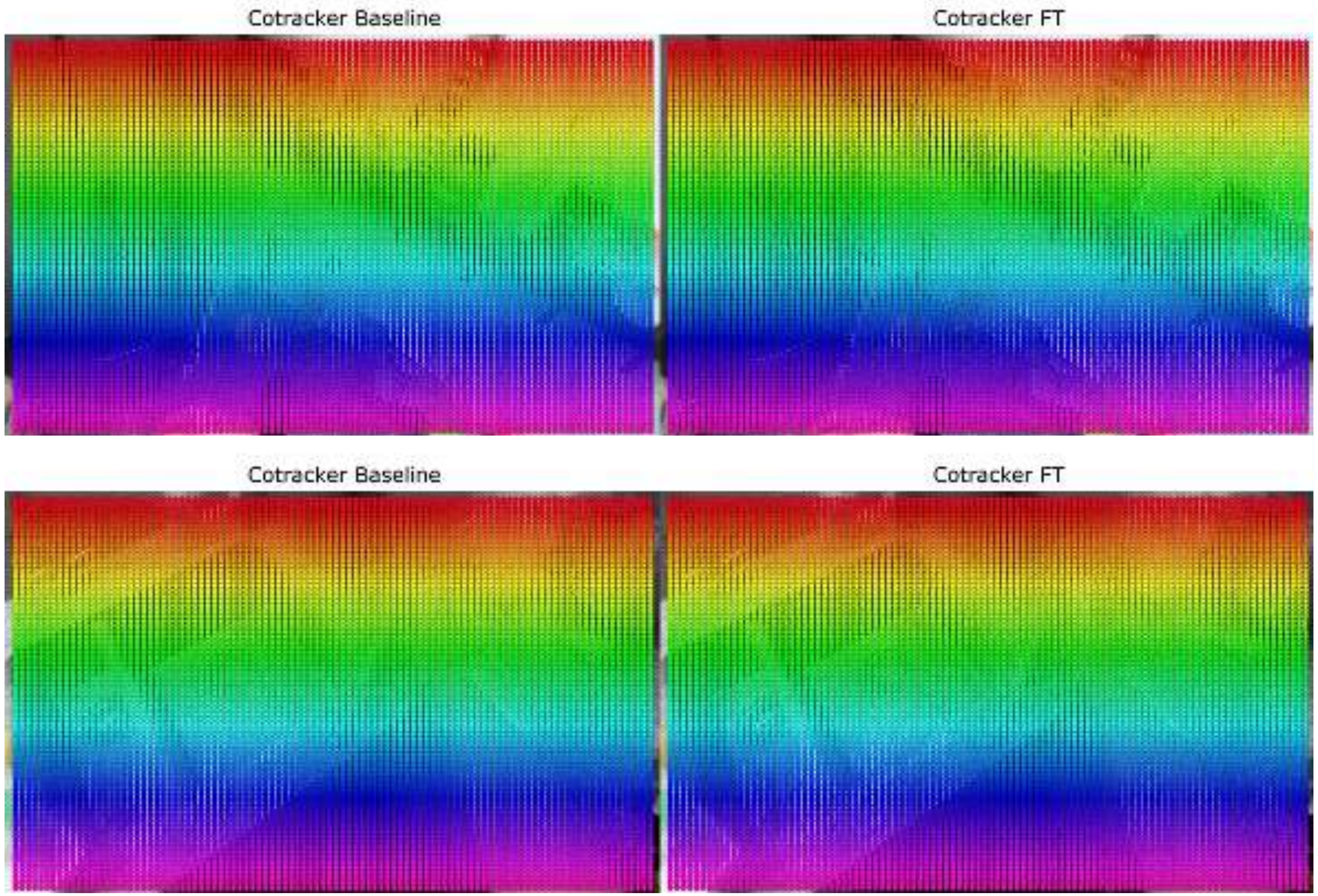




Darkhalil et al (2025). EgoPoints: Advancing Point Tracking for Egocentric Videos. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)

Qualitative Examples of CoTracker

with: Ahmad Darkhalil
Rhodri Guerrier
Adam W Harley



University of BRISTOL

Darkhalil et al (2025). EgoPoints: Advancing Point Tracking for Egocentric Videos. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)

AllTracker – newest work

Table 1. Comparison against recent point trackers and optical flow models, across nine datasets. We evaluate δ_{avg} (higher is better), using an input resolution of 384×512. The benchmarks are BADJA [3], CroHD [46], TAPVid-DAVIS [11], DriveTrack [1], EgoPoints [10], Horse10 [33], TAPVid-Kinetics [11], RGB-Stacking [28], and RoboTAP [52].

Method	Params.	Training	Bad.	Cro.	Dav.	Dri.	Ego.	Hor.	Kin.	Rgb.	Rob.	Avg.
RAFT [47]	5.26	Flow mix	23.7	29.3	48.5	44.8	41.0	27.8	64.3	82.8	72.2	48.3
SEA-RAFT [54]	19.66	Flow mix	23.9	21.9	48.7	49.4	44.0	33.1	64.3	85.7	67.6	48.7
AccFlow [55]	11.76	Flow mix	10.3	22.2	23.5	26.4	4.0	12.1	38.8	63.2	57.9	28.7
PIPs++ [59]	17.57	PointOdyssey	34.1	27.5	62.5	51.3	38.5	21.4	64.2	70.4	73.4	49.3
LocoTrack [6]	11.52	Kubric	41.4	43.1	68.0	66.5	58.4	48.9	70.0	80.3	76.9	61.5
BootsTAPIR [13]	54.70	Kubric+15M	42.7	34.9	67.9	66.9	56.8	48.8	70.6	81.0	78.2	60.9
DELTA [37]	59.17	Kubric	44.6	42.9	75.3	67.8	40.3	41.8	66.5	83.0	74.8	59.7
CoTracker2 [23]	45.43	Kubric	40.0	31.7	70.9	67.8	43.2	33.9	65.8	73.4	73	55.5
CoTracker3-Kub [25]	25.39	Kubric	47.5	**48.9**	**77.4**	**69.8**	58.0	47.5	70.6	83.4	77.2	64.5
CoTracker3 [25]	25.39	Kubric+15k	48.3	44.5	77.1	**69.8**	60.4	47.1	71.8	84.2	81.6	65.0
AllTracker-Tiny-Kub	6.29	Kubric	45.4	39.6	73.7	65.1	55.9	45.2	70.6	86.1	79.3	62.3
AllTracker-Tiny	6.29	Kubric+mix	47.5	39.8	74.3	63.9	58.3	45.5	71.5	88.1	80.7	63.3
AllTracker-Kub	16.48	Kubric	46.4	42.3	75.2	66.1	60.3	**49.0**	71.3	**90.1**	82.2	64.8
AllTracker	16.48	Kubric+mix	**51.5**	44.0	76.3	65.8	**62.5**	**49.0**	**72.3**	90.0	**83.4**	**66.1**

Harley et al (2025). AllTracker: Efficient Dense Point Tracking at High Resolution. ArXiv

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

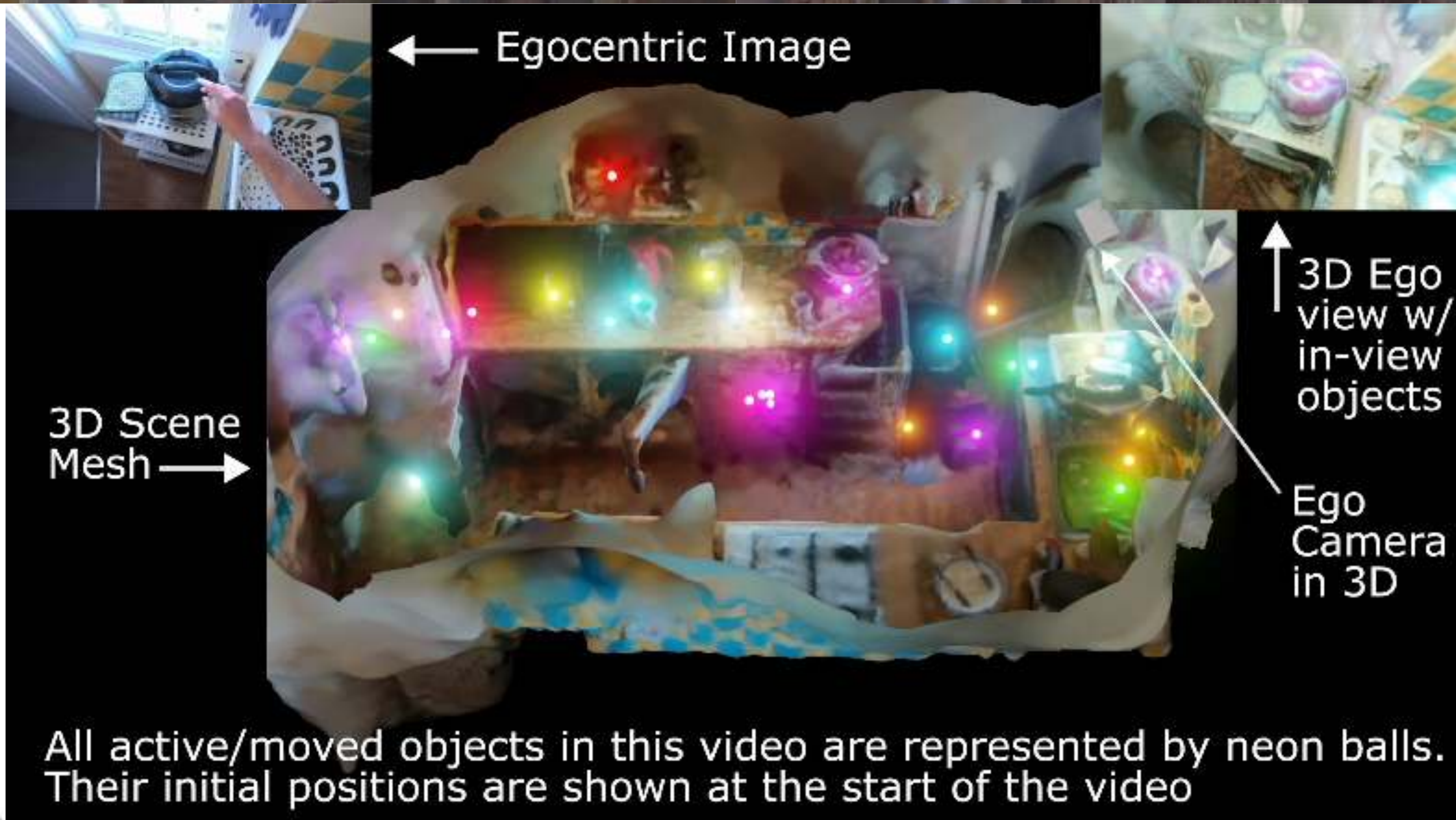
Politecnico di Torino

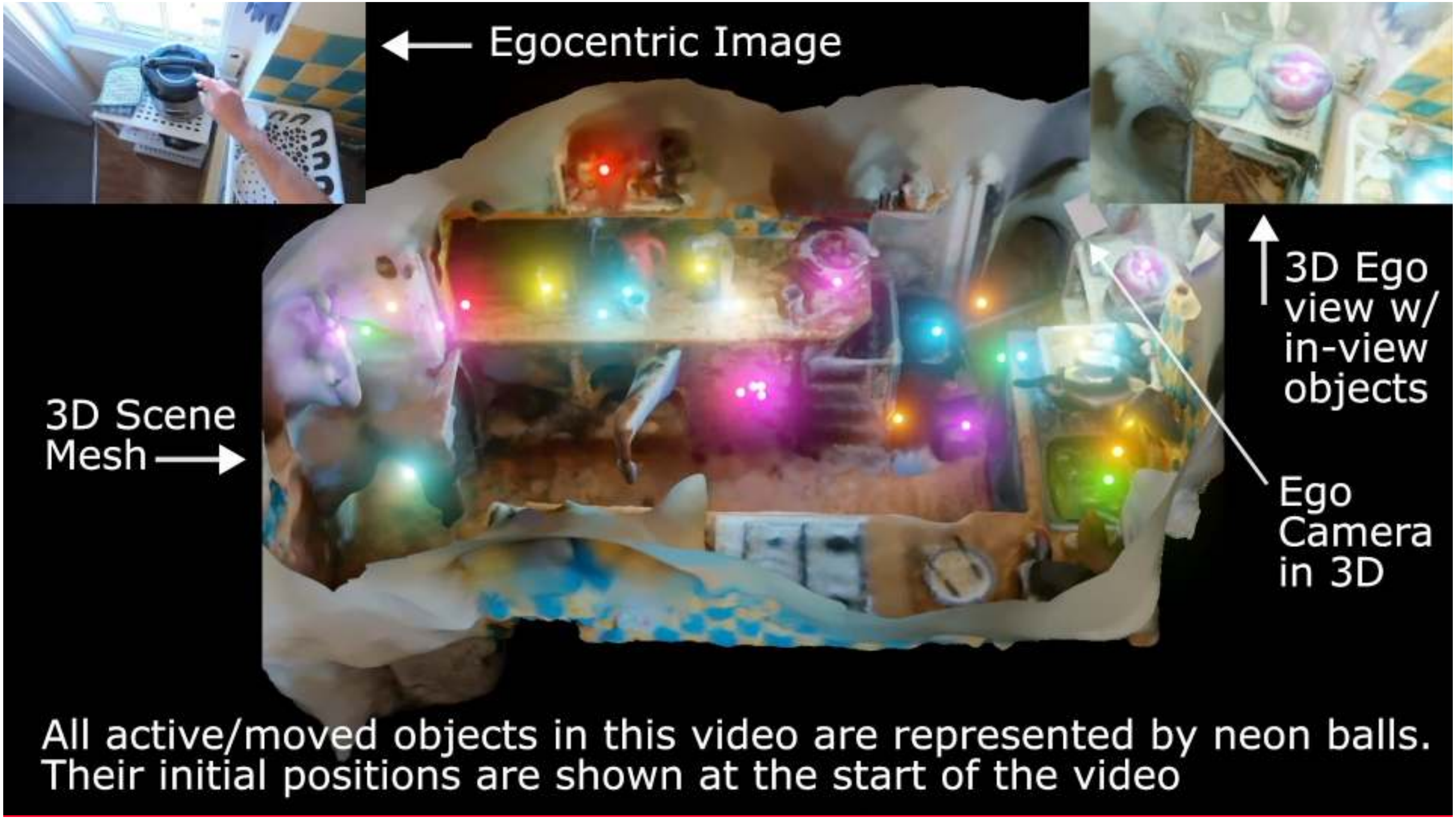
Berkeley UNIVERSITY OF CALIFORNIA

University of BRISTOL

Out of Sight, not Out of Mind

with: Chiara Plizzari
Toby Perrett
Shubham Goel
Angjoo Kanazawa

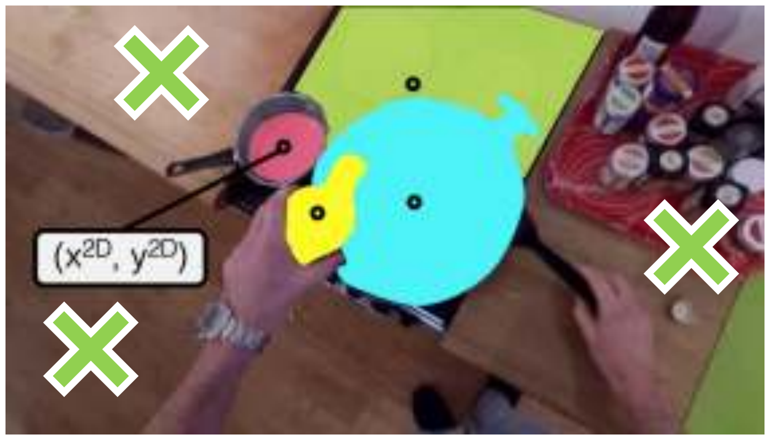


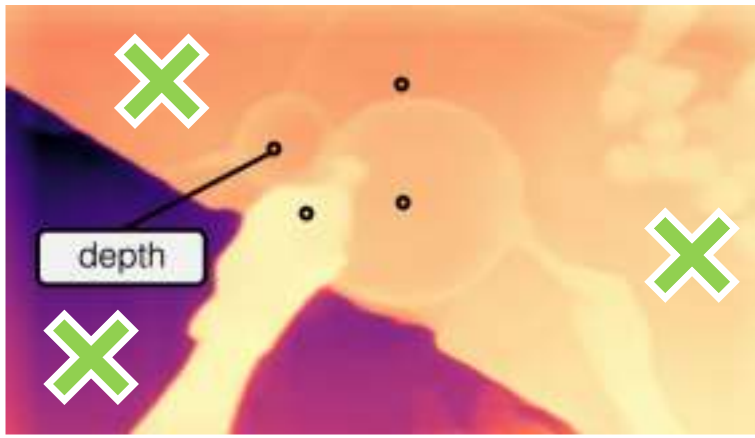

Egocentric Image
3D Ego view w/ in-view objects
3D Scene Mesh
Ego Camera in 3D
All active/moved objects in this video are represented by neon balls.
Their initial positions are shown at the start of the video

Out of Sight, not Out of Mind

with: Chiara Plizzari
Toby Perrett
Shubham Goel
Angjoo Kanazawa

Lift | Match | Keep





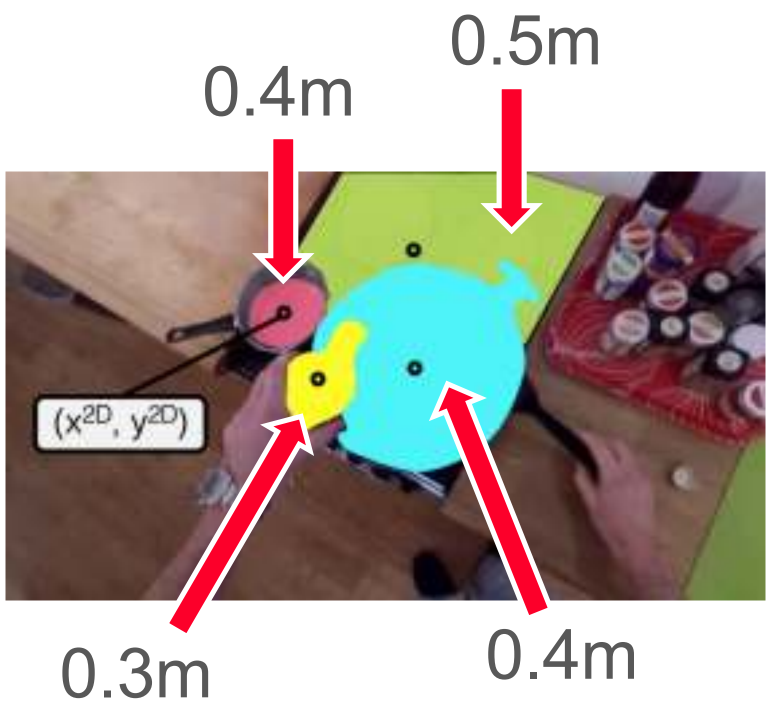
0.0 … 1.0

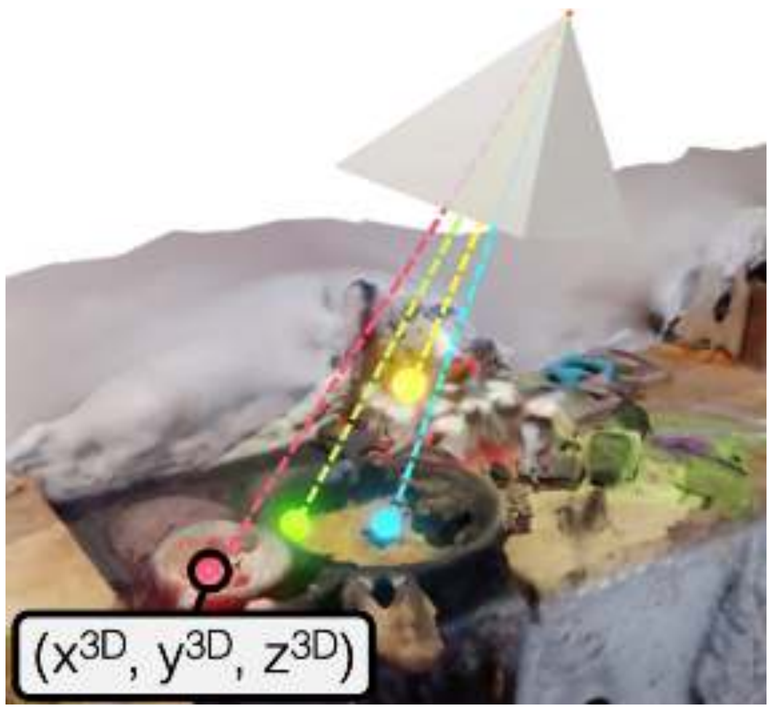
0.3m … 1.8m

Out of Sight, not Out of Mind

with: Chiara Plizzari
Toby Perrett
Shubham Goel
Angjoo Kanazawa

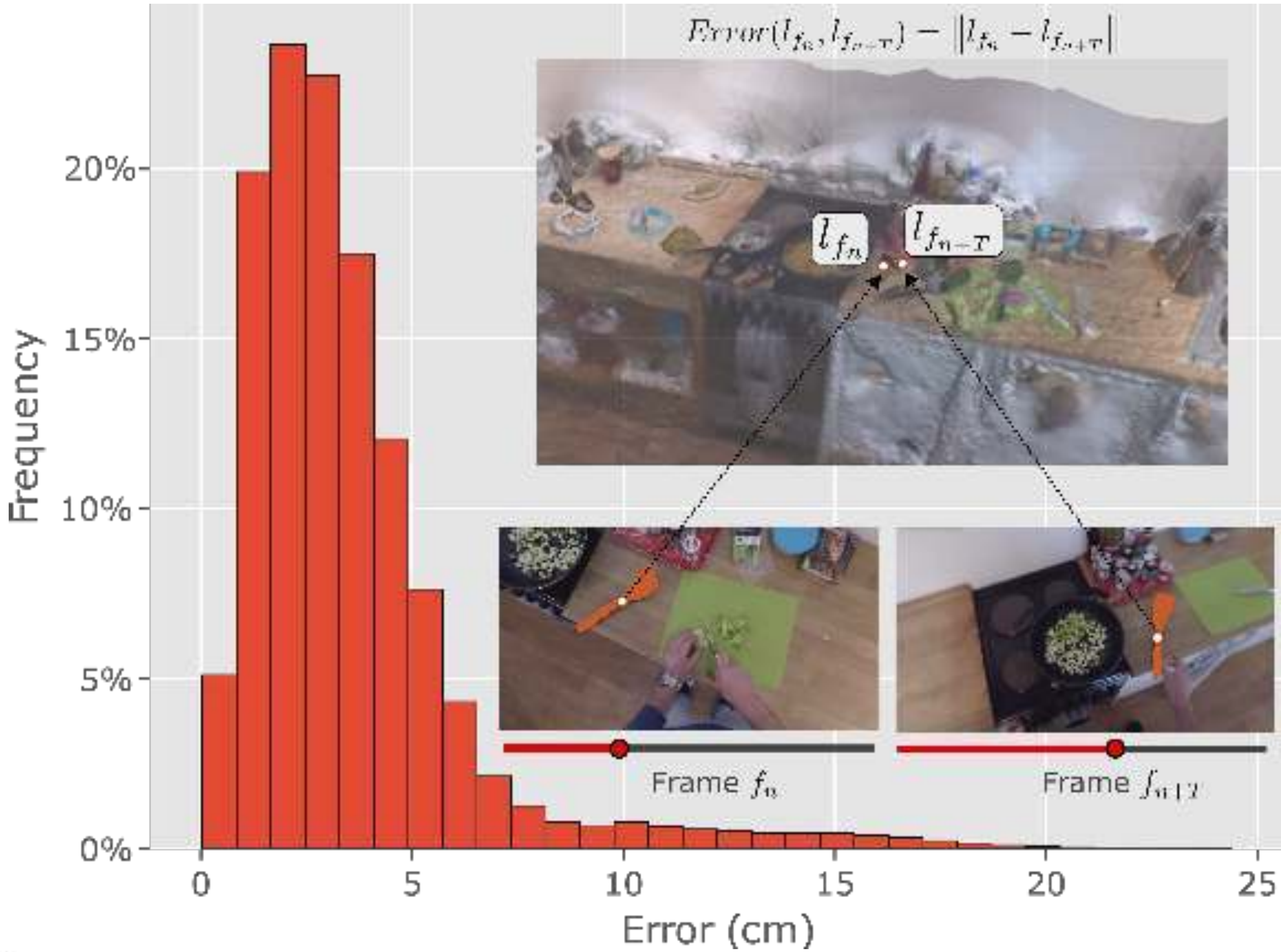
Lift → Match → Keep





Out of Sight, not Out of Mind

with: Chiara Plizzari Shubham Goel
Toby Perrett Angjoo Kanazawa



University of BRISTOL

Plizzari et al (2025). Spatial Cognition from Egocentric Video: Out of Sight, Not Out of Mind. 3DV

Out of Sight, not Out of Mind

with: Chiara Plizzari
Toby Perrett
Shubham Goel
Angjoo Kanazawa

Lift → Match → Keep

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

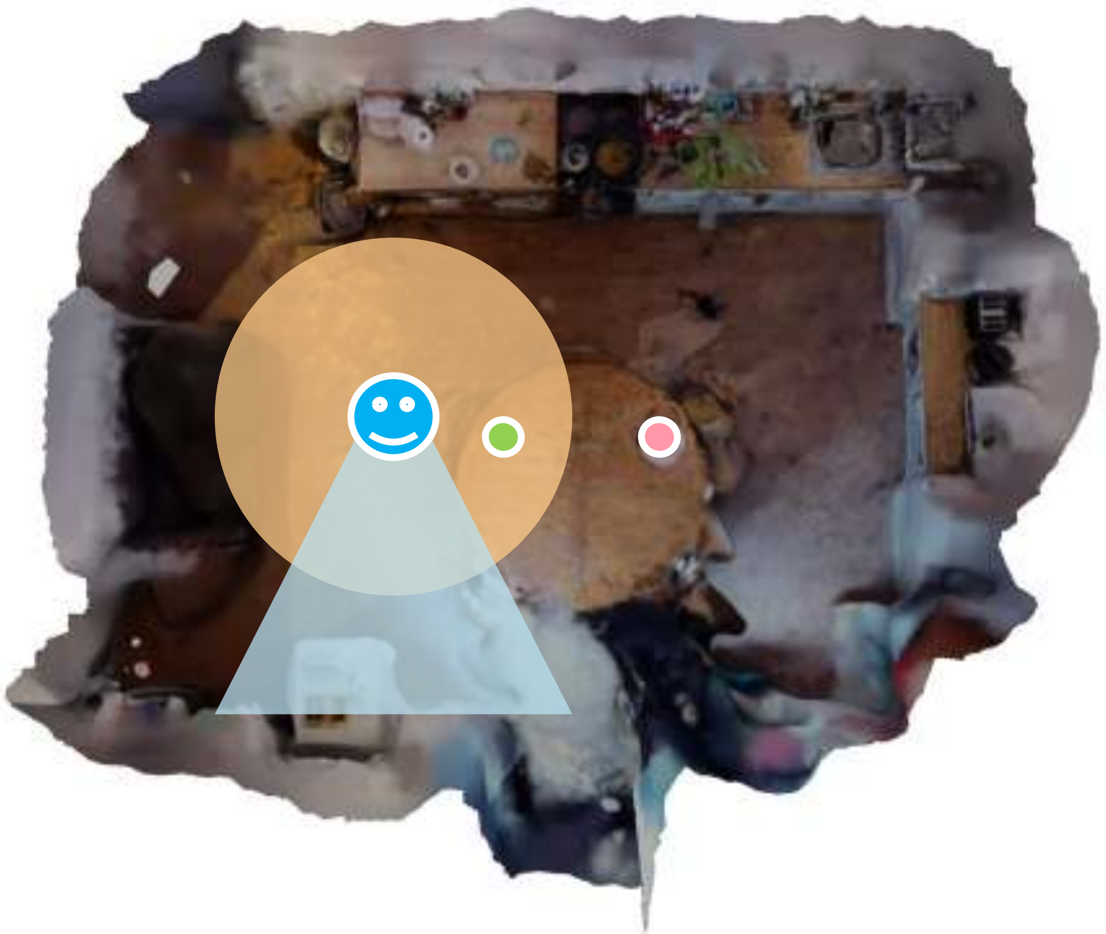
- In-View vs Out-of-View
- In-Sight vs Out-of-Sight (Occluded)
- Within-Reach vs Out-of-Reach (defining the camera wearer's near space)

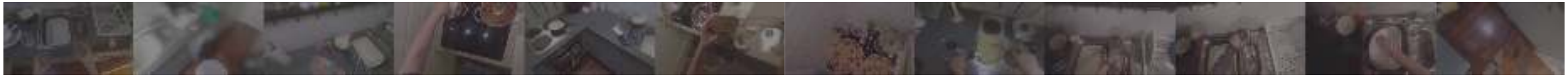
Out of Sight, not Out of Mind

with: Chiara Plizzari
Toby Perrett
Shubham Goel
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

Ground-Truth??

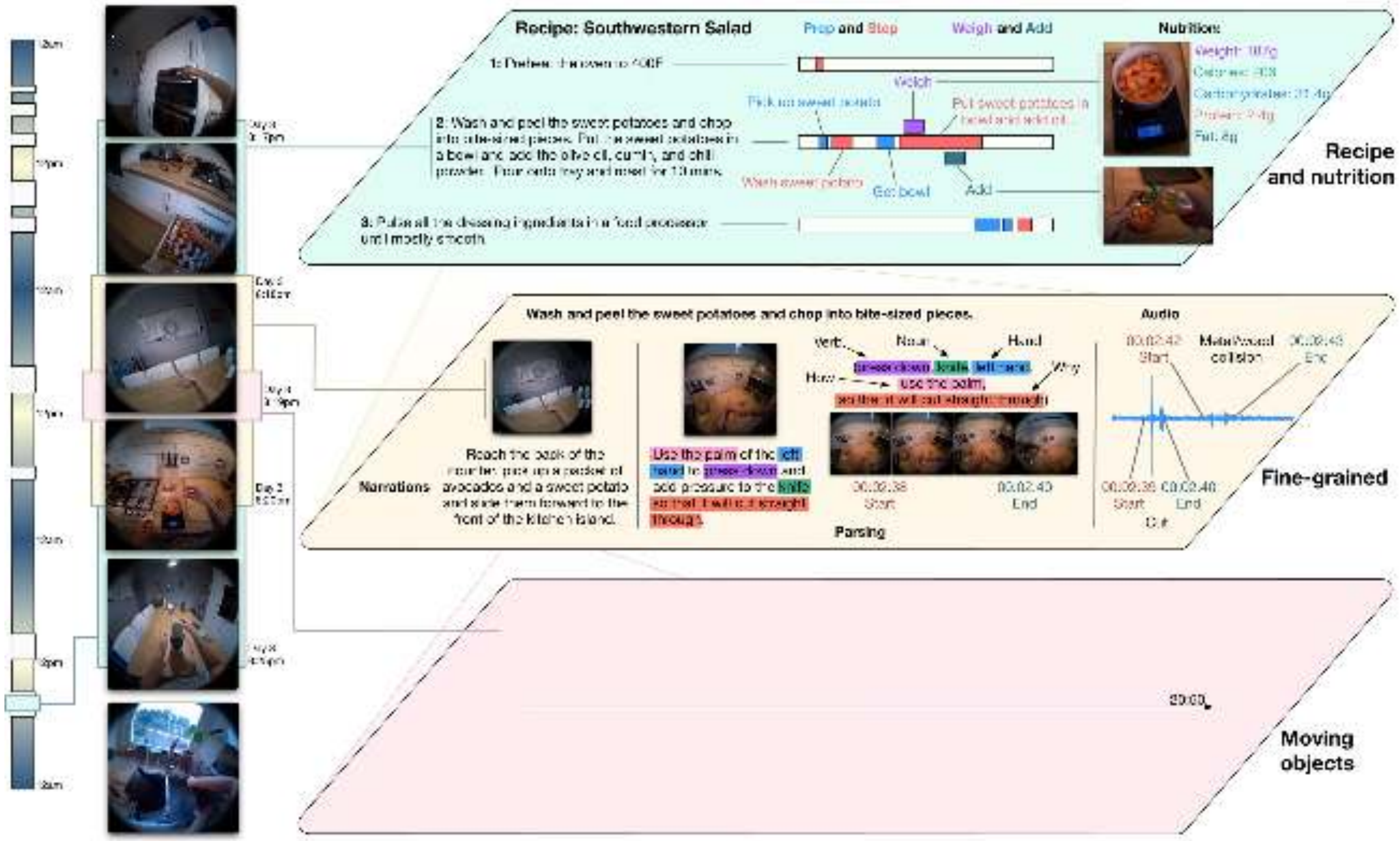
Chiara Plizzari   Shubham   oby Perrett   Jacob Chalk

Ang   a   Dima Damen

http://dimadamen.github.io/OSNOM

Politecnico di Torino   Berkeley UNIVERSITY OF CALIFORNIA   University of BRISTOL

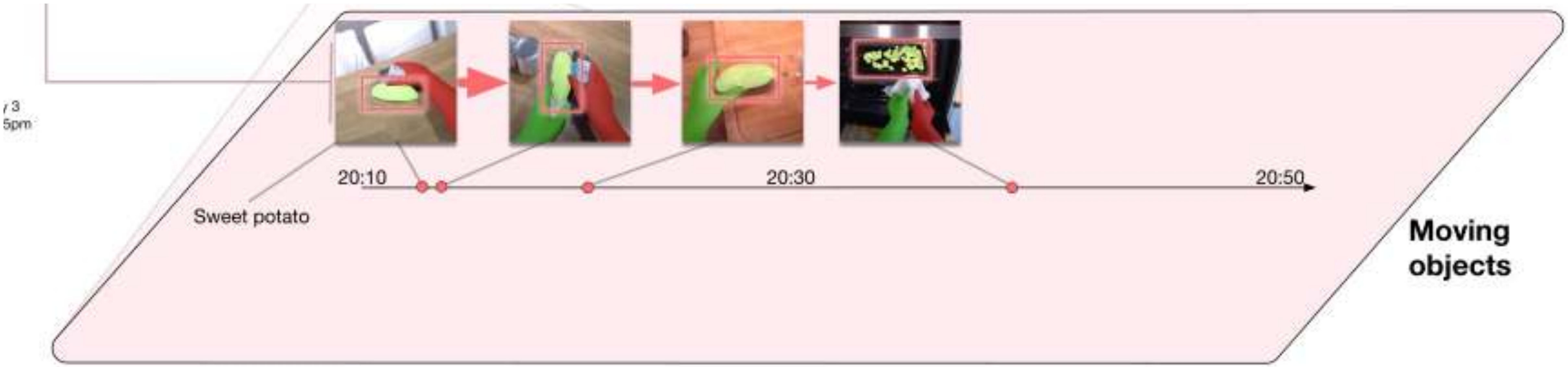
HD-EPIC



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

Dima Damen
ICVSS2025

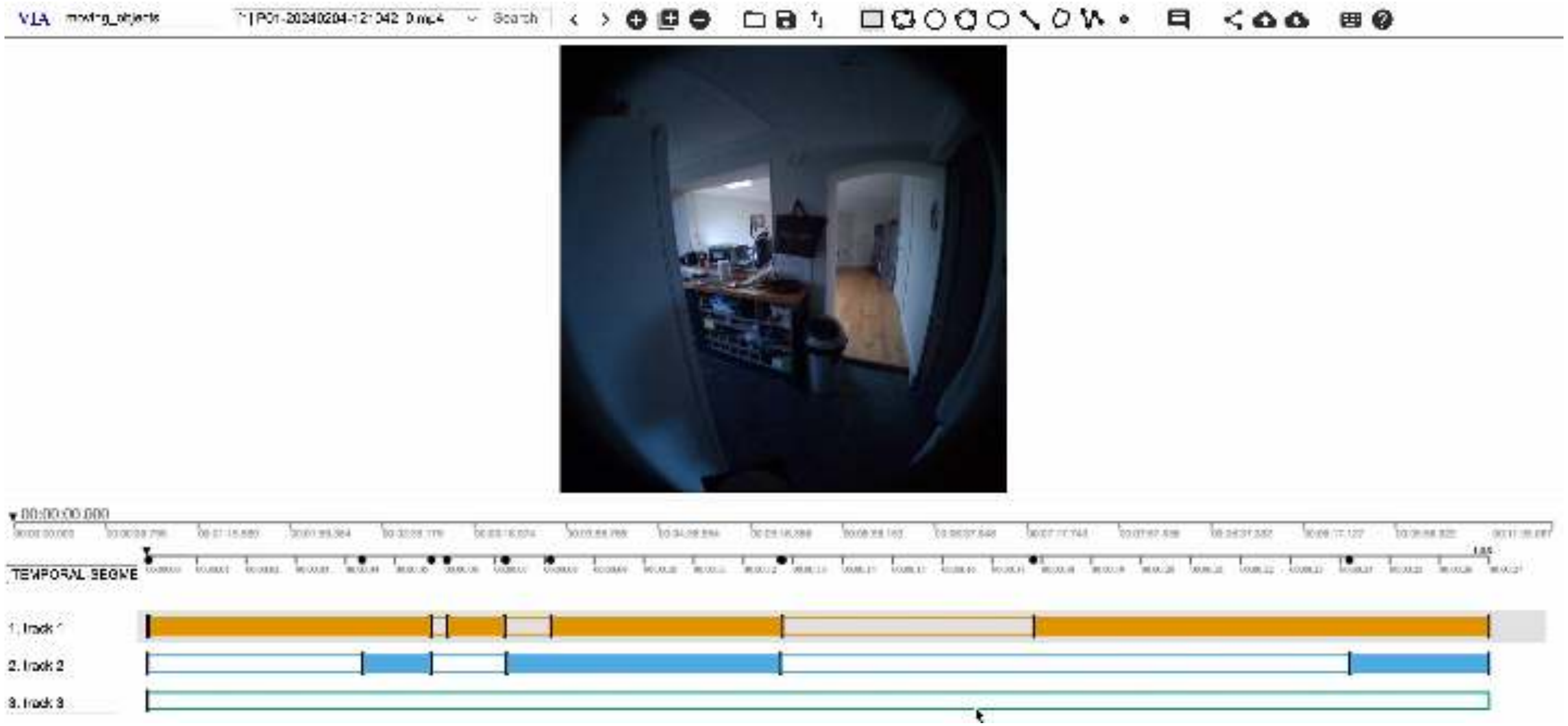
HD-EPIC



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

HD-EPIC

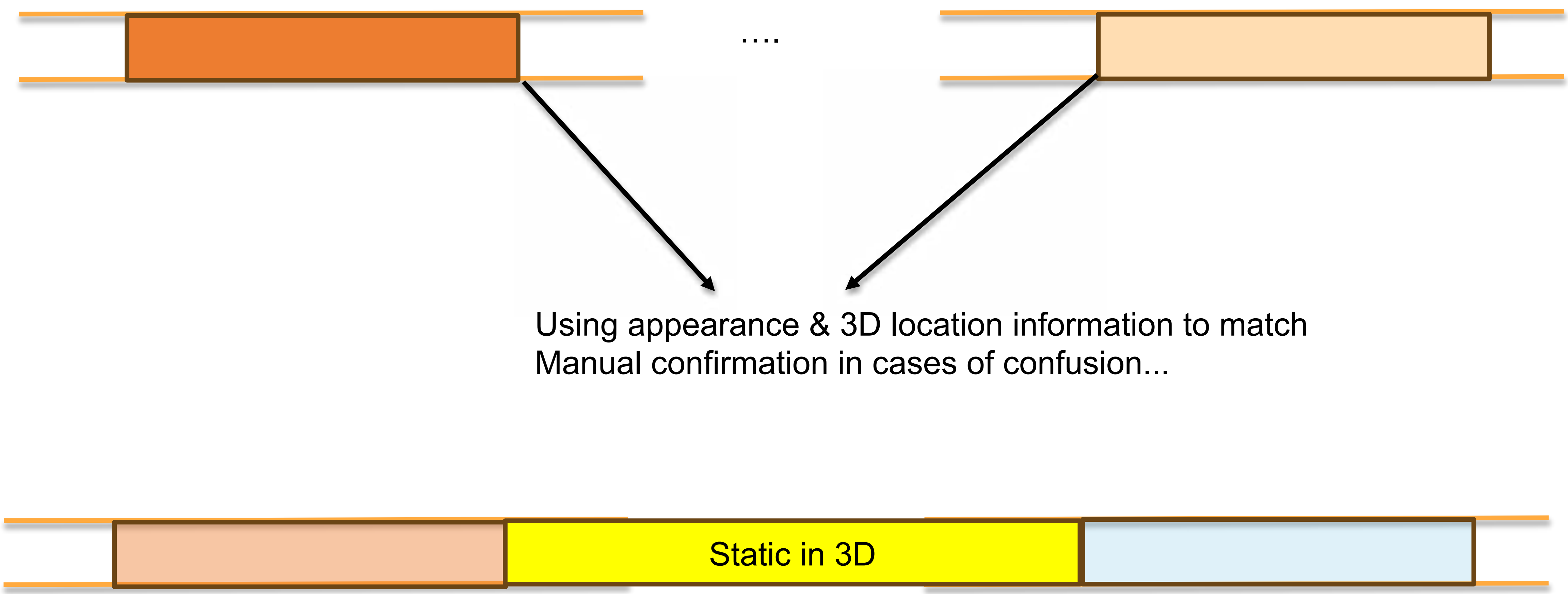
- How to minimize the annotations for tracking objects…



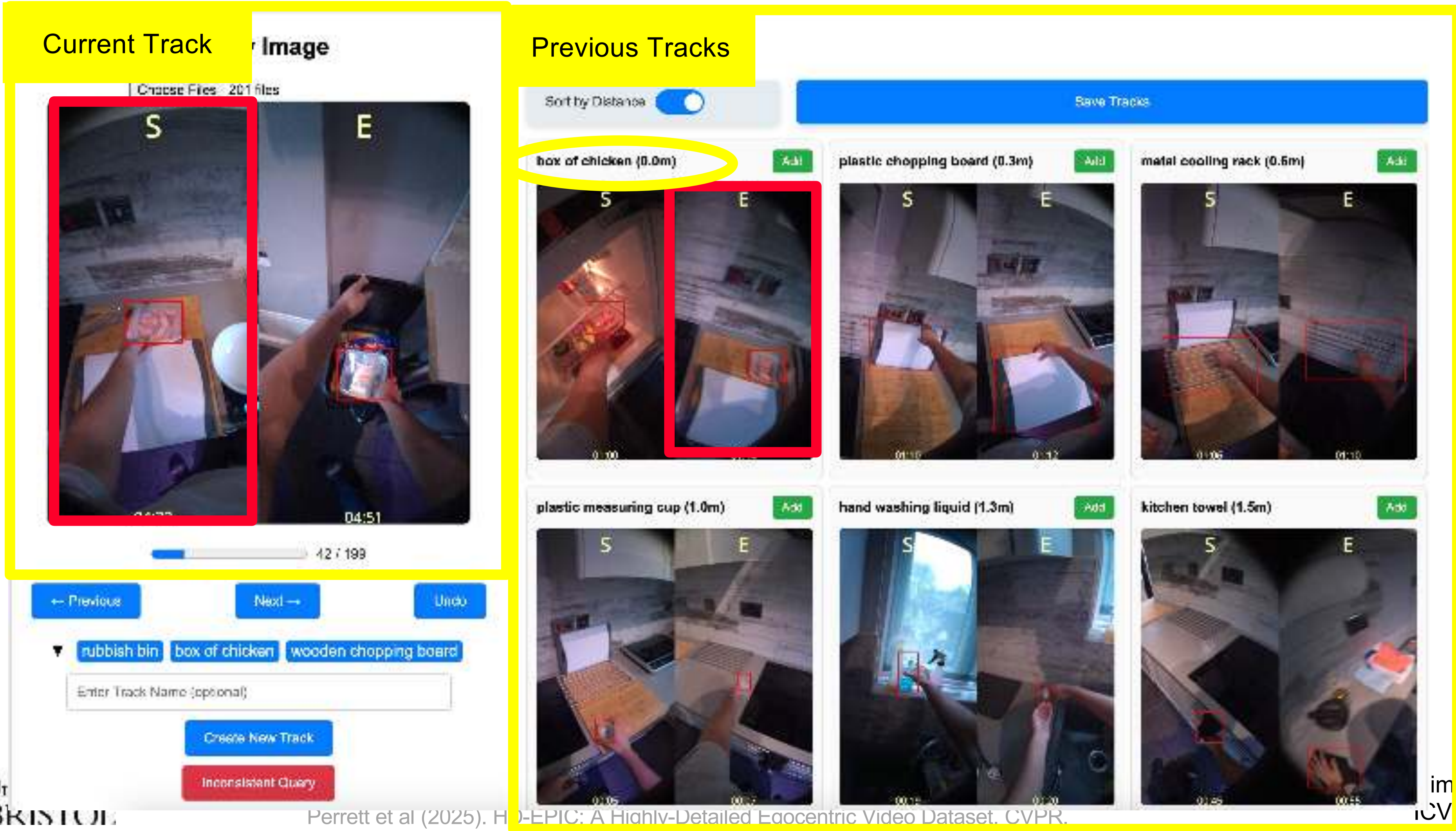
Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

HD-EPIC

- How to minimize the annotations for tracking objects…

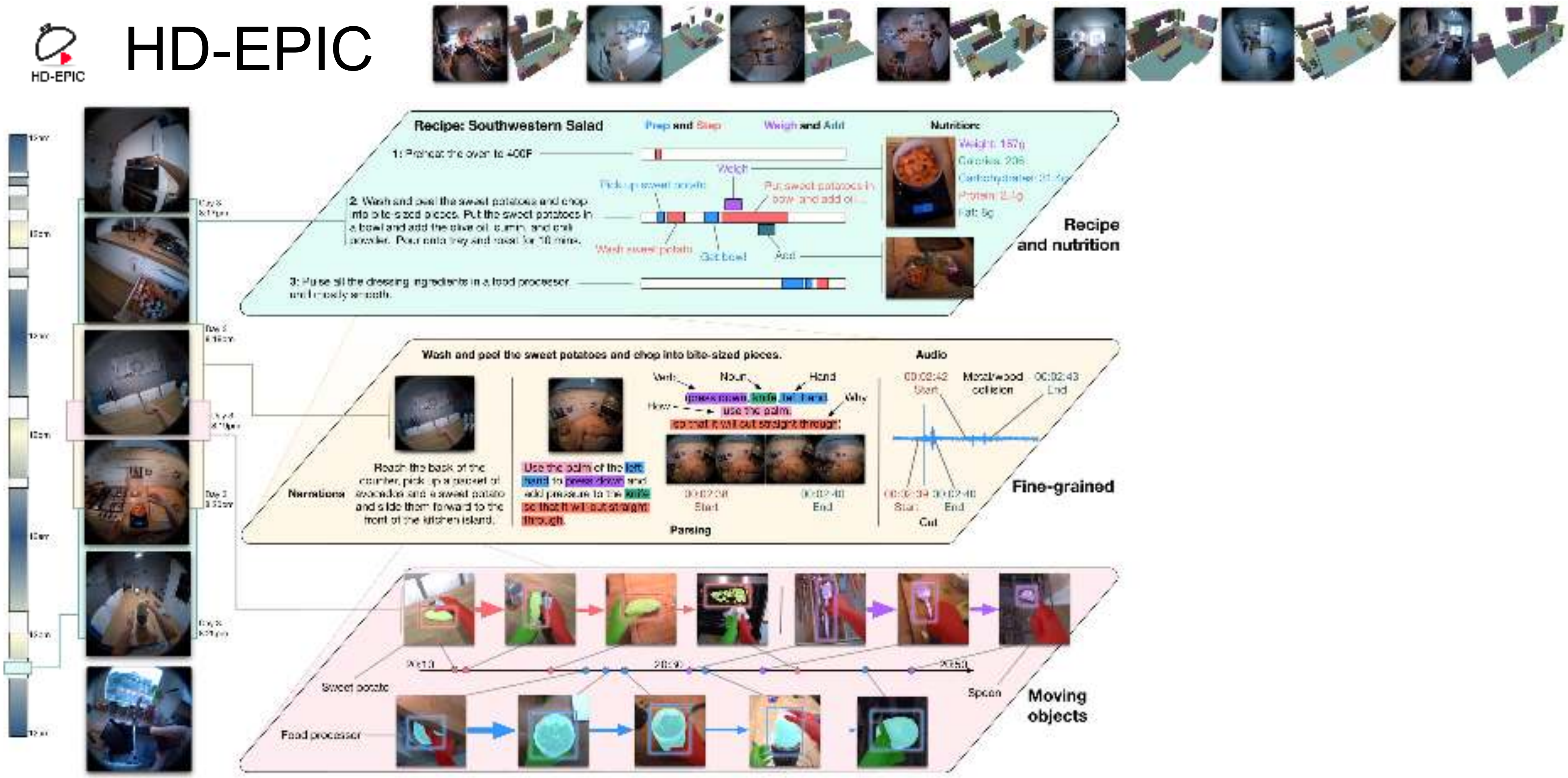


HD-EPIC



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

HD-EPIC



University of BRISTOL

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

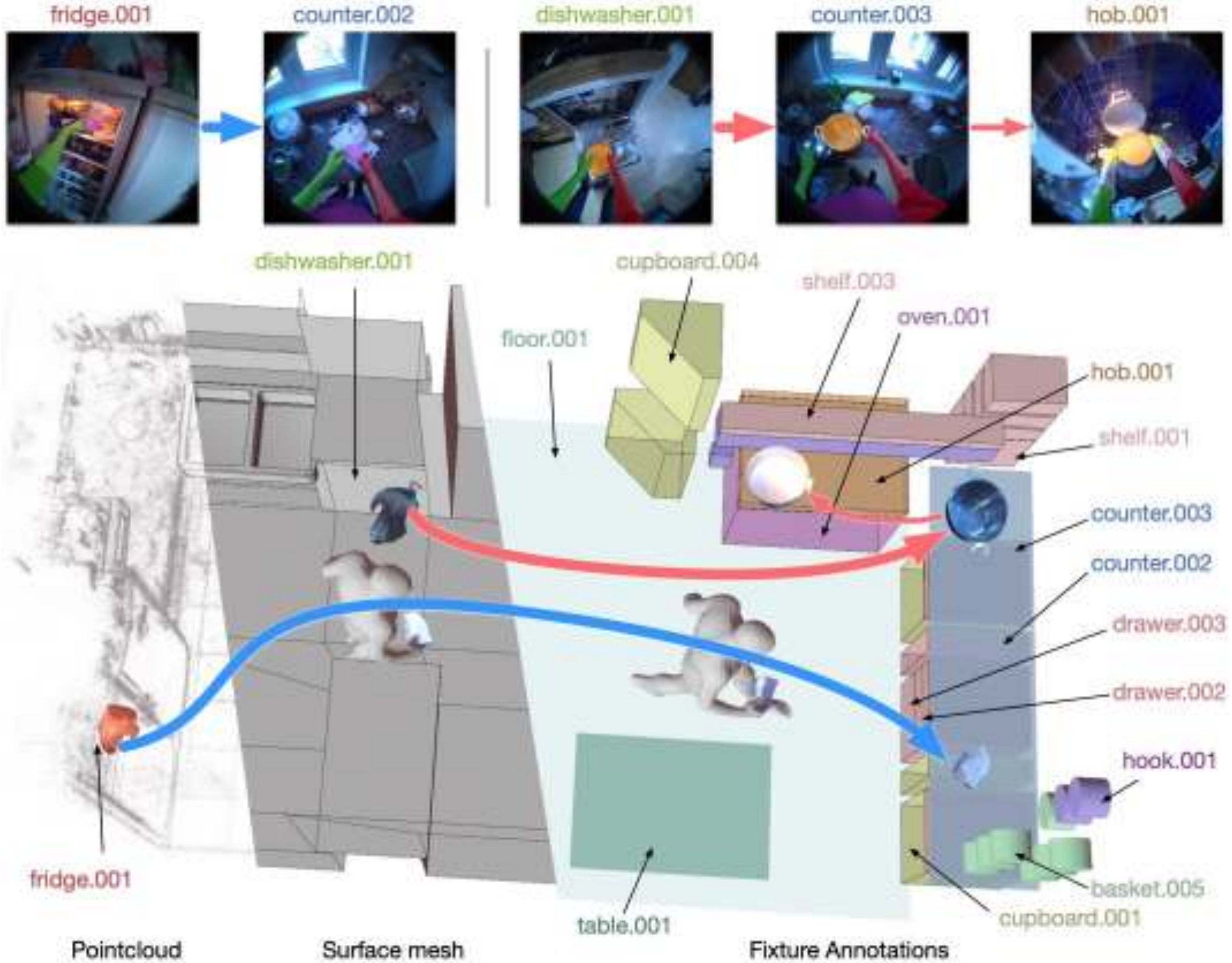
Dima Damen
ICVSS2025

HD-EPIC

Digital Twin

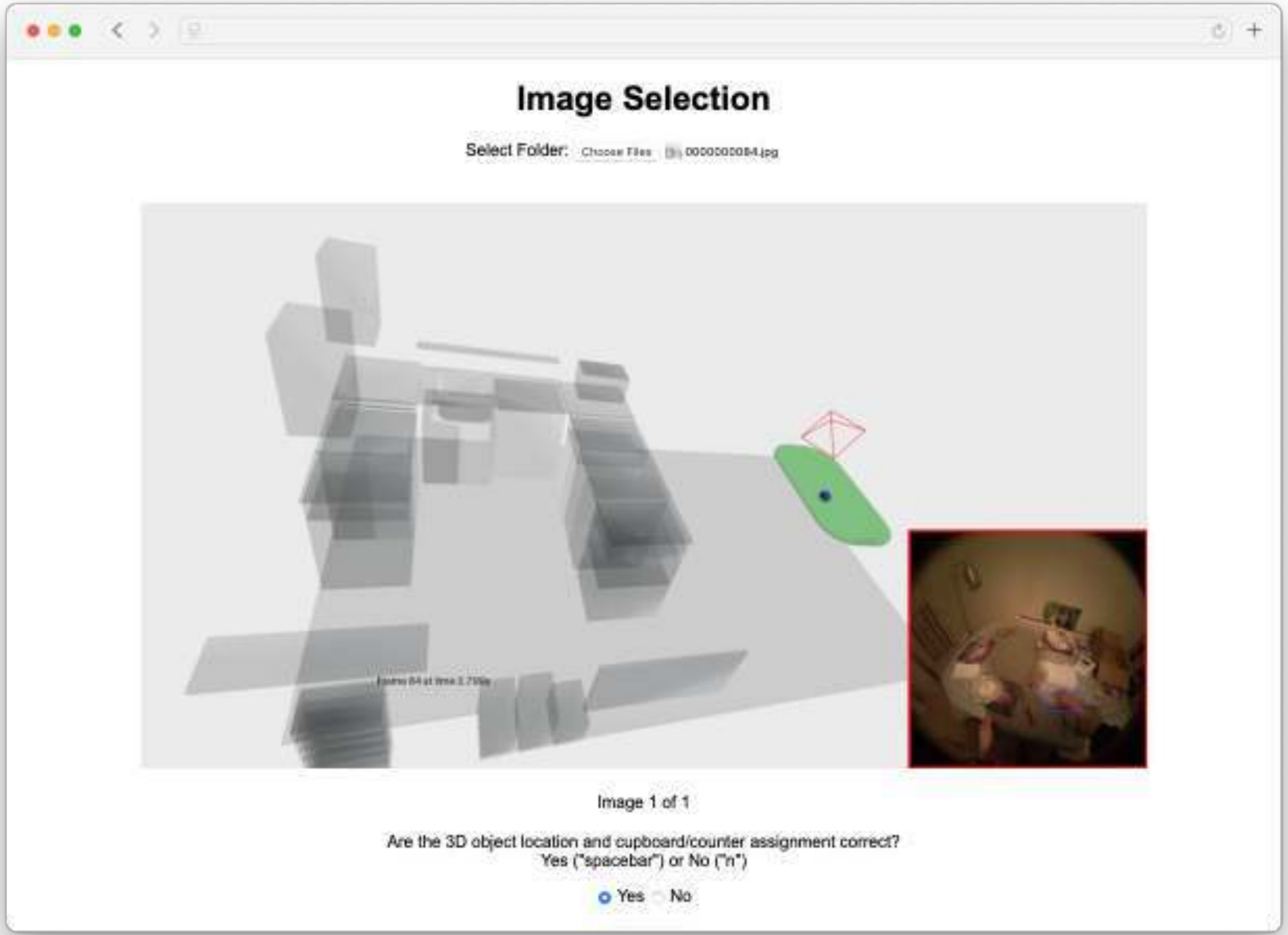
Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

HD-EPIC



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

HD-EPIC



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

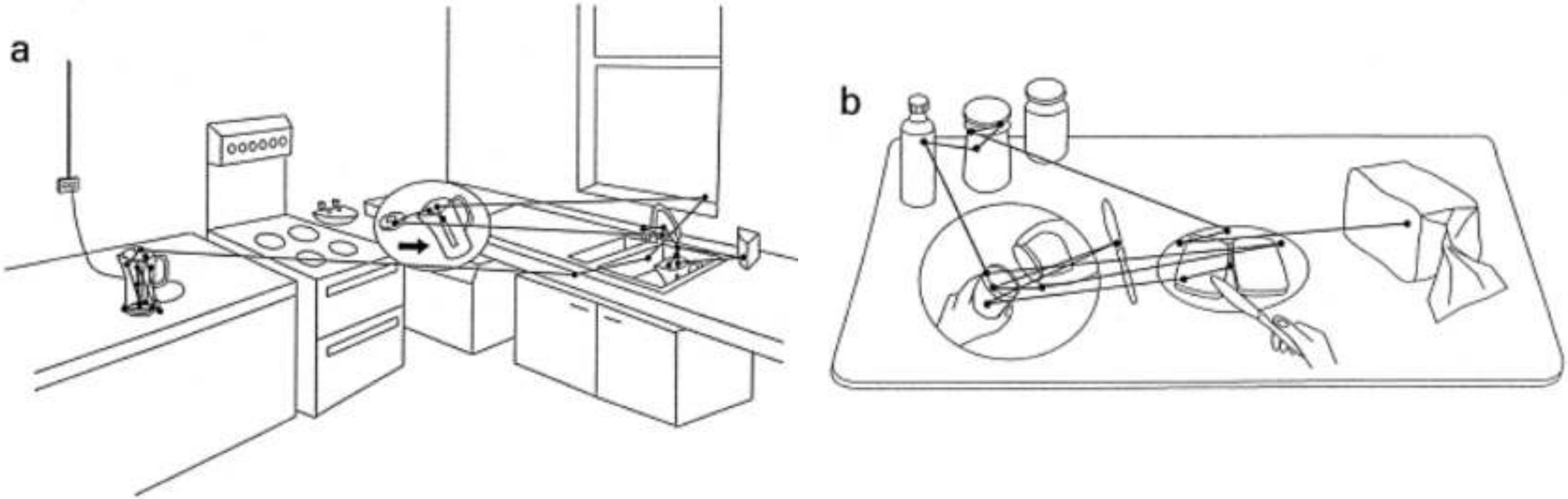
[redacted]

Video Understanding Out of the Frame

What can we now do with these reconstructions:

- Point Tracking
- Object Tracking
- Gaze Estimation

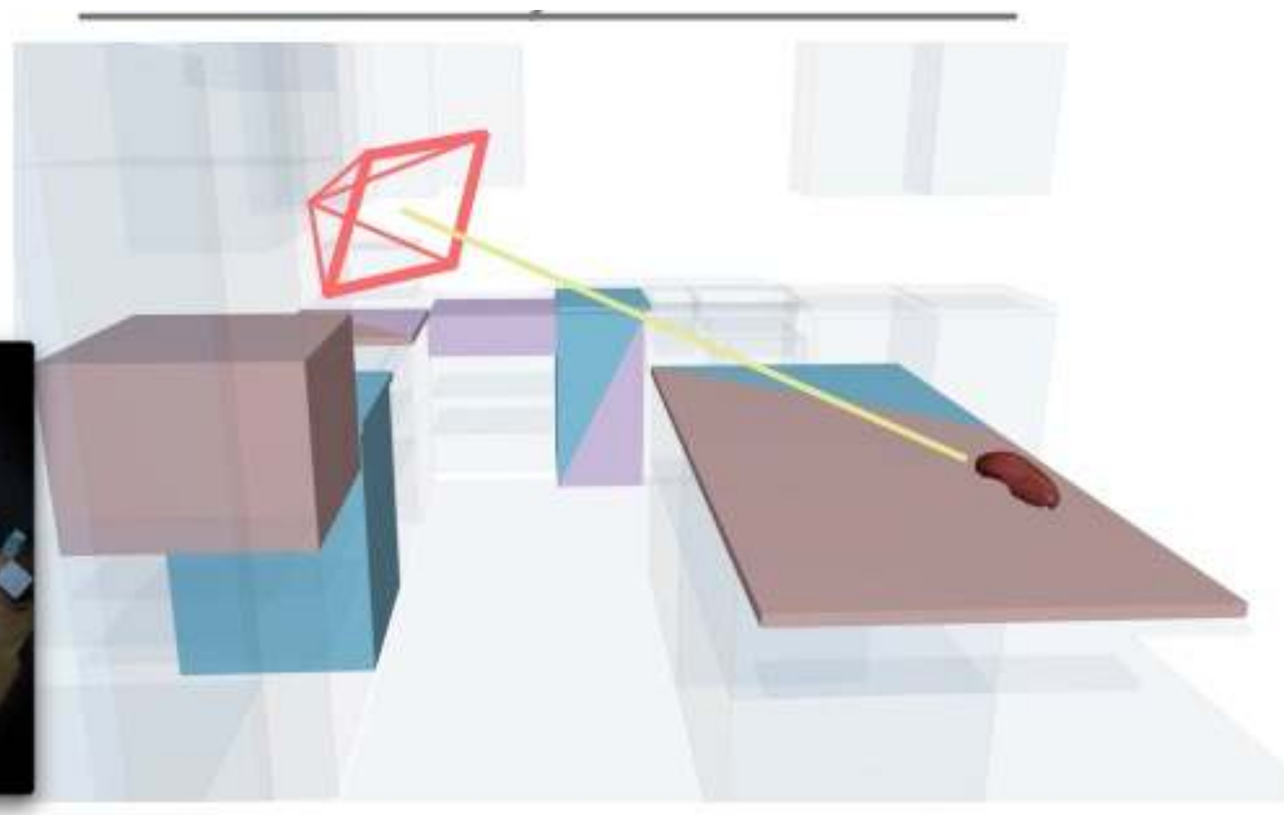
Gaze and Fixations



Land and Hayhoe (2001) In what ways do eye movements contribute to everyday activities? Vision Research

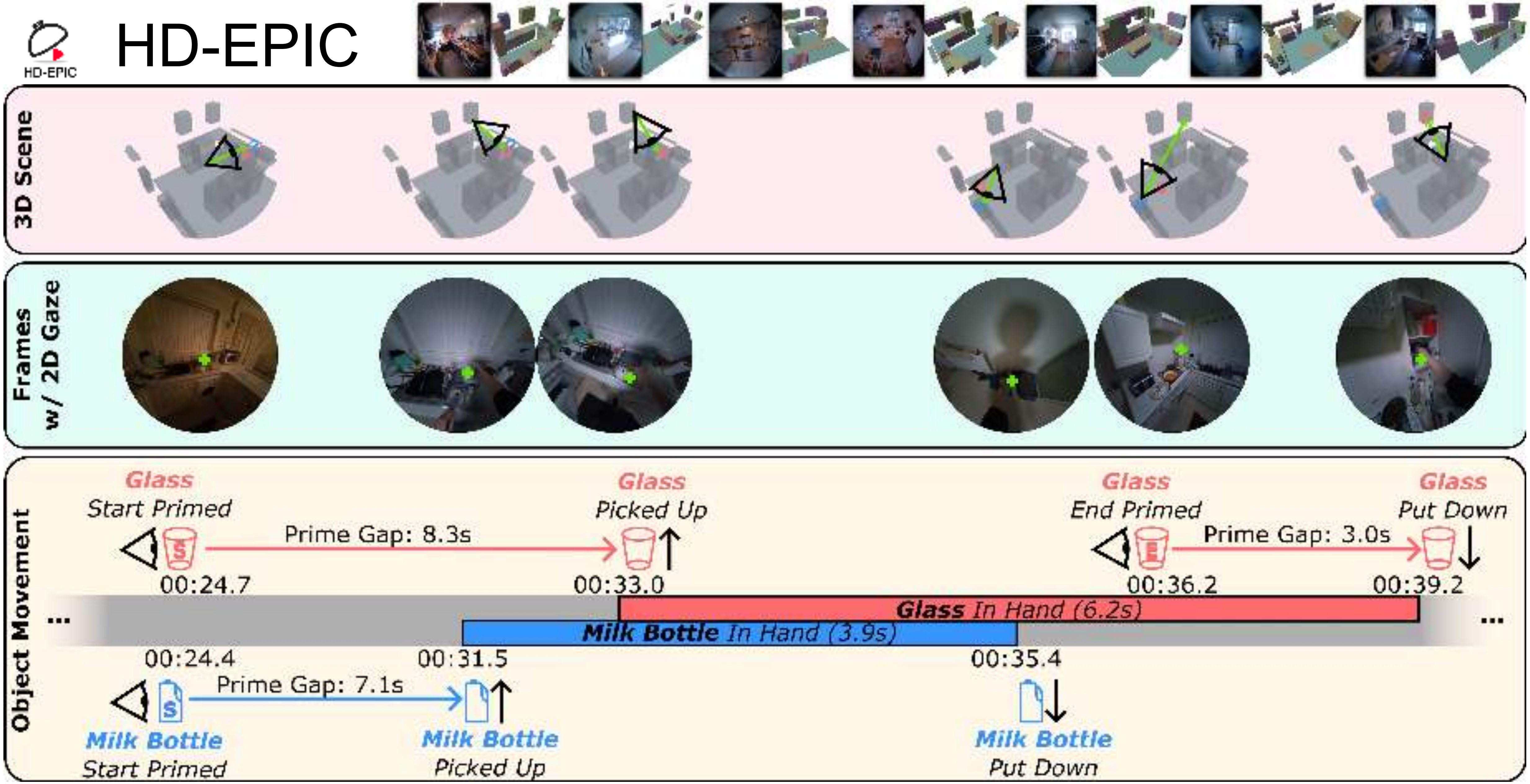
HD-EPIC

Gaze priming



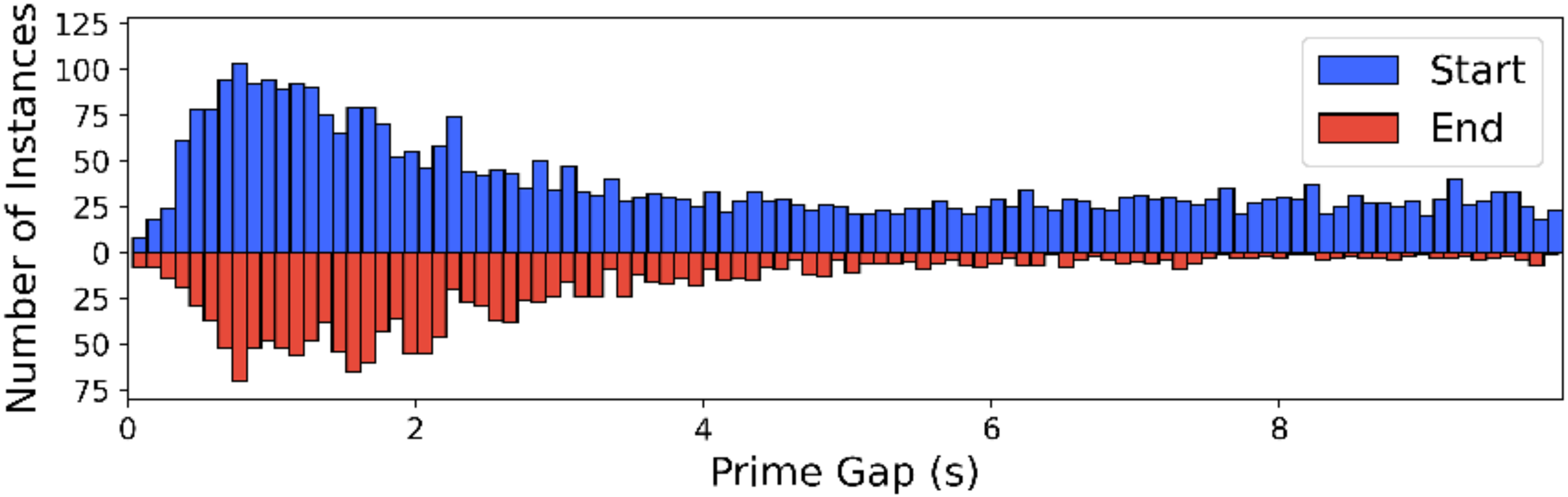
Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

HD-EPIC



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

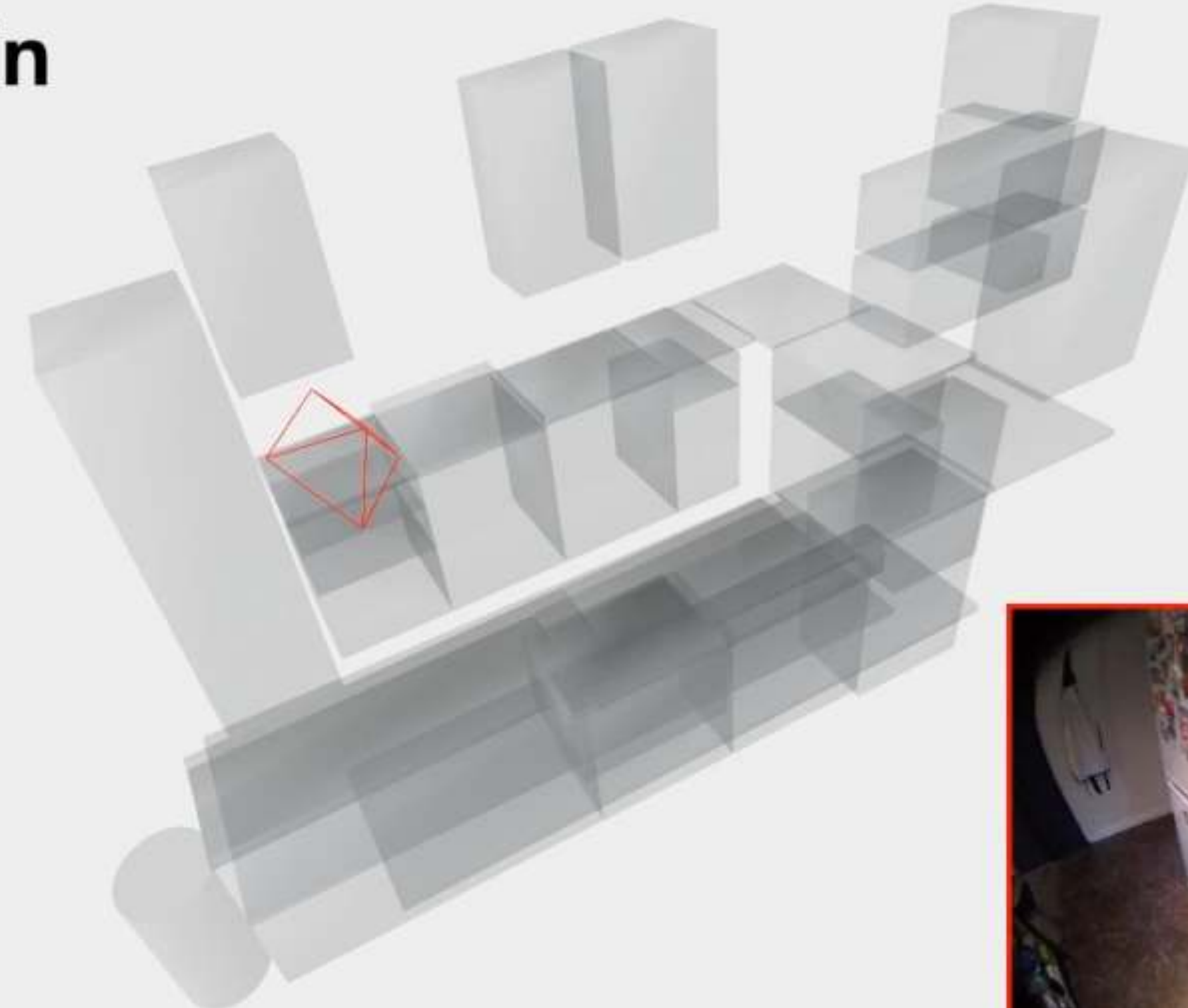
HD-EPIC


Number of Instances
125
100
75
50
25
0
25
50
75
0
2
4
6
8
Prime Gap (s)
Start
End

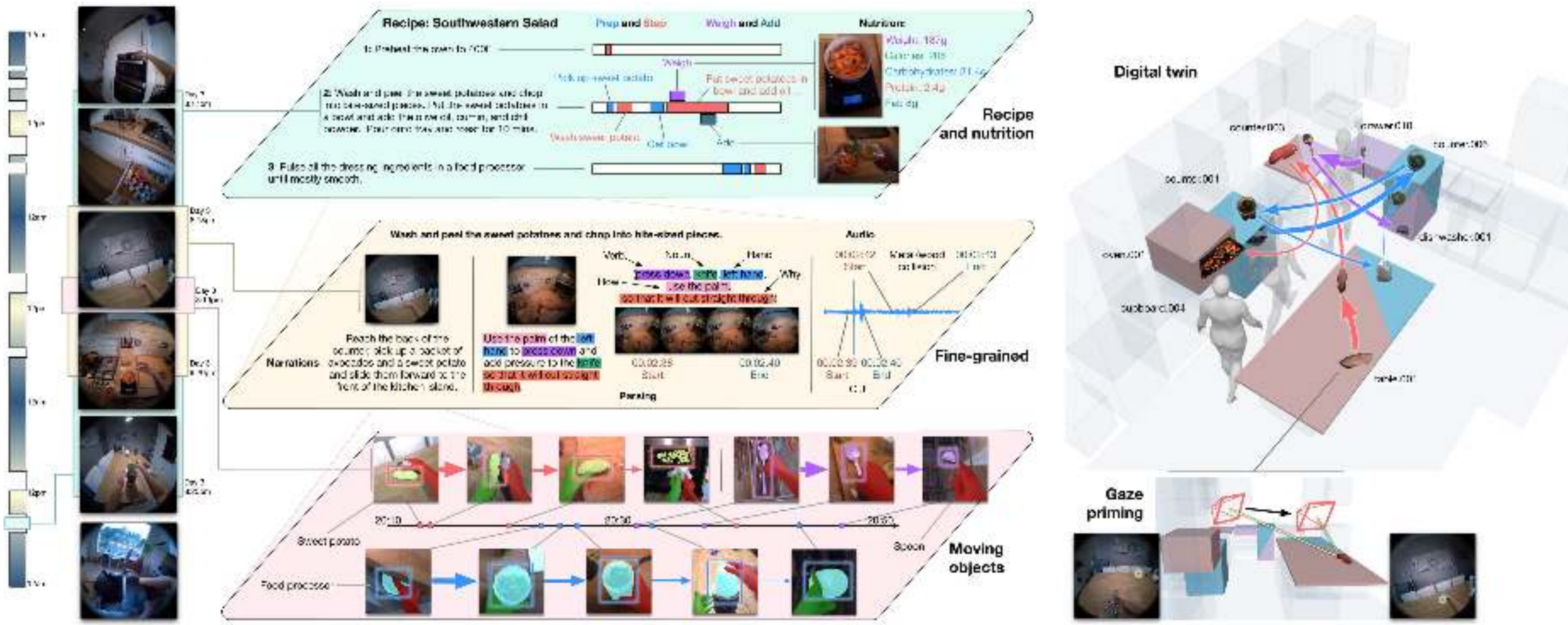
Digital Twin

Fixtures

Open drawer



HD-EPIC



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

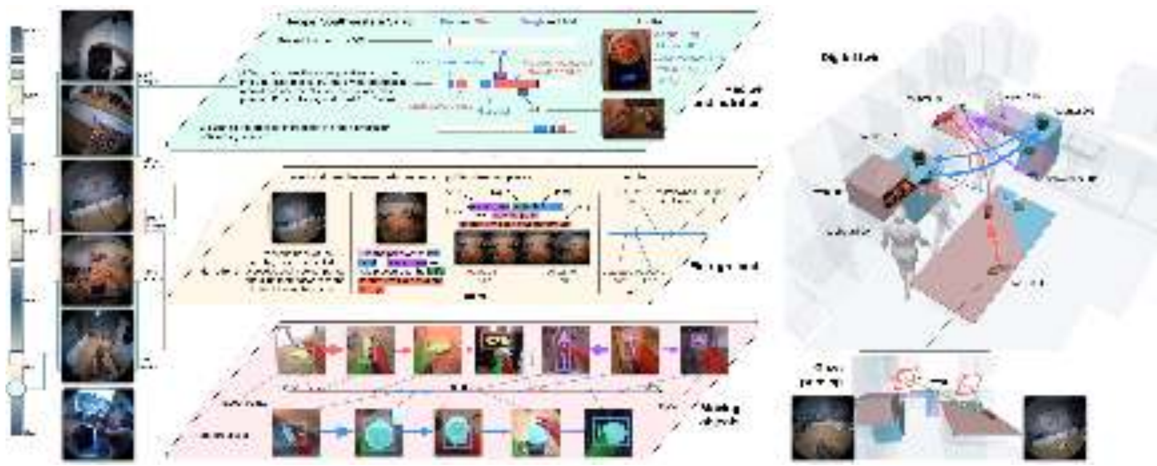
Dima Damen
ICVSS2025

HD-EPIC

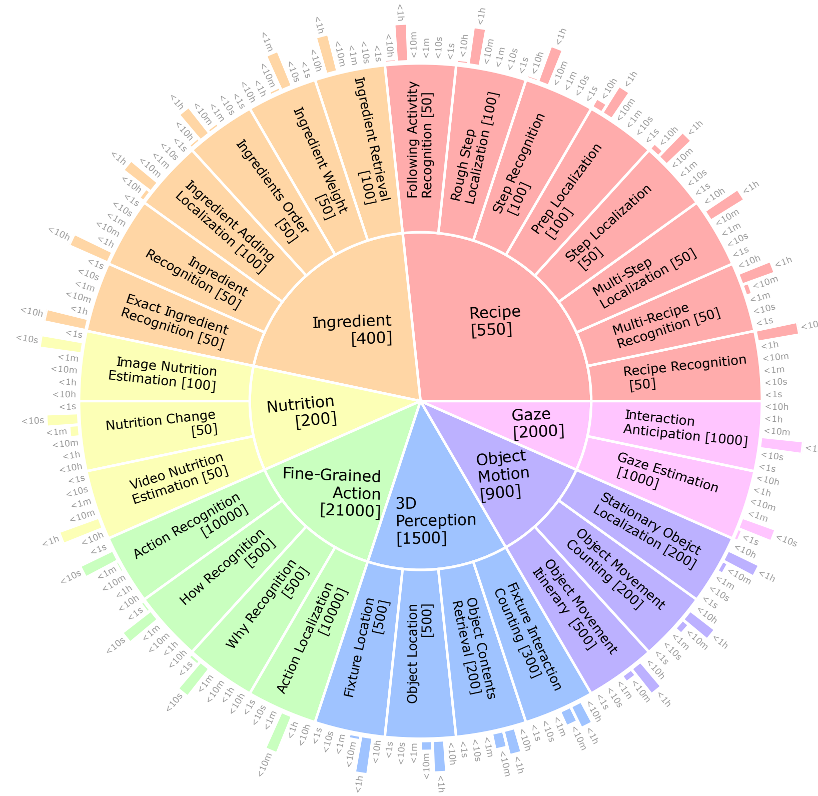
Annotation Type	Total annotations	Annotations/min
Narrations	59,454	24.0
Parsing (Verbs + Nouns + Hands + How + Why)	303,968	122.7
Recipes (Preps + Steps)	4,052	1.6
Sound	50,968	20.6
Action boundaries	59,454	24.0
Object Motion (Pick up + Put down + Fixtures + Bboxes + Masks)	153,480	62.0
Object Itinerary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

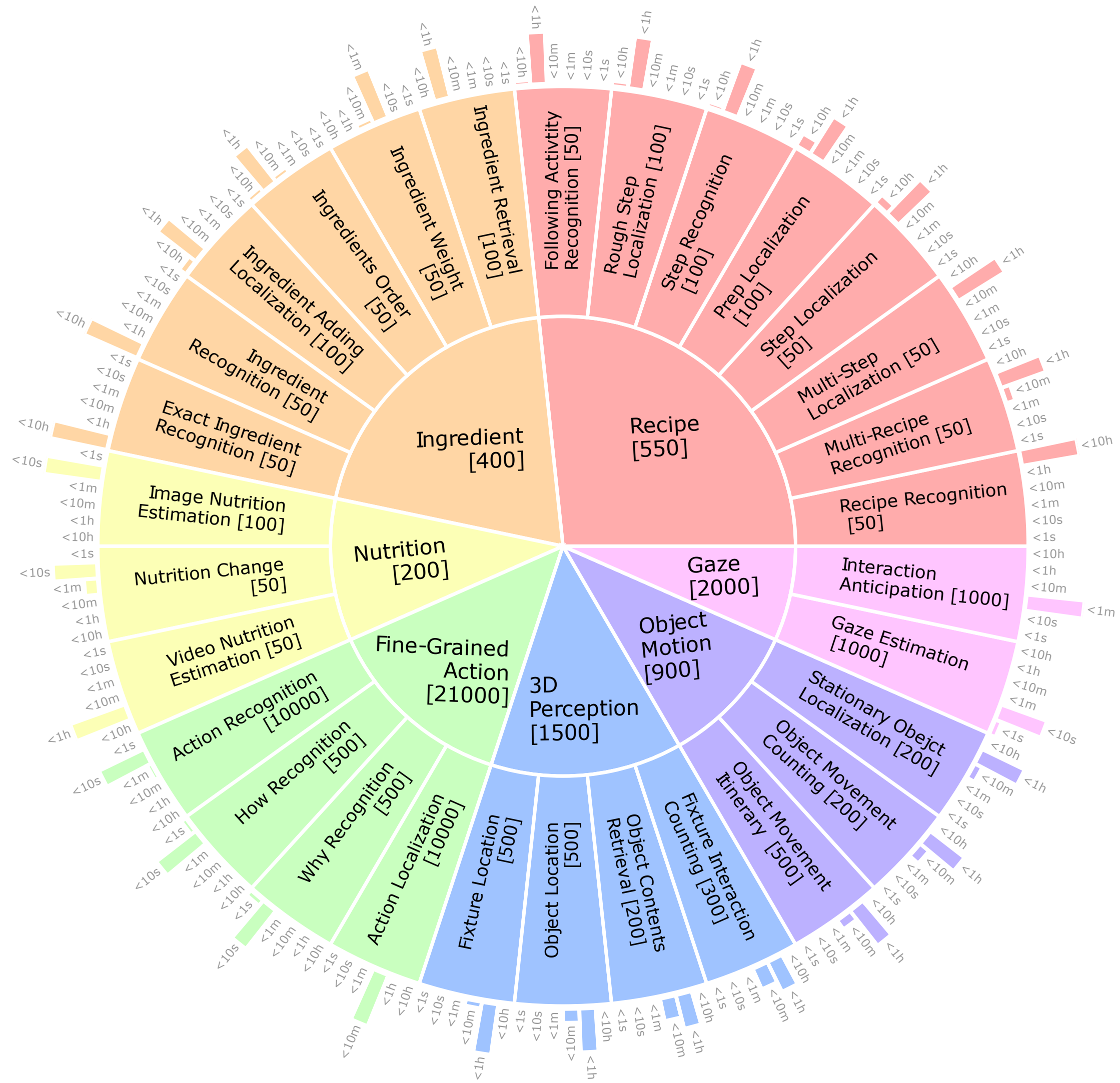
Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

University of Bristol

Dima Damen
ICVSS2025

HD-EPIC

1. Recipe. Questions on temporally localising, retrieving, or recognising recipes and their steps.
2. Ingredient. Questions on the ingredients used, their weight, their adding time and order.
3. Nutrition. Questions on nutrition of ingredients and nutritional changes as ingredients are added to recipes.
4. Fine-grained action. What, how, and why of actions and their temporal localisation.
5. 3D perception. Questions that require the understanding of relative positions of objects in the 3D scene.
6. Object motion. Questions on where, when and how many times objects are moved across long videos.
7. Gaze. Questions on estimating the fixation on large landmarks and anticipating future object interactions.



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

HD-EPIC

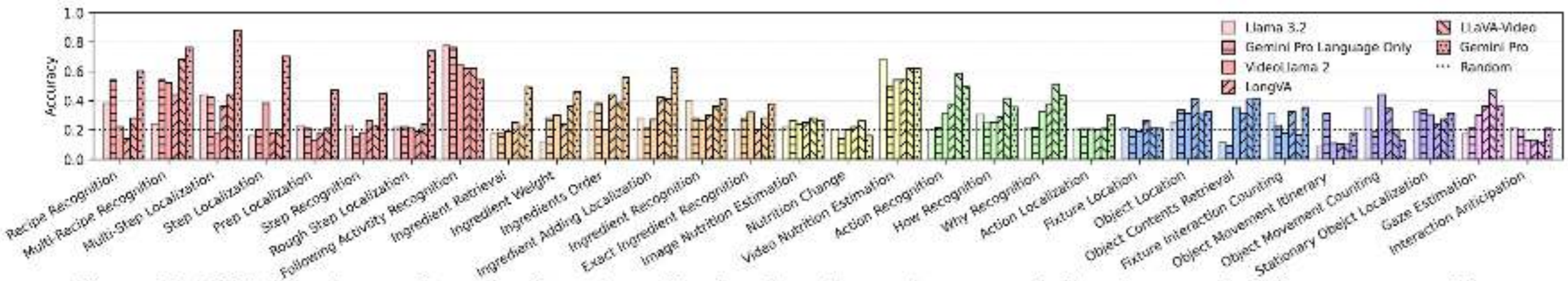


Figure 11. **VQA Results per Question Prototype**. Our benchmark contains many challenging questions for current models.

Model	Recipe	Ingredient	Nutrition	Action	3D	Motion	Gaze	Avg.
Blind - Language Only								
Llama 3.2	33.5	25.0	36.7	23.3	22.3	25.5	19.5	26.5
Gemini Pro	38.0	26.8	30.0	22.1	21.5	27.7	20.5	26.7
Video-Language								
VideoLlama 2	30.8	25.7	32.7	27.2	25.7	28.5	21.2	27.4
LongVA	29.6	30.8	33.7	30.7	32.9	22.7	24.5	29.3
LLaVA-Video	36.3	33.5	38.7	43.0	27.3	18.9	29.3	32.4
Gemini Pro	64.3	48.6	34.7	39.6	32.5	20.8	28.7	38.5
Sample Human Baseline	*96.7*	*96.7*	*85.0*	*92.5*	*93.8*	*92.7*	*75.0*	*90.3*

University of BRISTOL

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

HD-EPIC

How many times did I **open** the item at bounding box **(165, 452, 1408, 1408)** in **00:00:57**?

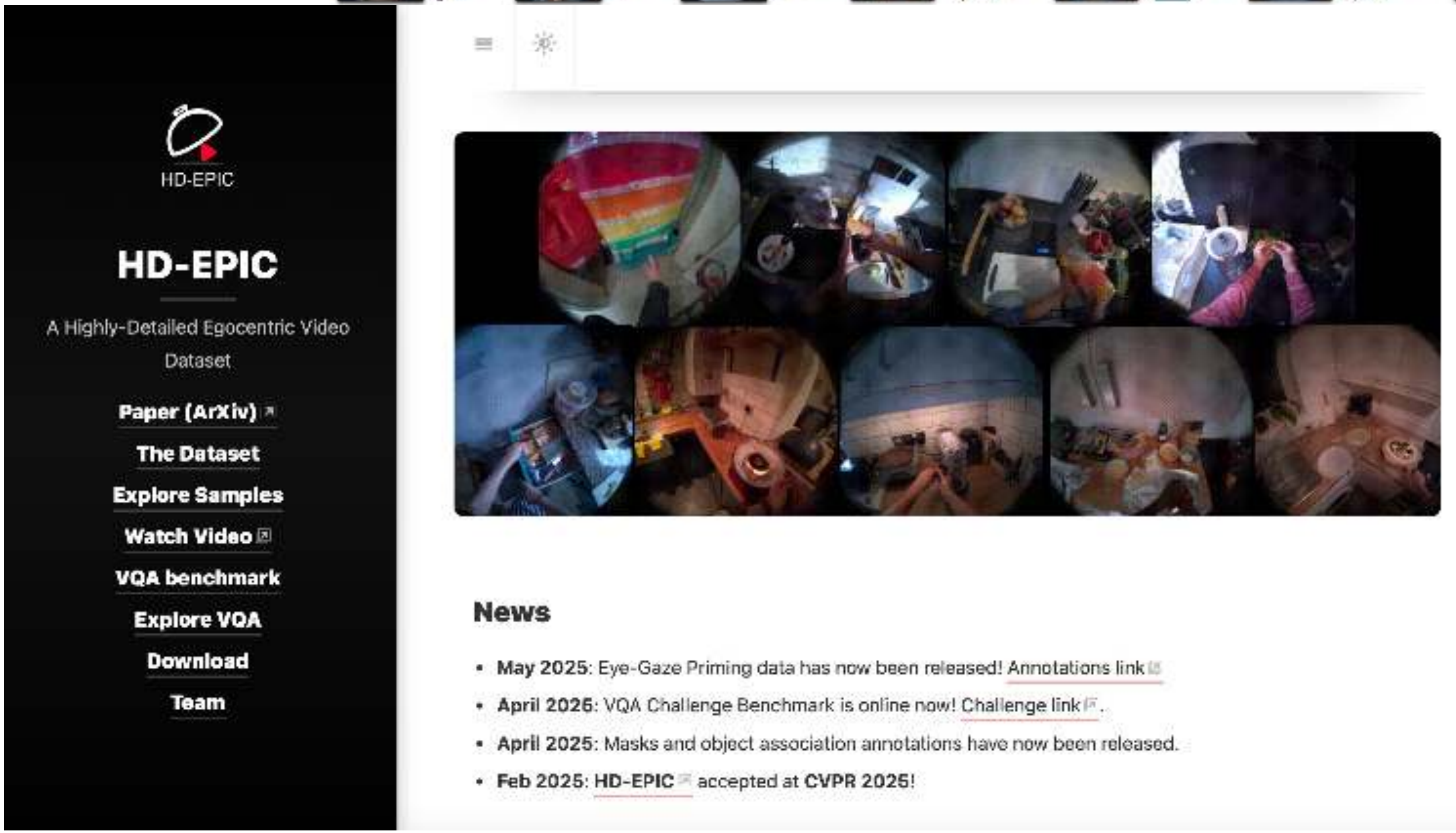
A. 3 **B.** 1 **C.** 4 **D.** 5 **E.** 2

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

HD-EPIC

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

HD-EPIC


HD-EPIC
HD-EPIC
A Highly-Detailed Egocentric Video Dataset
Paper (ArXiv)
The Dataset
Explore Samples
Watch Video
VQA benchmark
Explore VQA
Download
Team
News
May 2025: Eye-Gaze Priming data has now been released! Annotations link
April 2025: VQA Challenge Benchmark is online now! Challenge link.
April 2025: Masks and object association annotations have now been released.
Feb 2025: HD-EPIC accepted at CVPR 2025!

Try it Yourself

12 VQAs
Focusing on 3D

SCAN ME

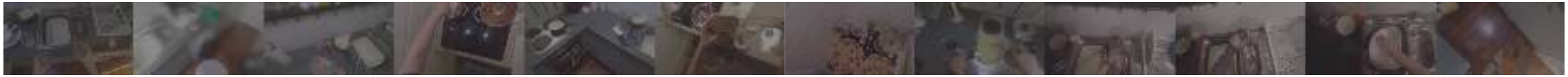
https://hd-epic.github.io/icvss_demo.html

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

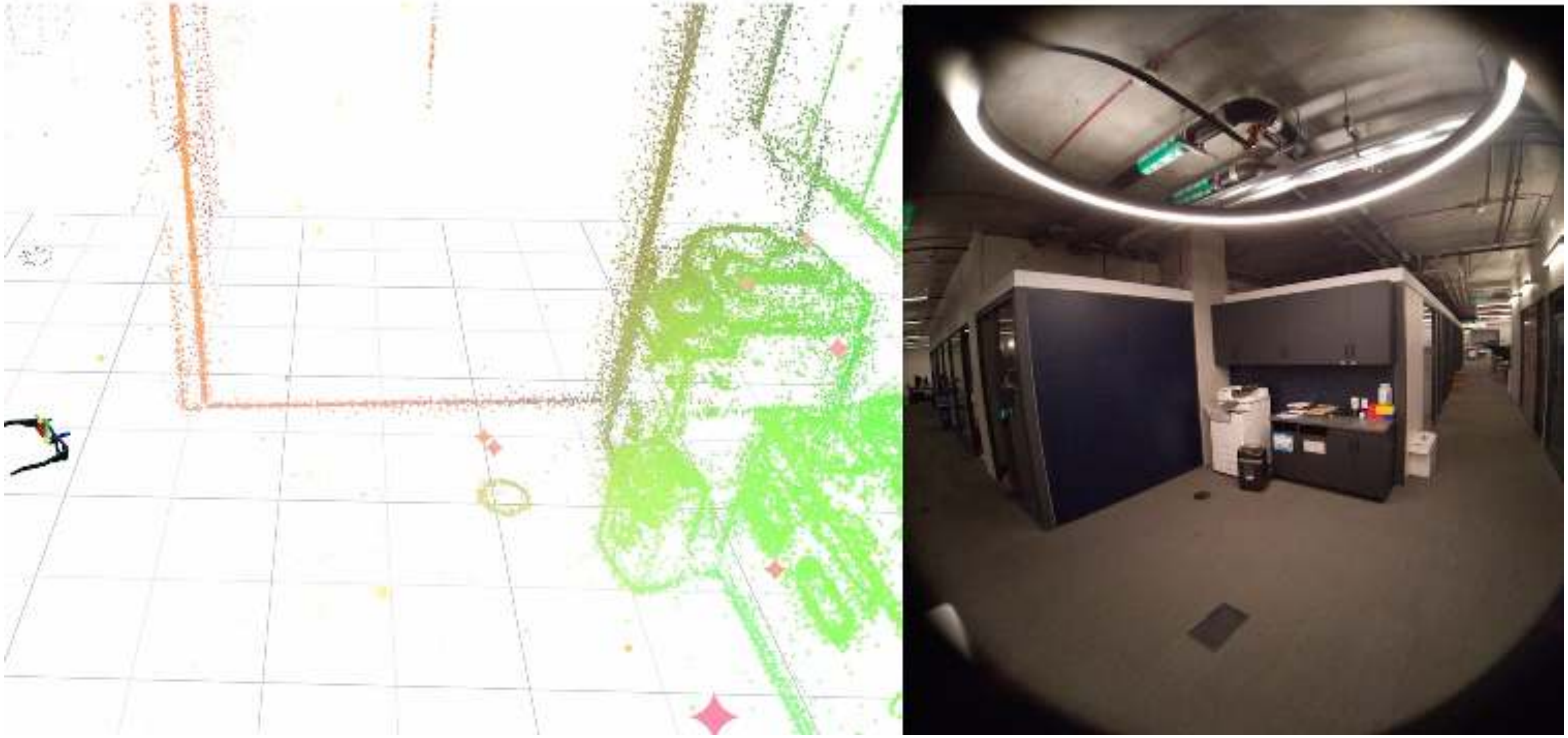
Body and Hand Motion Estimation “out of the frame”

University of BRISTOL

Dima Damen
ICVSS2025

EgoBody

EgoAllo uses egocentric (👓) SLAM poses and images

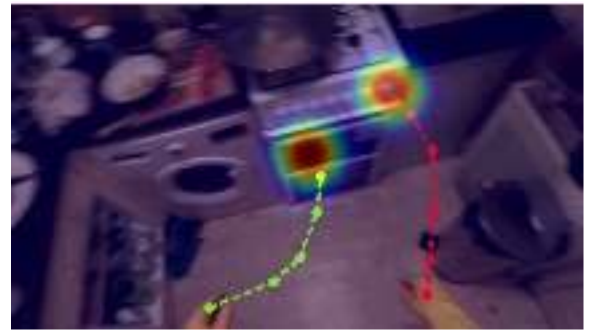


with: Masashi Hatano
Zhifan Zhu
Hideo Saito

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]

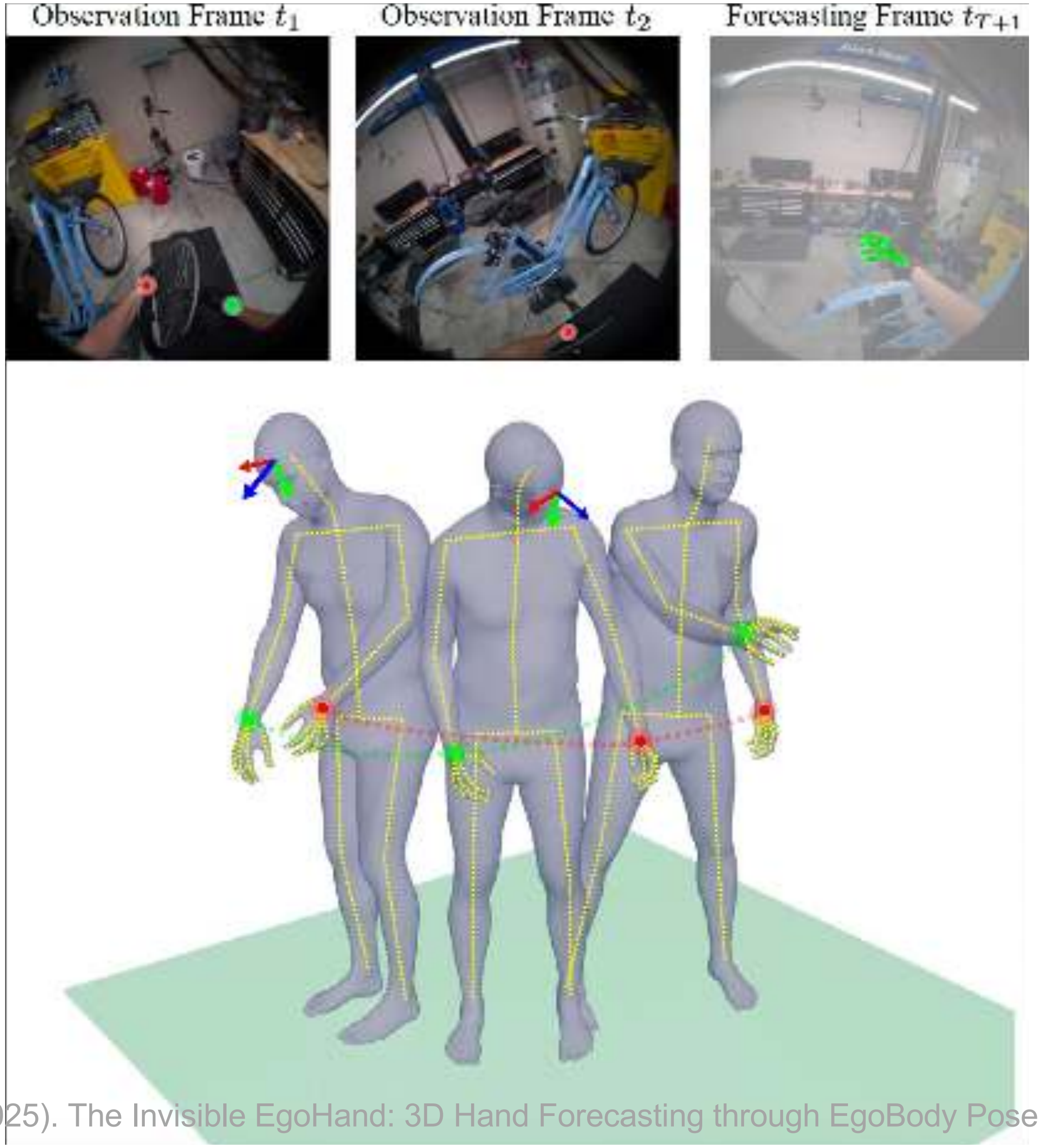
3D Hand Forecasting

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

USST [ICCV'23]

The Invisible EgoHand

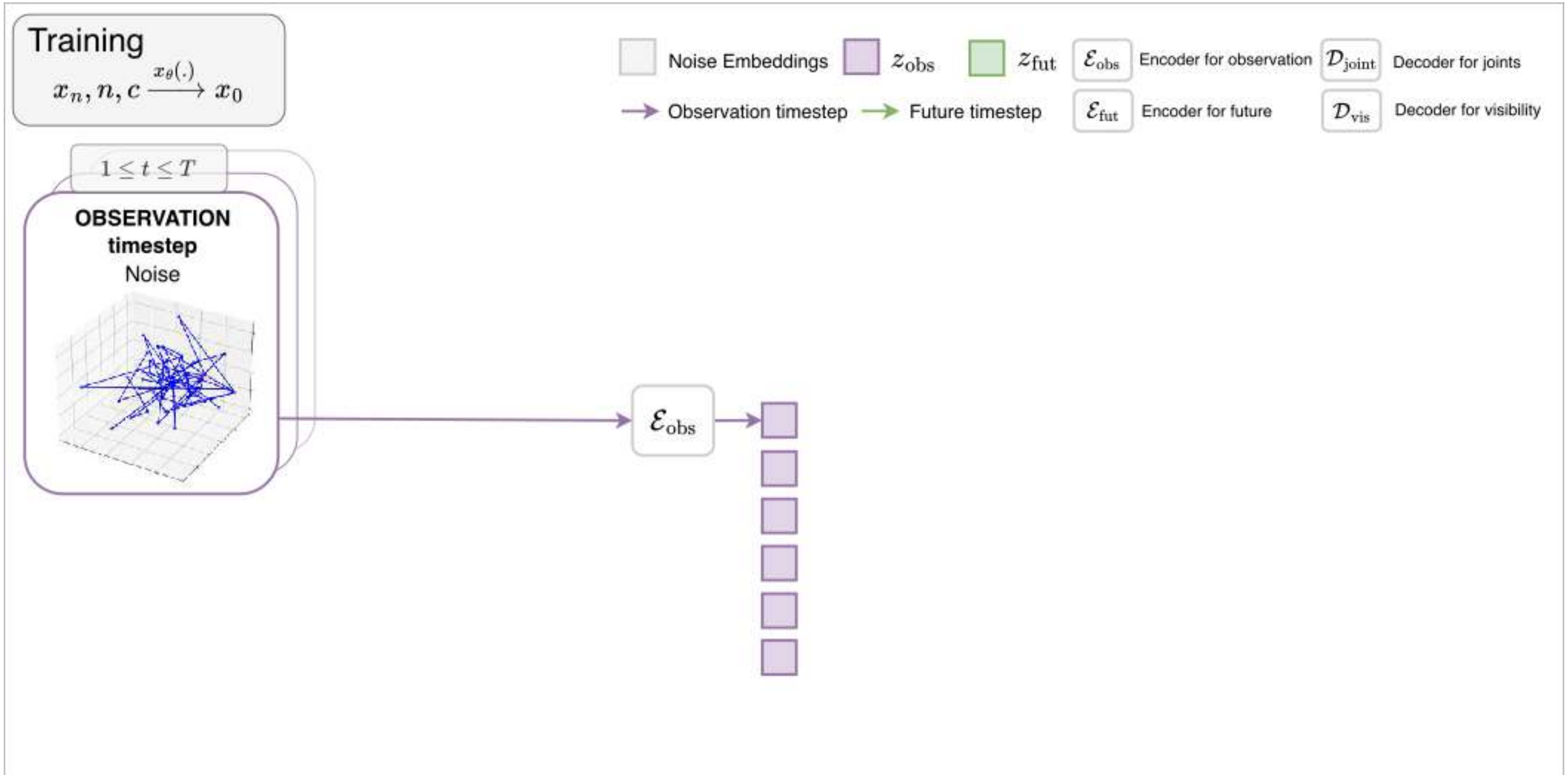
with: Masashi Hatano
Zhifan Zhu
Hideo Saito



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**

University of BRISTOL

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

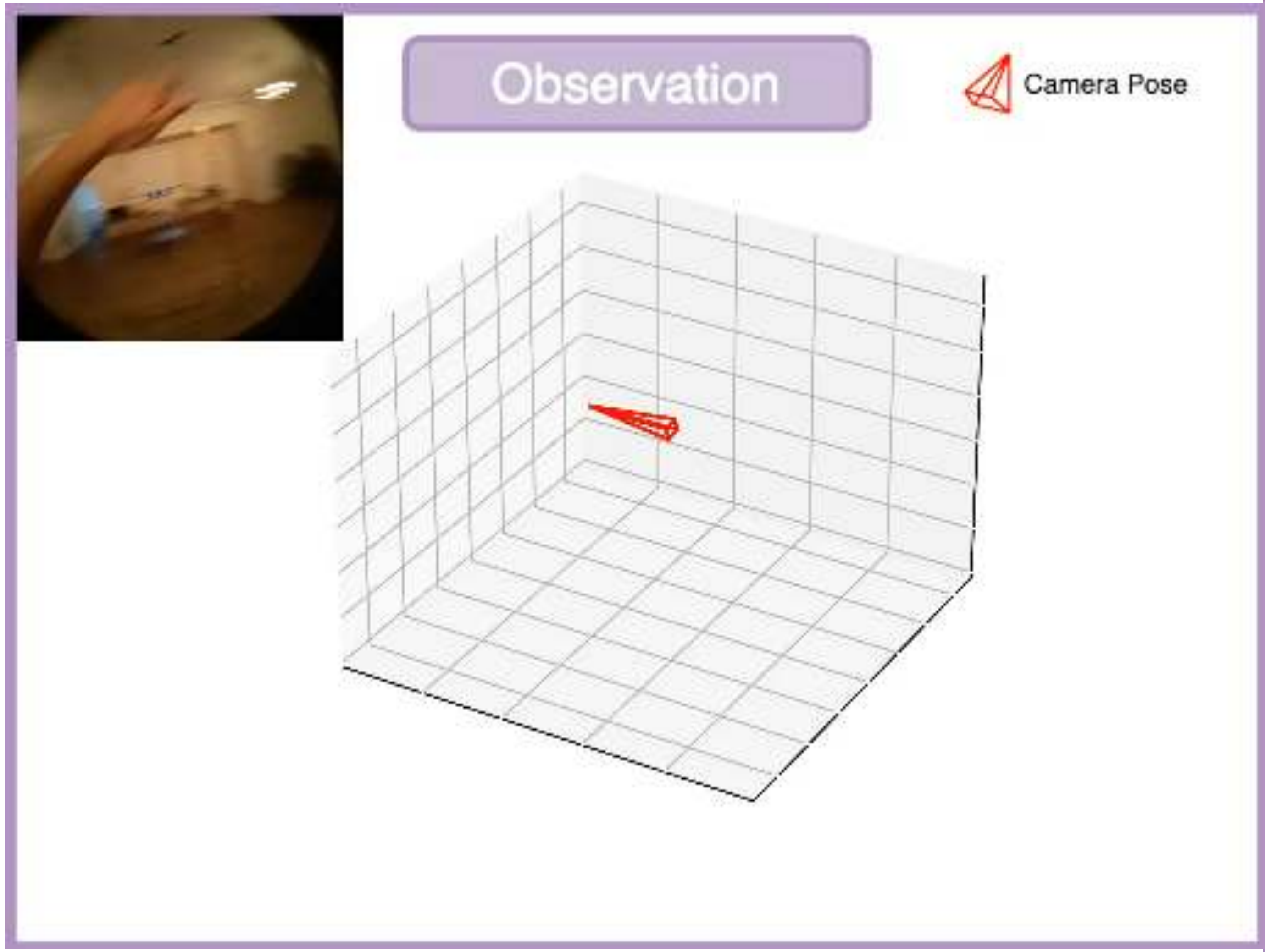
Method	Hand Trajectory Forecasting			Hand Pose Forecasting		
	In-view	Out-of-view	All	In-view	Out-of-view	All
EgoEgoForecast	0.171	0.385	0.295	0.162	0.299	0.166
Ours w/o. 2D joint	0.151	0.377	0.282	0.139	0.269	0.142
Ours w/o. image	**0.116**	0.367	**0.261**	0.117	**0.234**	0.120
Ours w/o. $\mathcal{L}_{reproj}$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

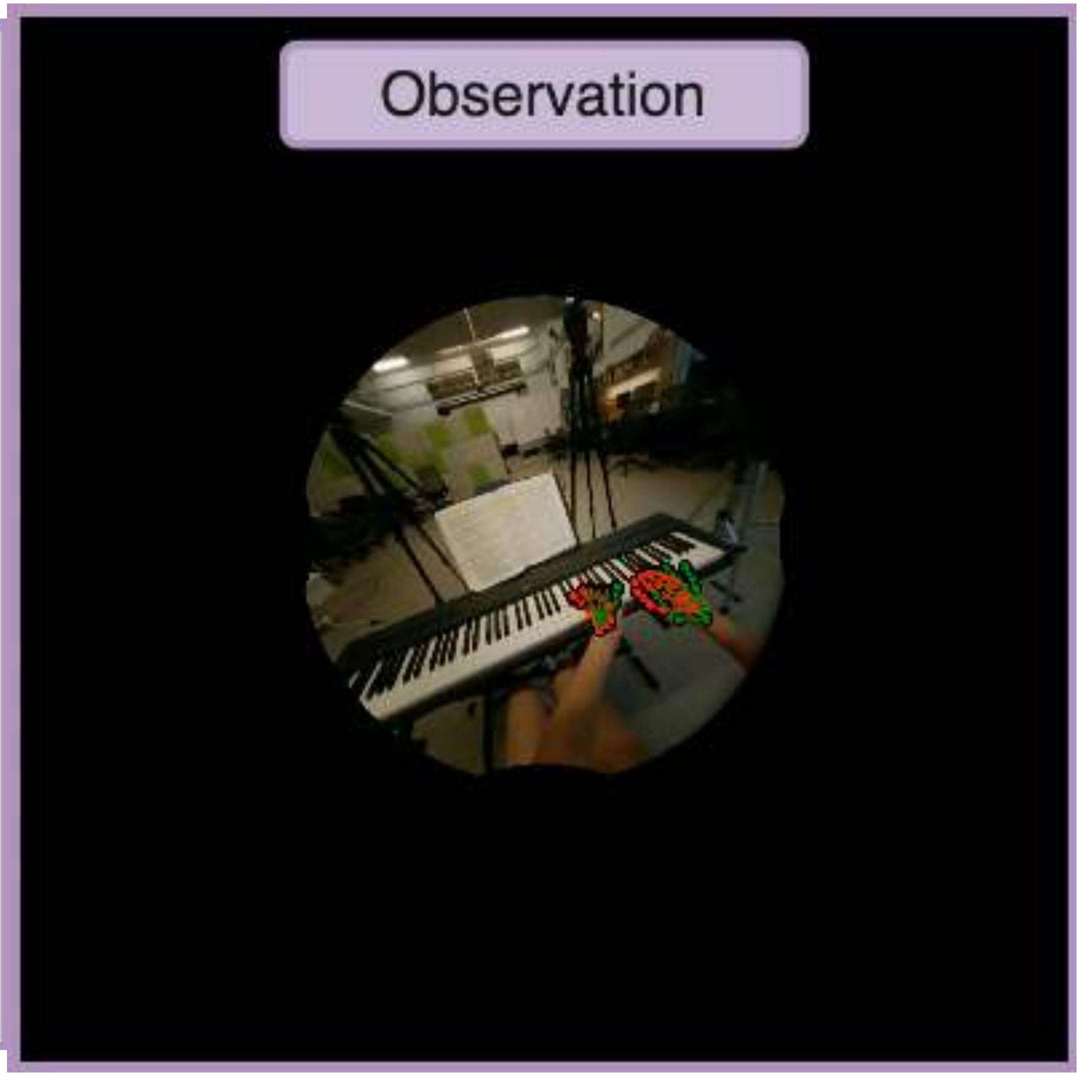
- Without visible 2D joints, significant performance drops can be seen
- 2D reprojection loss serves as effective regularization
- Visibility loss & Body joints loss contribute for out-of-view scenario

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

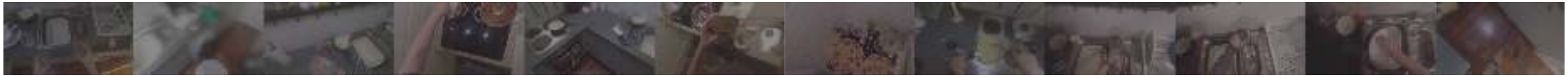
The Invisible EgoHand

with: Masashi Hatano
Zhifan Zhu
Hideo Saito





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



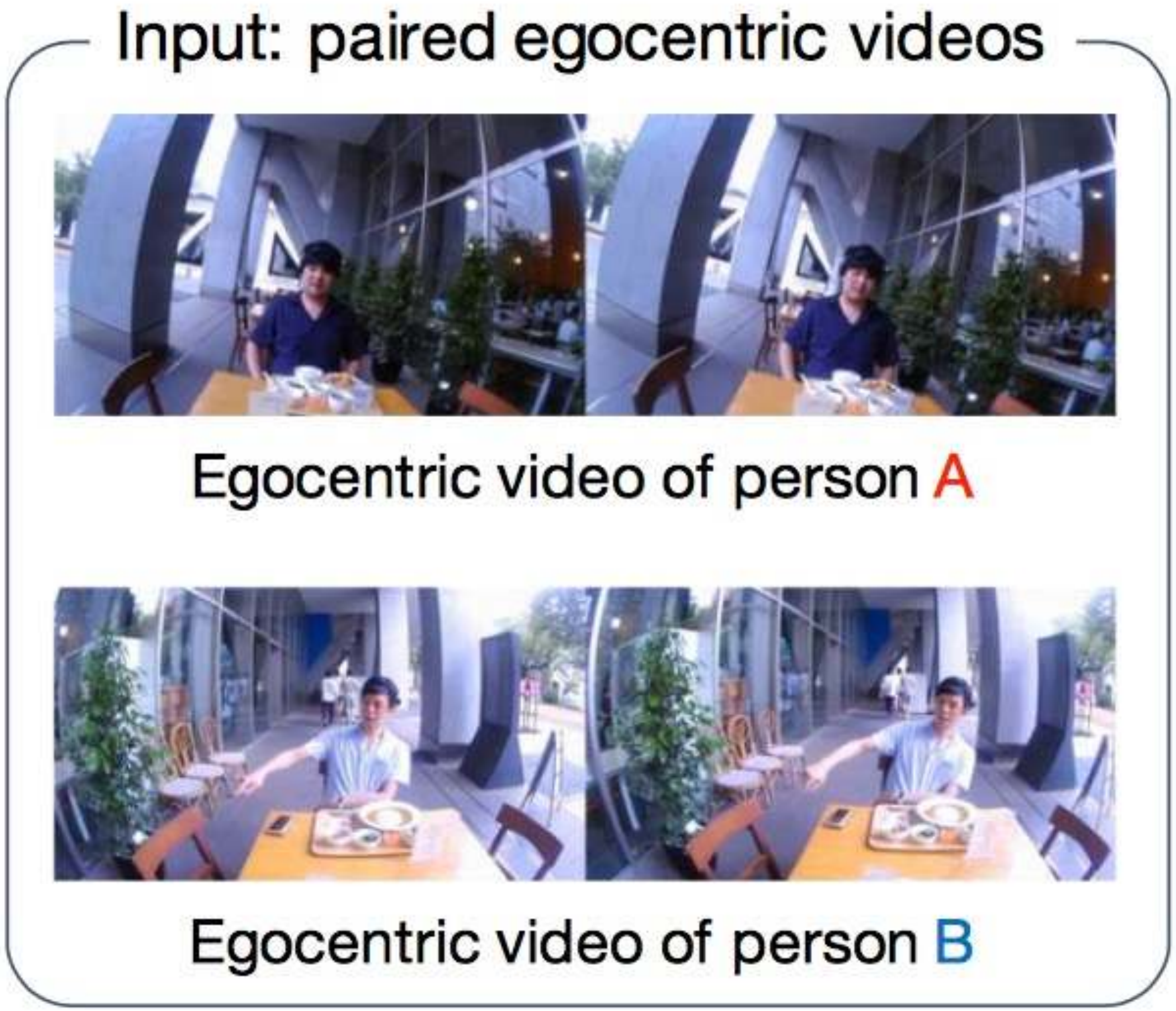
Video Understanding Out of the Frame

From First-Point View to Second- and Third-

University of BRISTOL

Dima Damen
ICVSS2025

FPV with SPV

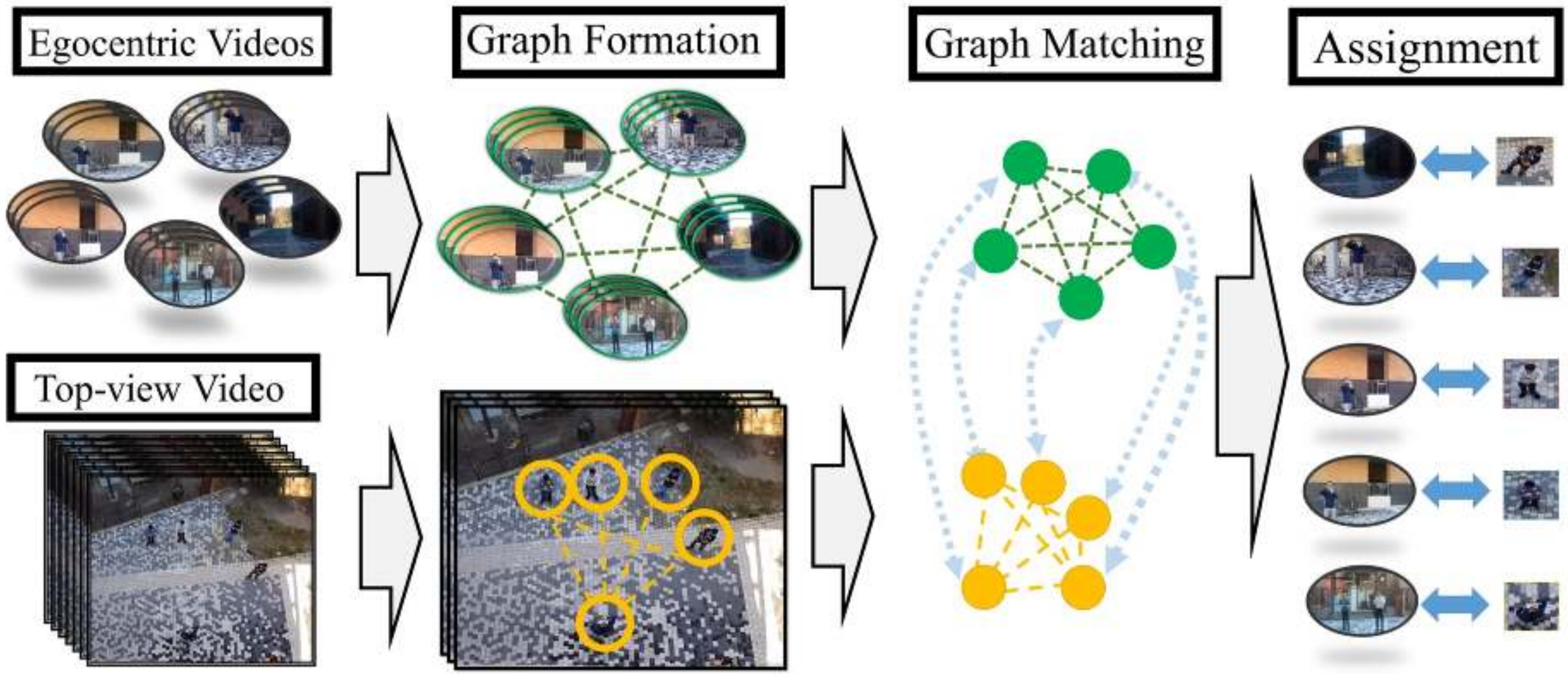


Yonetani et al (2016). Recognizing Micro-Actions and Reactions From Paired Egocentric Videos. CVPR

FPV with TPV (top-view)

Ardeshir and Borji (2016). Egocentric Meets Top-view CVPR

FPV with TPV (top-view)



Ego-Exo4D

with: Kristen Grauman
+102 authors

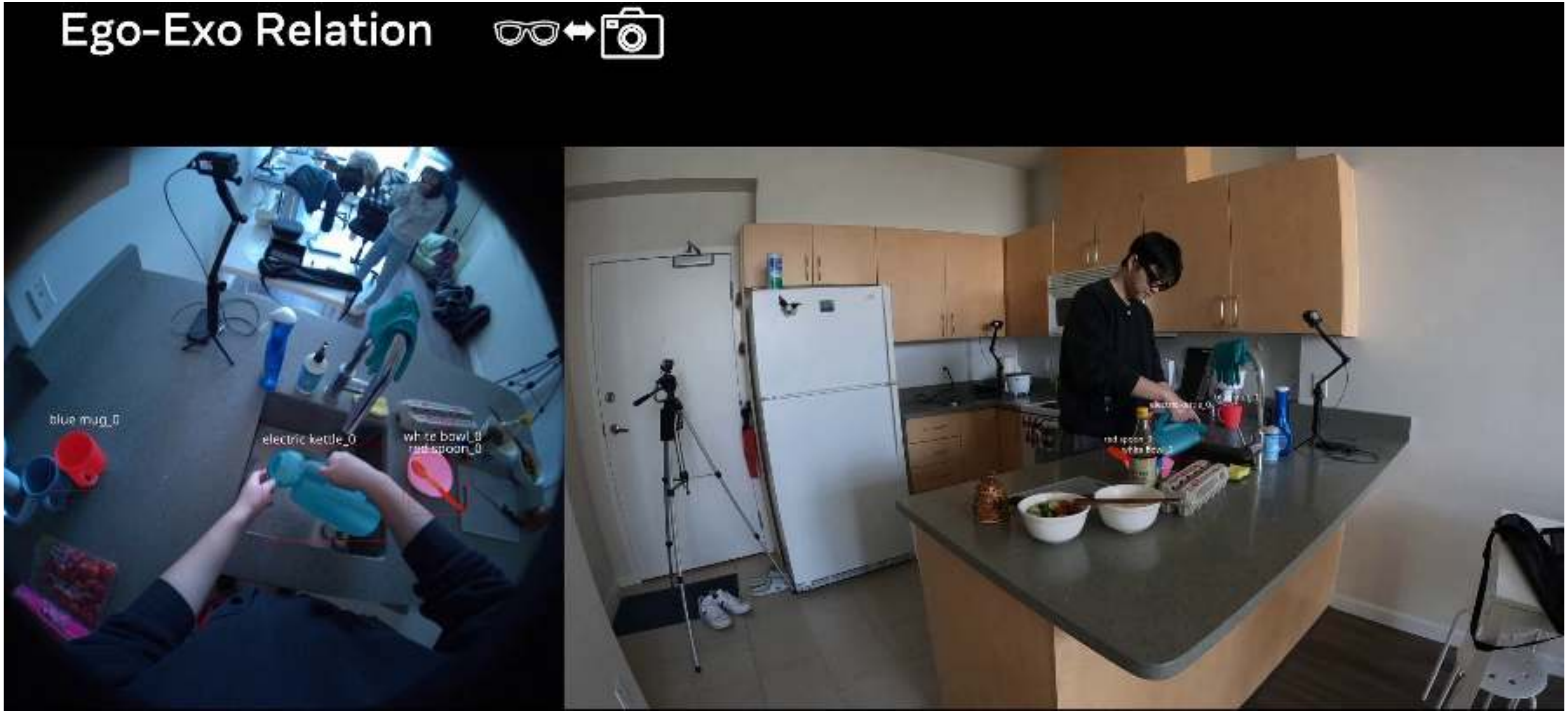
K Grauman et al (2024). Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives.. *CVPR*

University of BRISTOL

Dima Damen
ICVSS2025

Ego-Exo4D

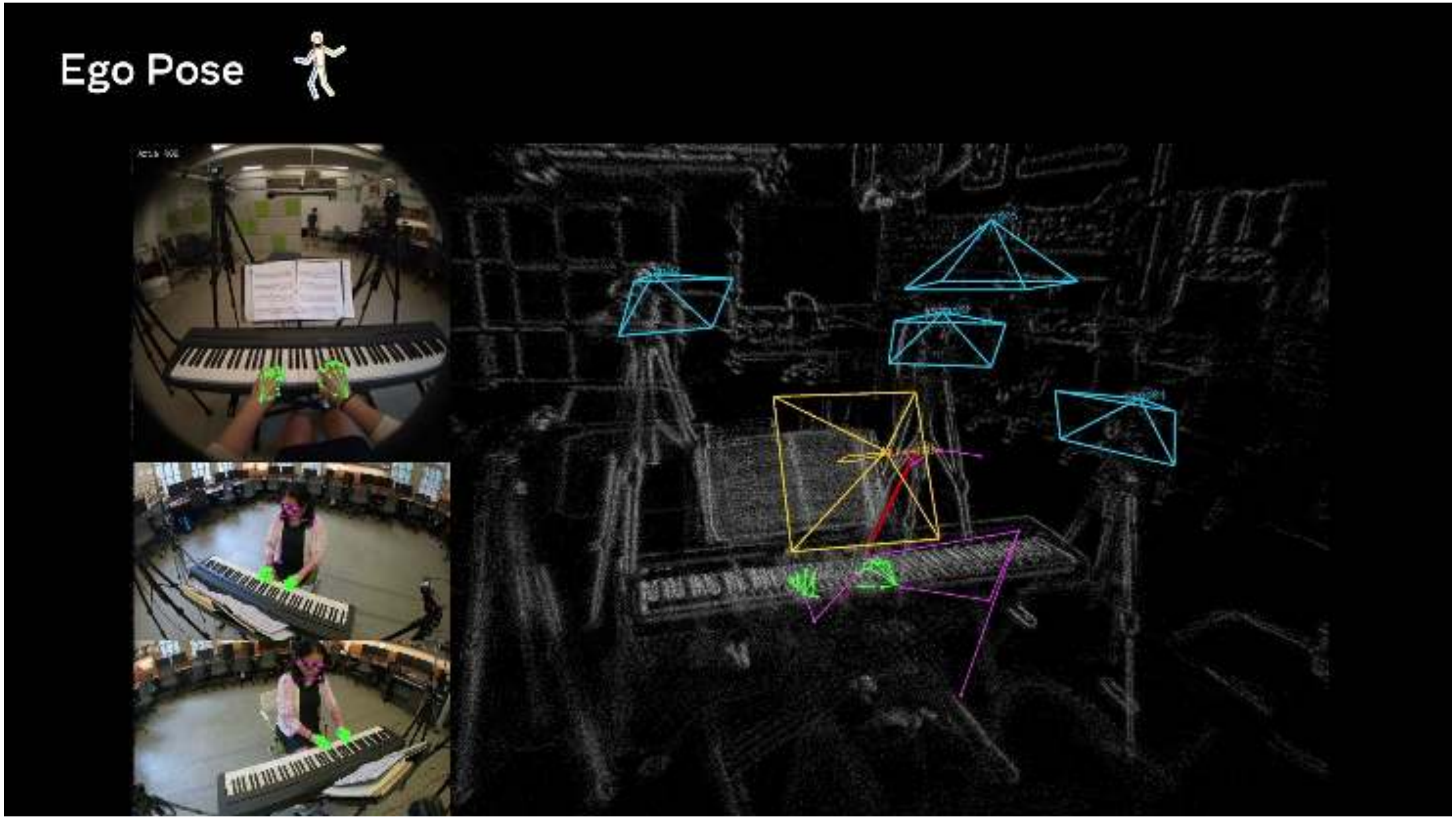
with: Kristen Grauman
+102 authors



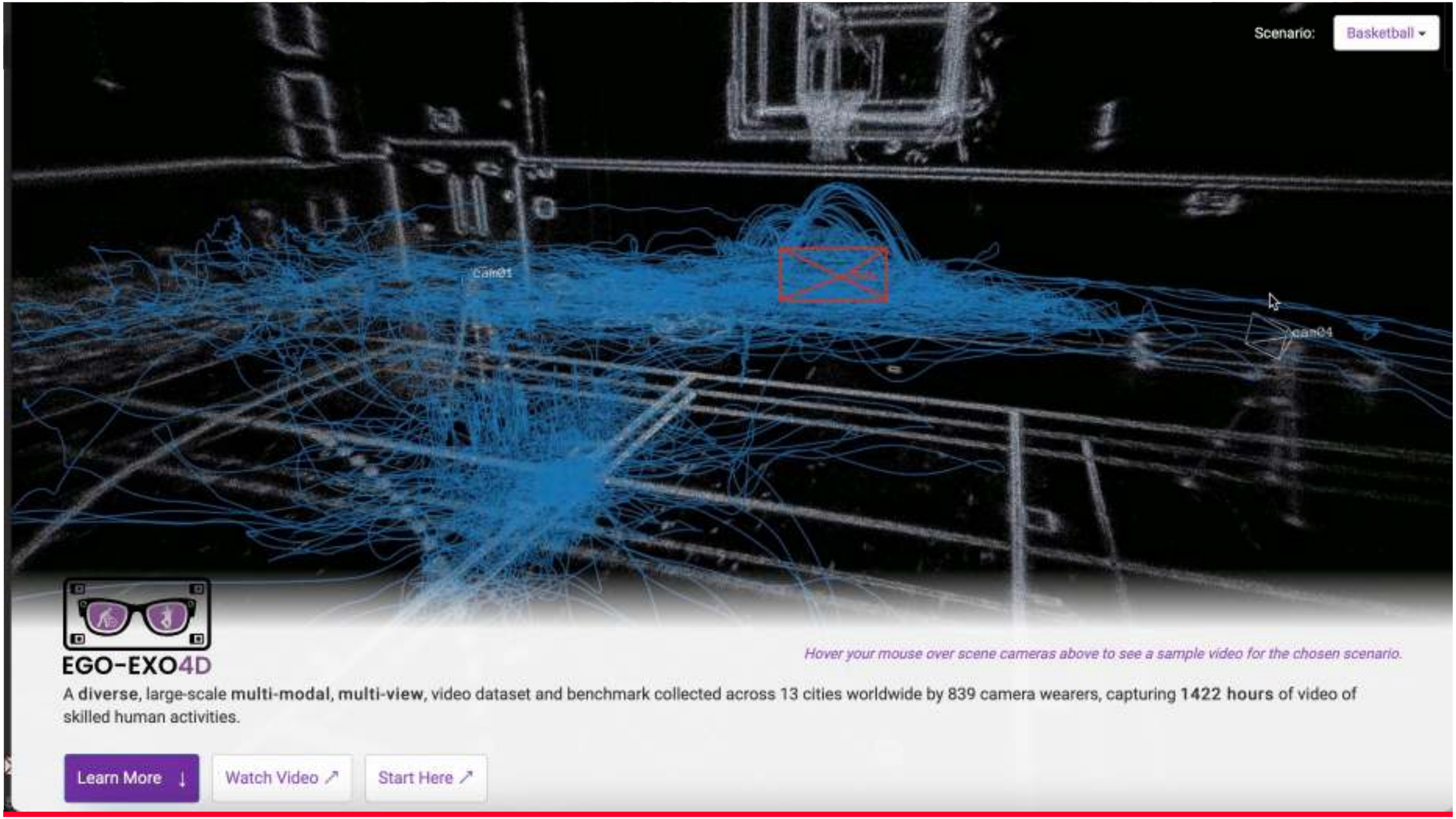
K Grauman et al (2024). Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives.. *CVPR*

Ego-Exo4D

with: Kristen Grauman
+102 authors



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


Scenario:
Basketball
cam01
cam04
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 at the Sphere
- Outlook into the Future of Egocentric Vision
- Conclusion

The Wizard of Oz at the Sphere

Coming in August 2025

https://behindthecurtain.withgoogle.com

Dima Damen
ICVSS2025

The Wizard of Oz @ The Sphere

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

[redacted]

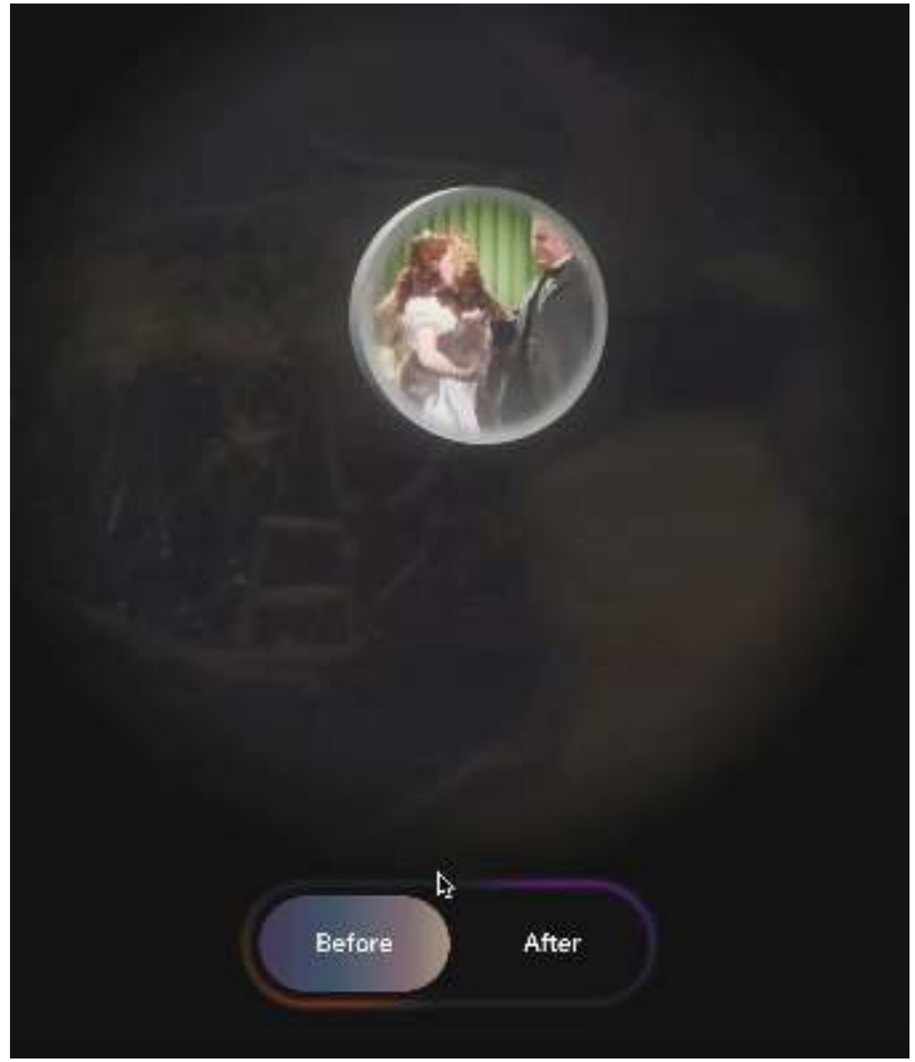
The Wizard of Oz @ The Sphere

Dima Damen
ICVSS2025

The Wizard of Oz @ The Sphere

The Wizard of Oz @ The Sphere

- Super-resolution,
- Outpainting…



https://behindthecurtain.withgoogle.com

Dima Damen
ICVSS2025

The Wizard of Oz @ The Sphere

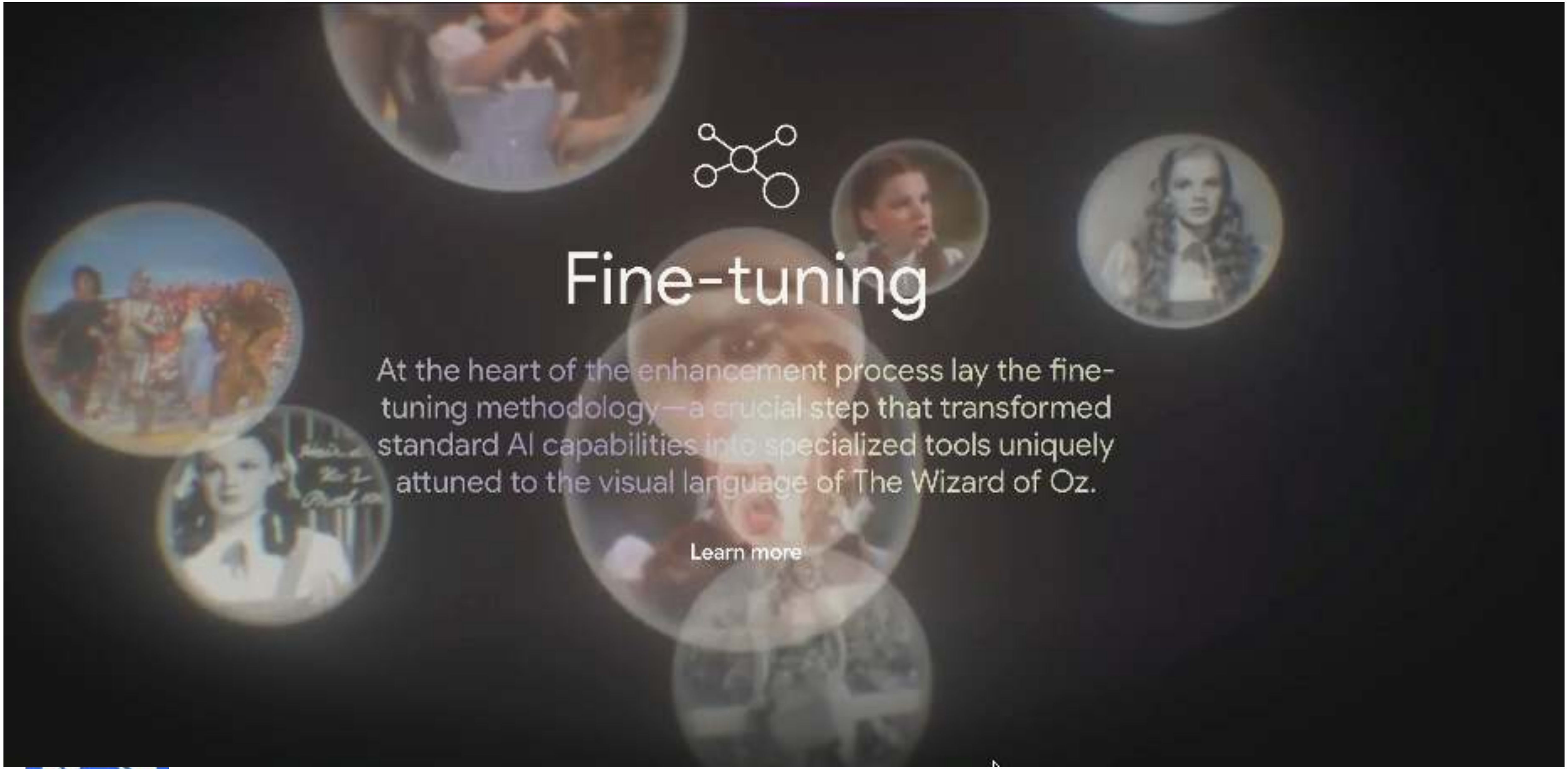
- Performance Interpolation,

The Wizard of Oz @ The Sphere

- Auto-Director

https://behindthecurtain.withgoogle.com

[redacted]



https://behindthecurtain.withgoogle.com

In today's tutorial

Motivation and Datasets in Egocentric Video Understanding

Video Understanding: Data and Tasks

Videos are Multimodal

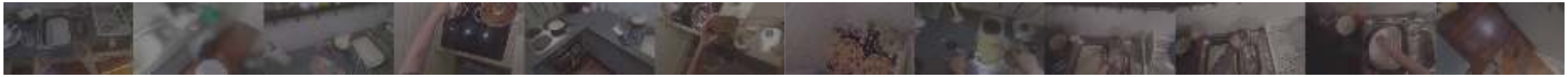
Connected Videos of One's Life

Video Understanding Out of the Frame

Teaser: The Wizard of Oz at the Sphere

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

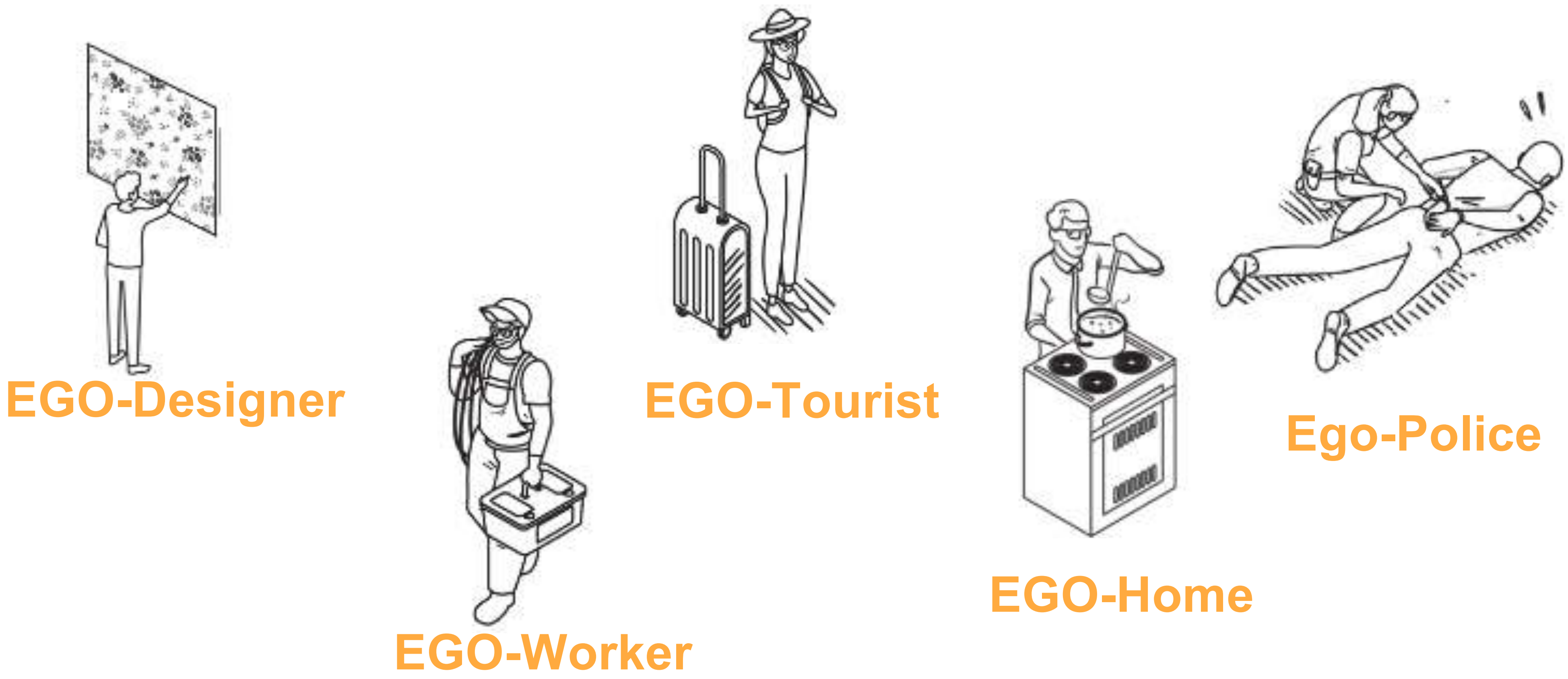
with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi

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

How did we do this?

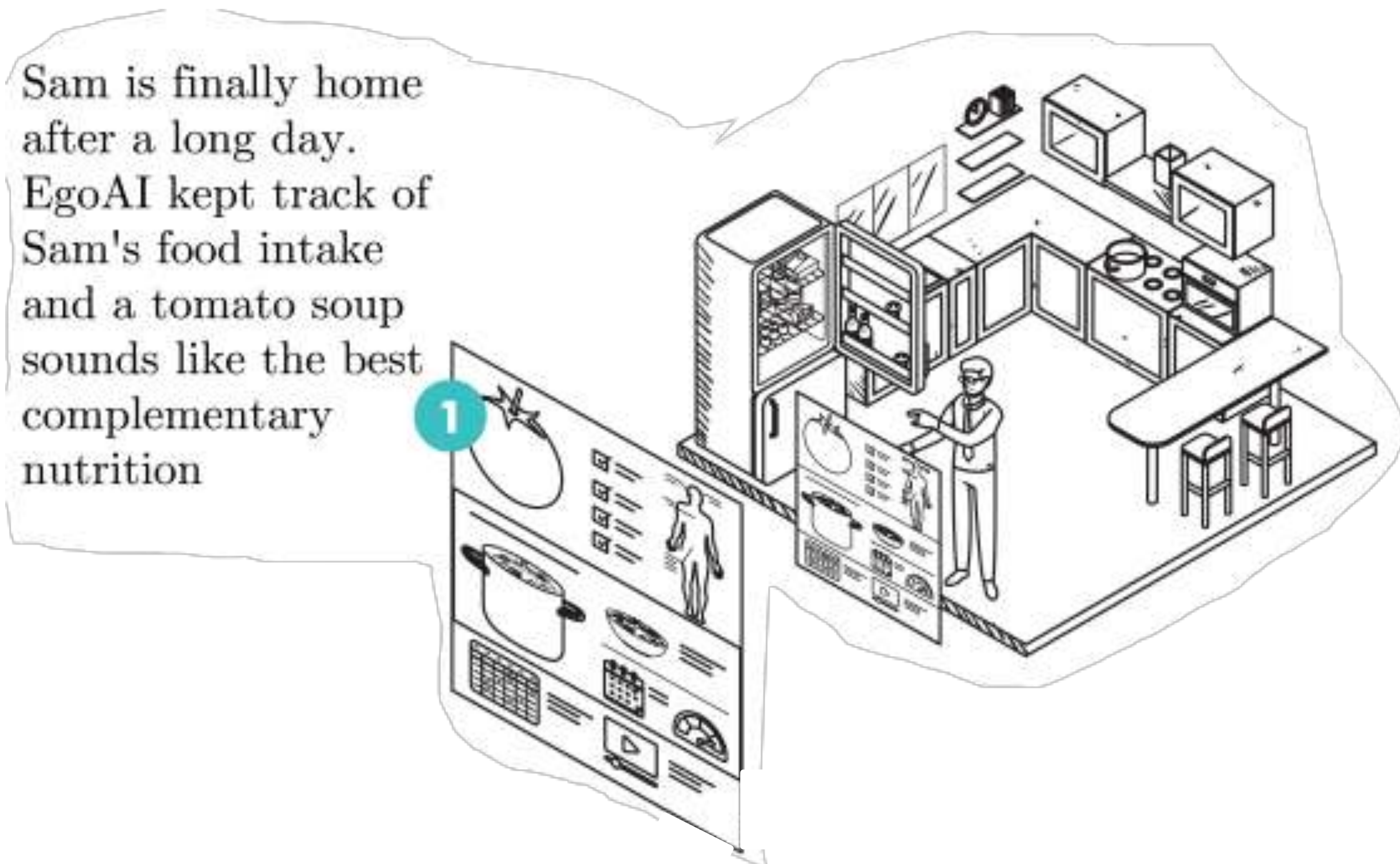
with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi

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



EGO-Home

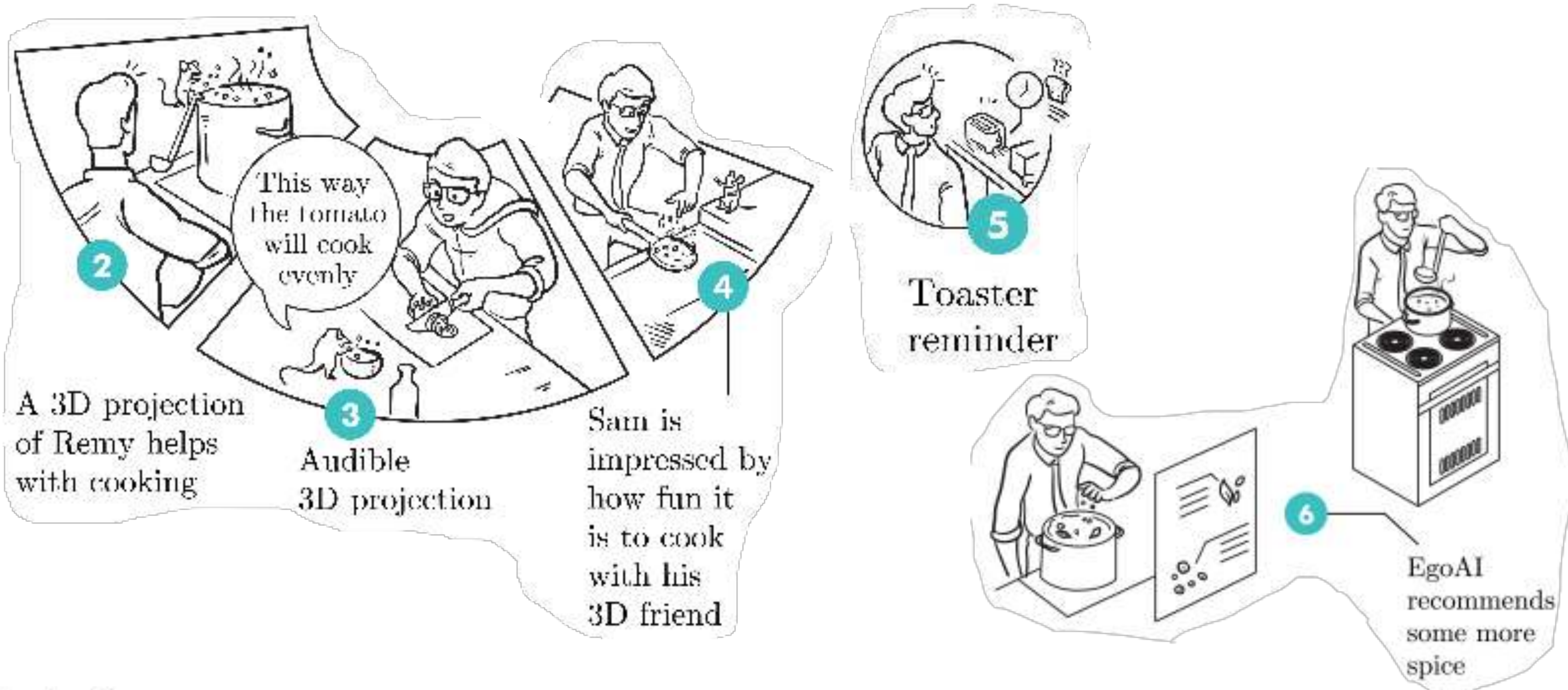
with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi

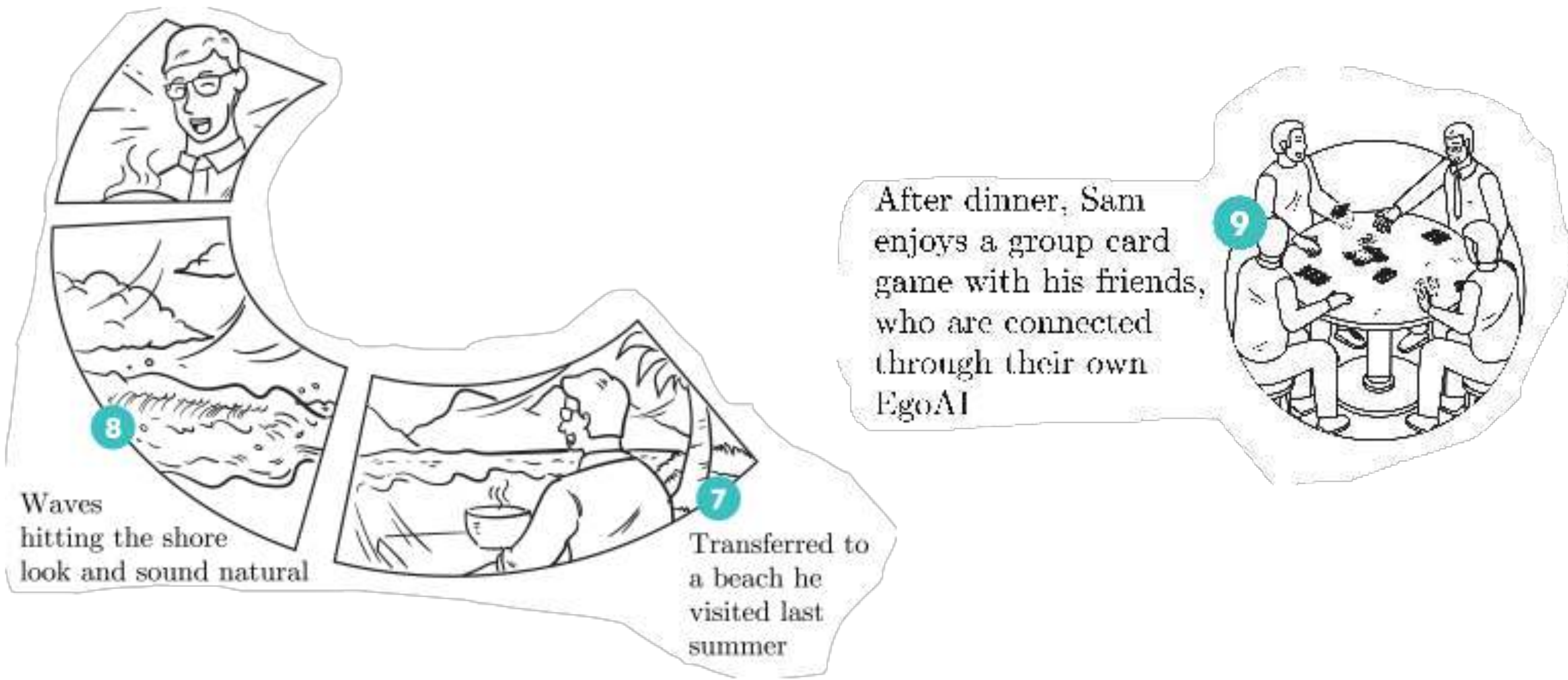


C Plizzari et al (2025). An Outlook into the Future of Egocentric Vision. IJCV

EGO-Home

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

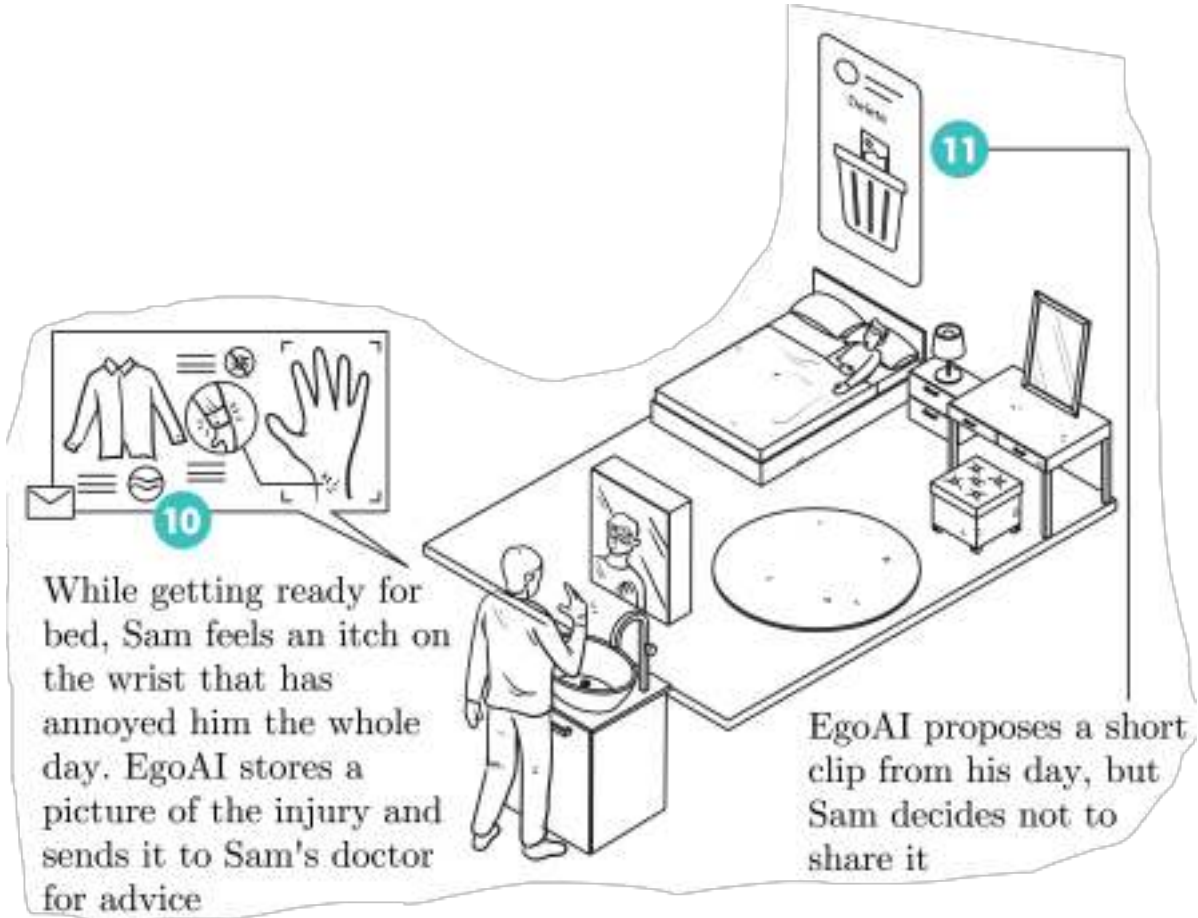




8
Waves hitting the shore look and sound natural
7
Transferred to a beach he visited last summer
After dinner, Sam enjoys a group card game with his friends, who are connected through their own EgoAI
9

EGO-Home

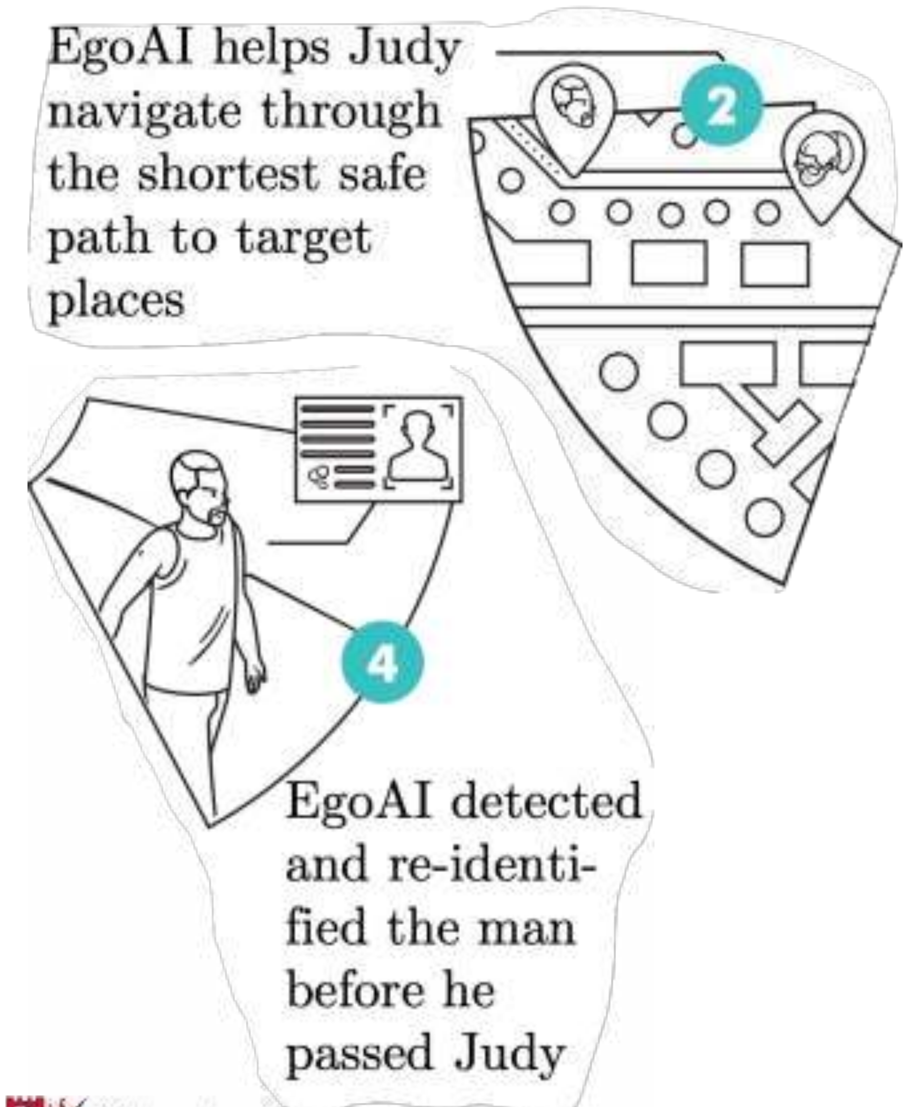
with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi

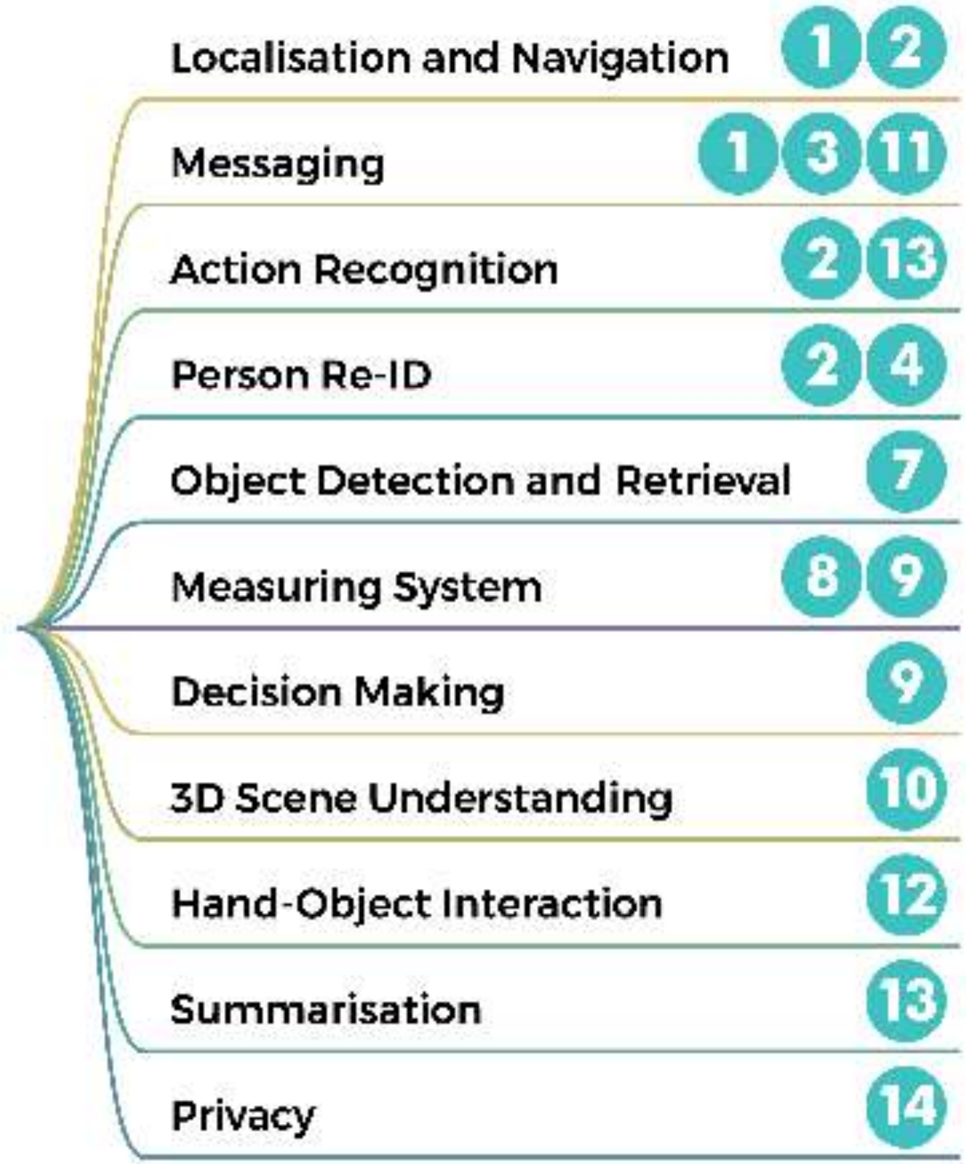


C Plizzari et al (2025). An Outlook into the Future of Egocentric Vision. IJCV

From Stories to Tasks

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



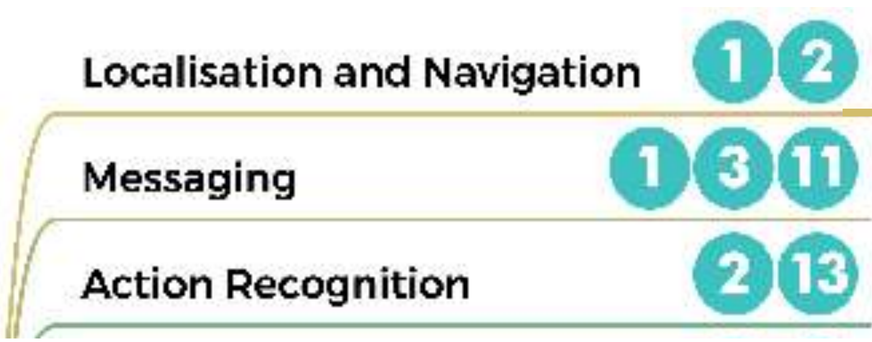


University of BRISTOL

C Plizzari et al (2025). An Outlook into the Future of Egocentric Vision. IJCV

The Survey Part

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



Seminal Works

State-of-the-art

Datasets

For the Future

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

C Plizzari et al (2025). An Outlook into the Future of Egocentric Vision. IJCV

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

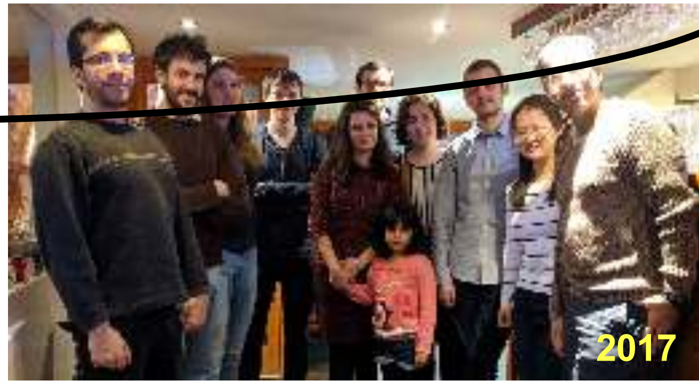
Teaser: The Wizard of Oz at the Sphere

Outlook into the Future of Egocentric Vision

Conclusion

The Team

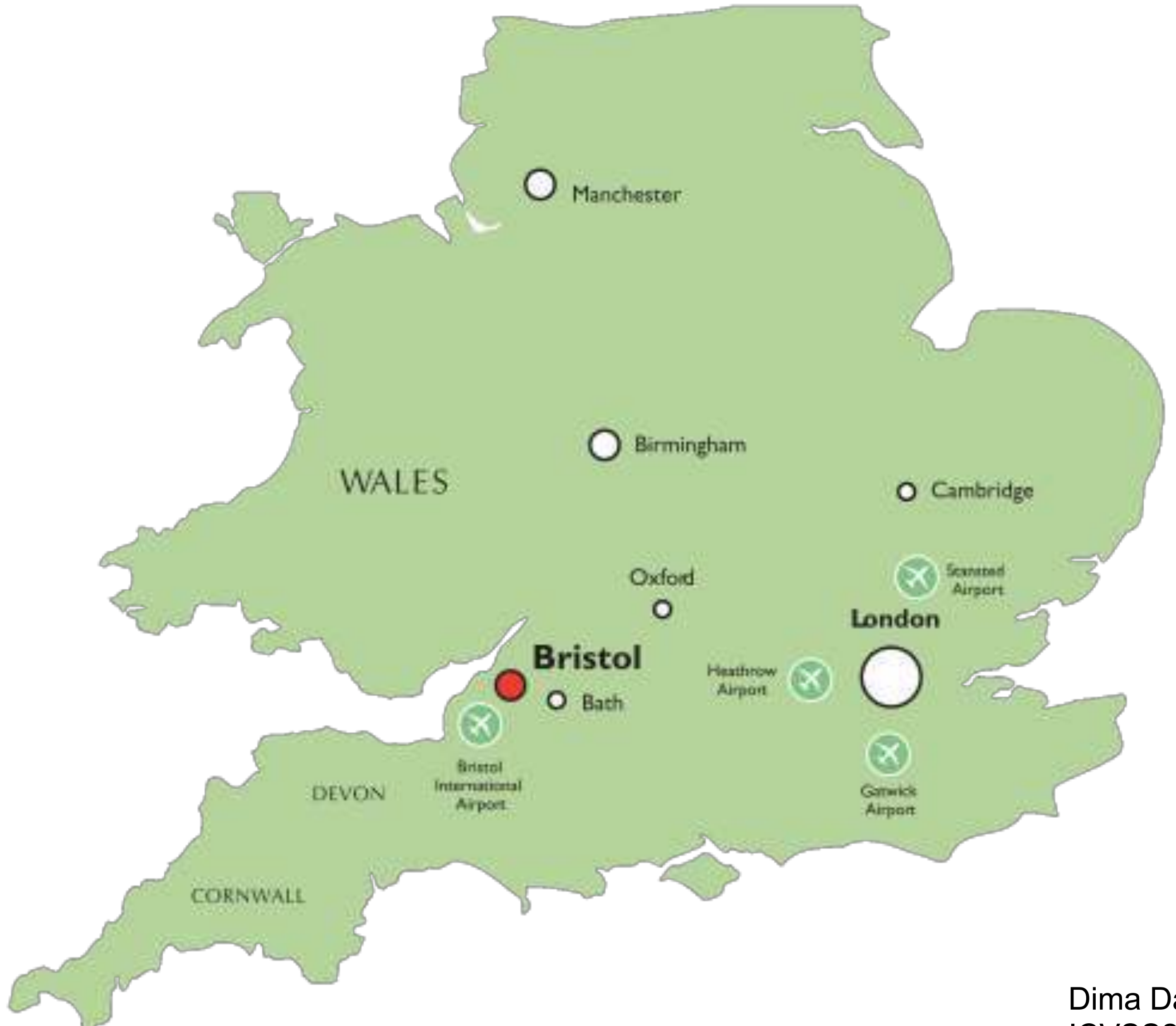
grateful



My research team…

grateful

Visit us…



University of BRISTOL

Thank you

For further info, datasets, code, publications…

http://dimadamen.github.io

@dimadamen

@dimadamen.bsky.social

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

Q&A