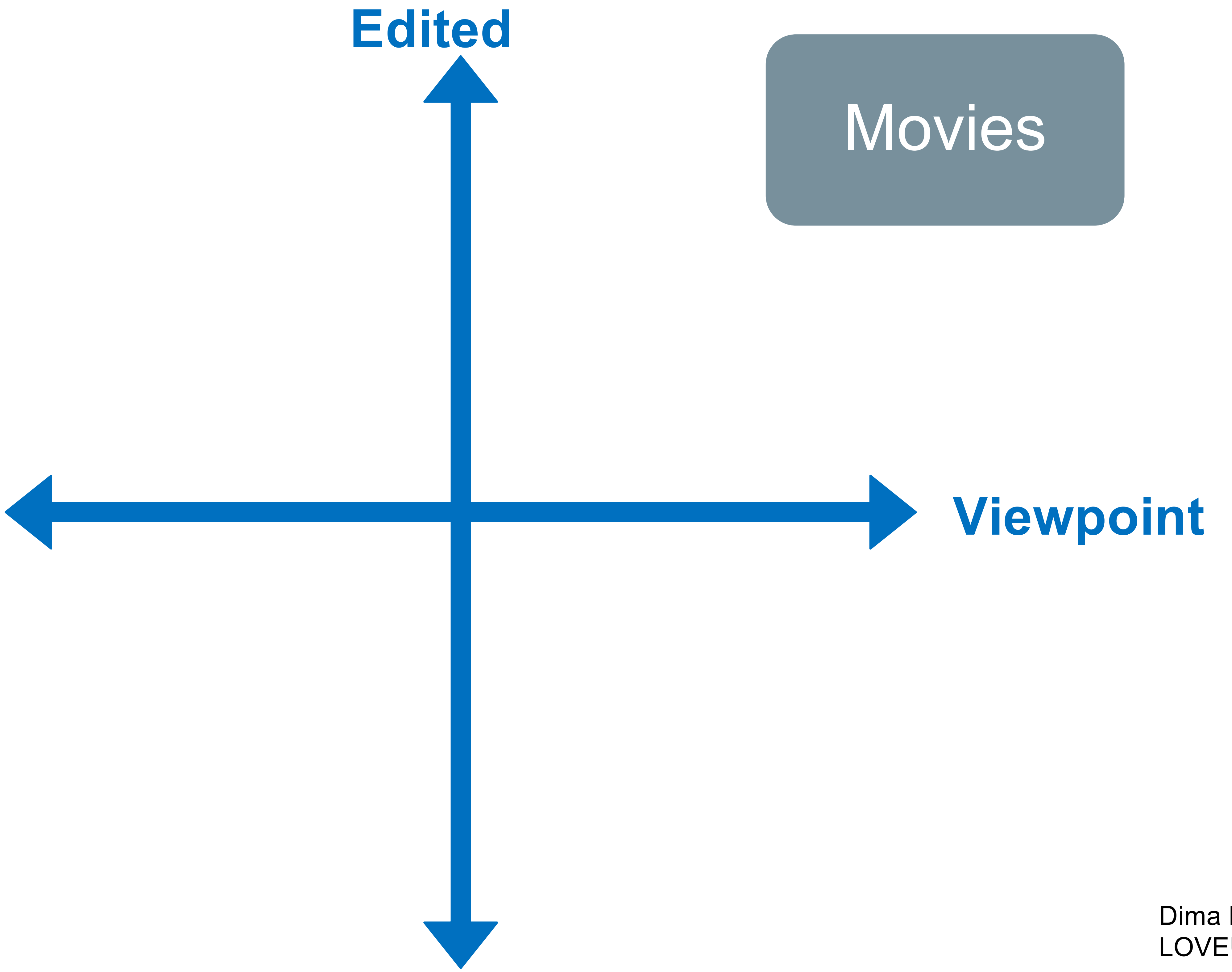
Long-Form Egocentric Video Understanding

Towards Multi-Modal AI Assistant

University of BRISTOL

Dima Damen
LOVEU @CVPR2024

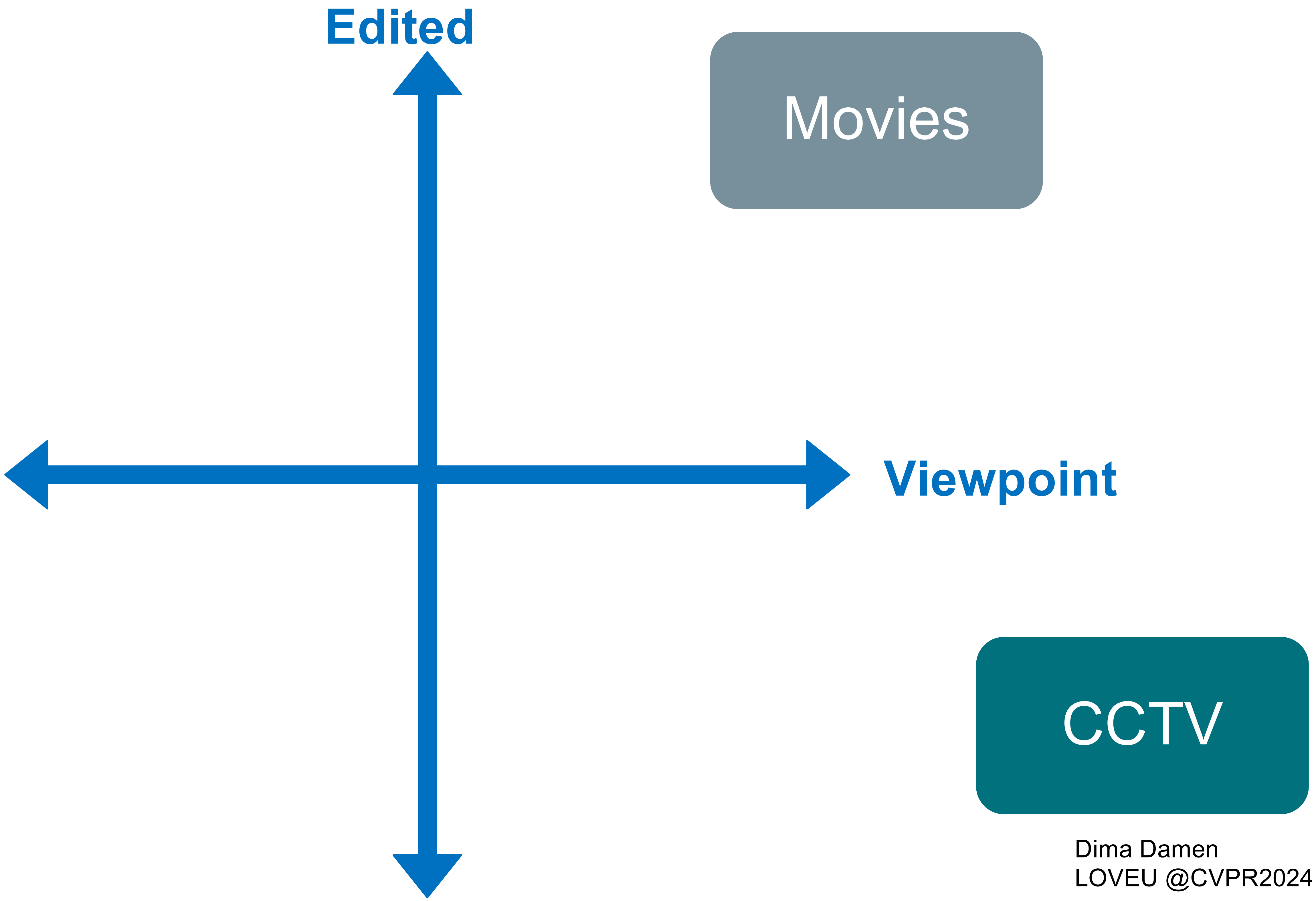
The history of Long-Form Video Understanding



The history of Long-Form Video Understanding

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

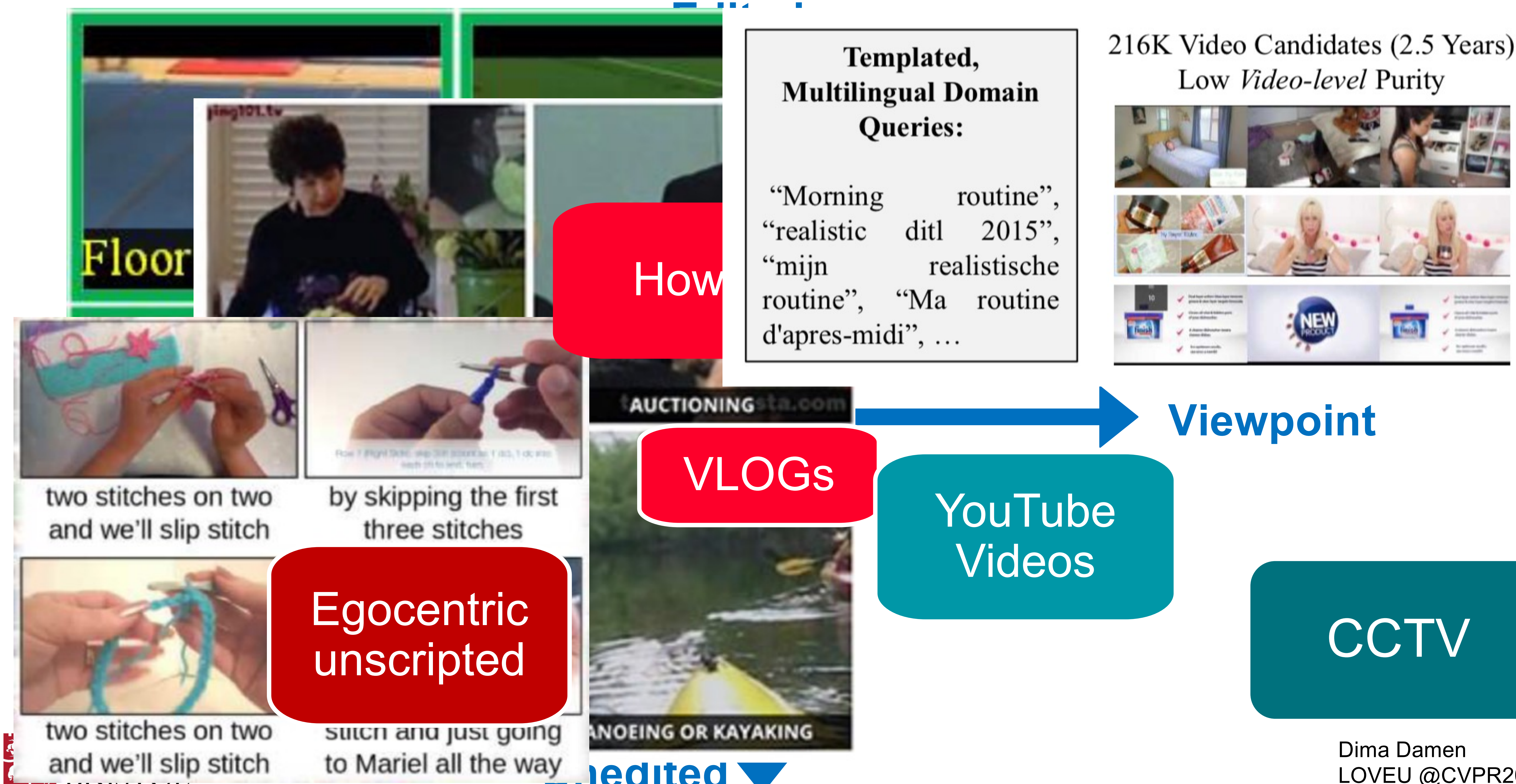
The history of Long-Form Video Understanding



The history of Long-Form Video Understanding

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

The history of Long-Form Video Understanding



Viewpoint

YouTube Videos

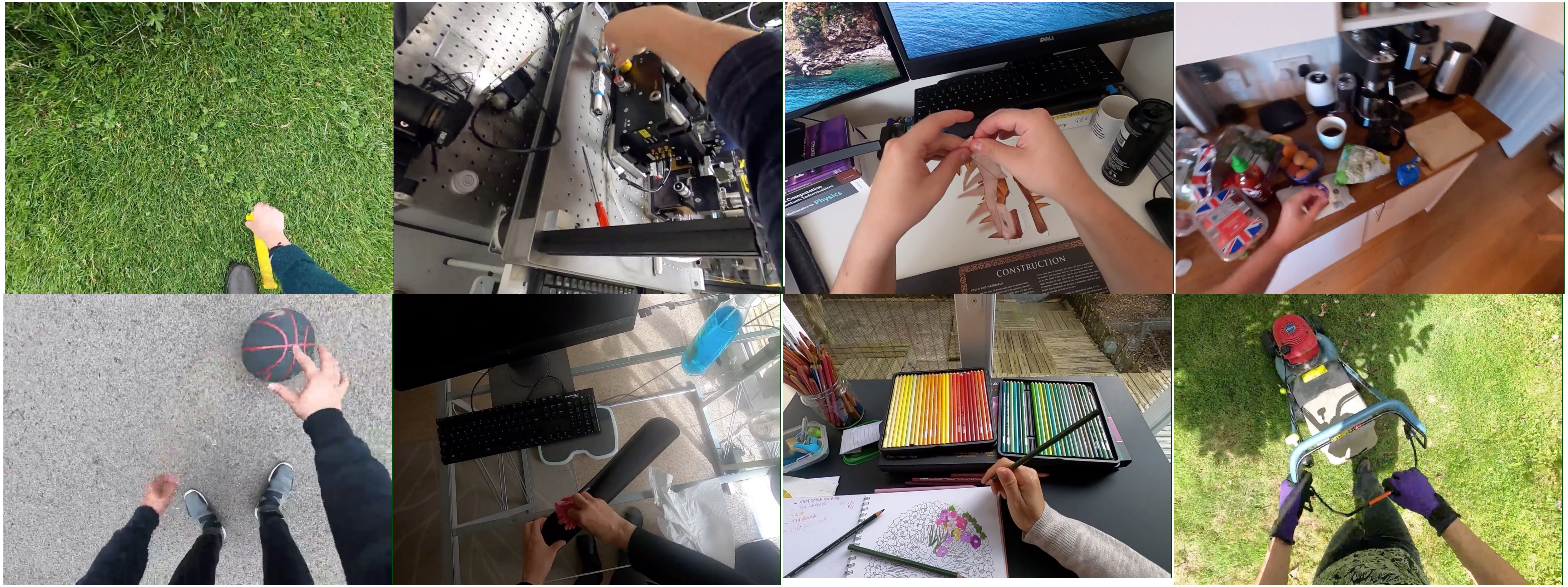
CCTV

Unedited

Dima Damen
LOVEU @CVPR2024

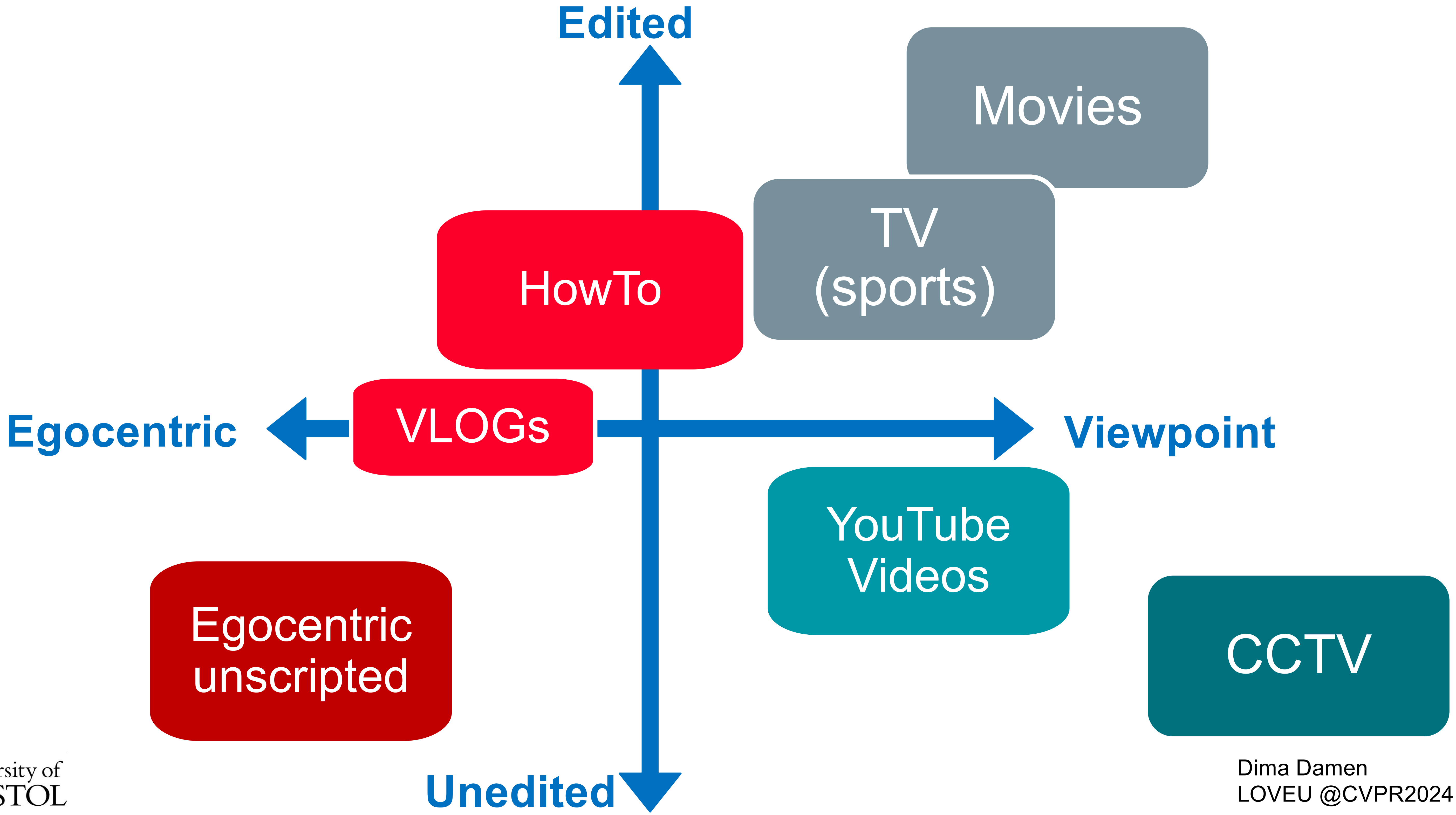
Ego4D

with: Kristen Grauman
+83 authors



K Grauman et al (2022). Around the World in 3,000 Hours of Egocentric Video. CVPR

The history of Long-Form Video Understanding



The history of Long-Form Video Understanding

Movies

TV (sports)

HowTo

VLOGs

YouTube Videos

Egocentric unscripted

CCTV

Long-Form Understanding

Speech/Plot	Movies	HowTo	VLOGs
Edits/Shots	Movies	HowTo	
Audio-Visual	Movies	YouTube	Egocentric
Hand-Obj		HowTo	Egocentric
Guidance/ Assistance		HowTo	Egocentric

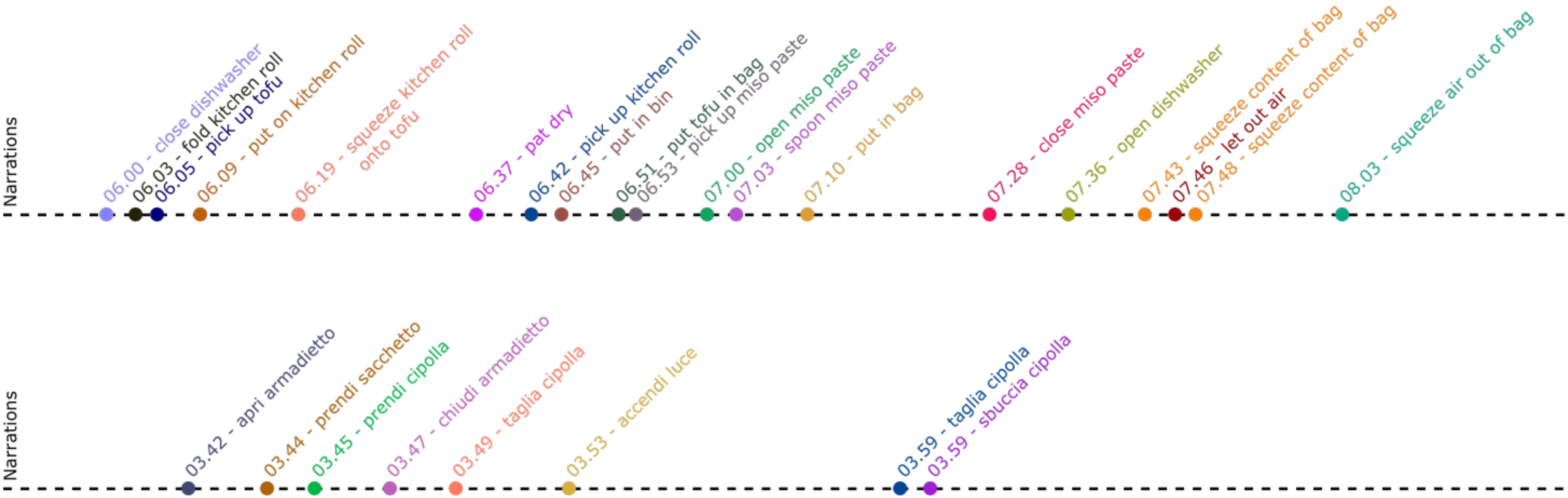
Long-Form Egocentric Video Understanding

- No Semantic Supervision
- No Shots - Temporal Alignment
- Audio-Visual Semantic Gap
- Quick View Changes
- Repeating Actions
- Long Continuous Streams

Long-Form Egocentric Video Understanding

- No Semantic Supervision
- No Shots - Temporal Alignment
- Audio-Visual Semantic Gap
- Quick View Changes
- Repeating Actions
- Long Continuous Streams

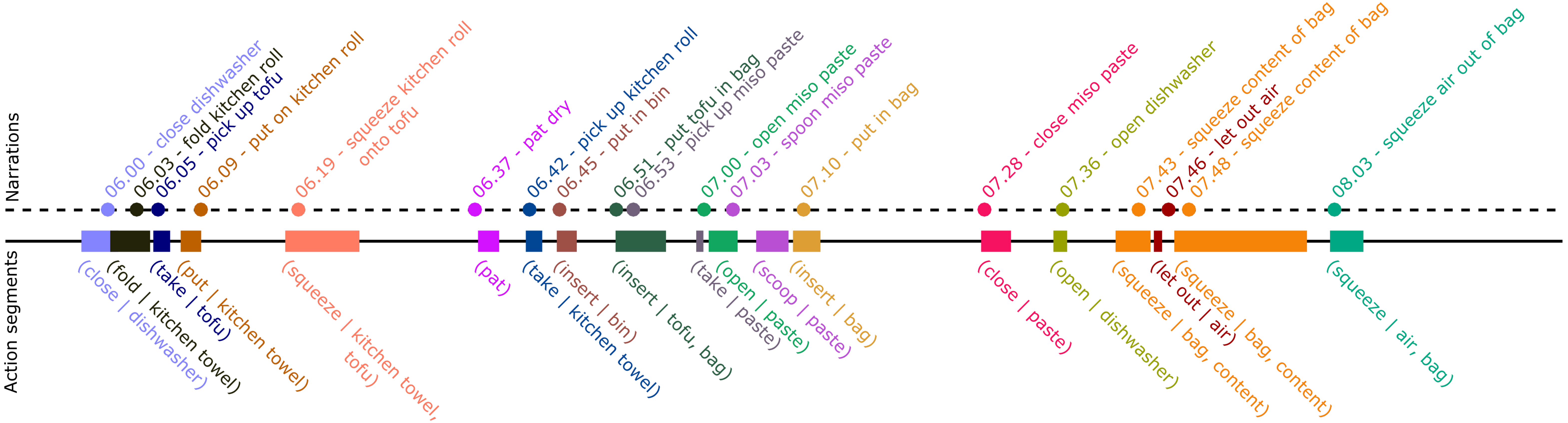
EPIC-KITCHENS Narrating Egocentric Videos



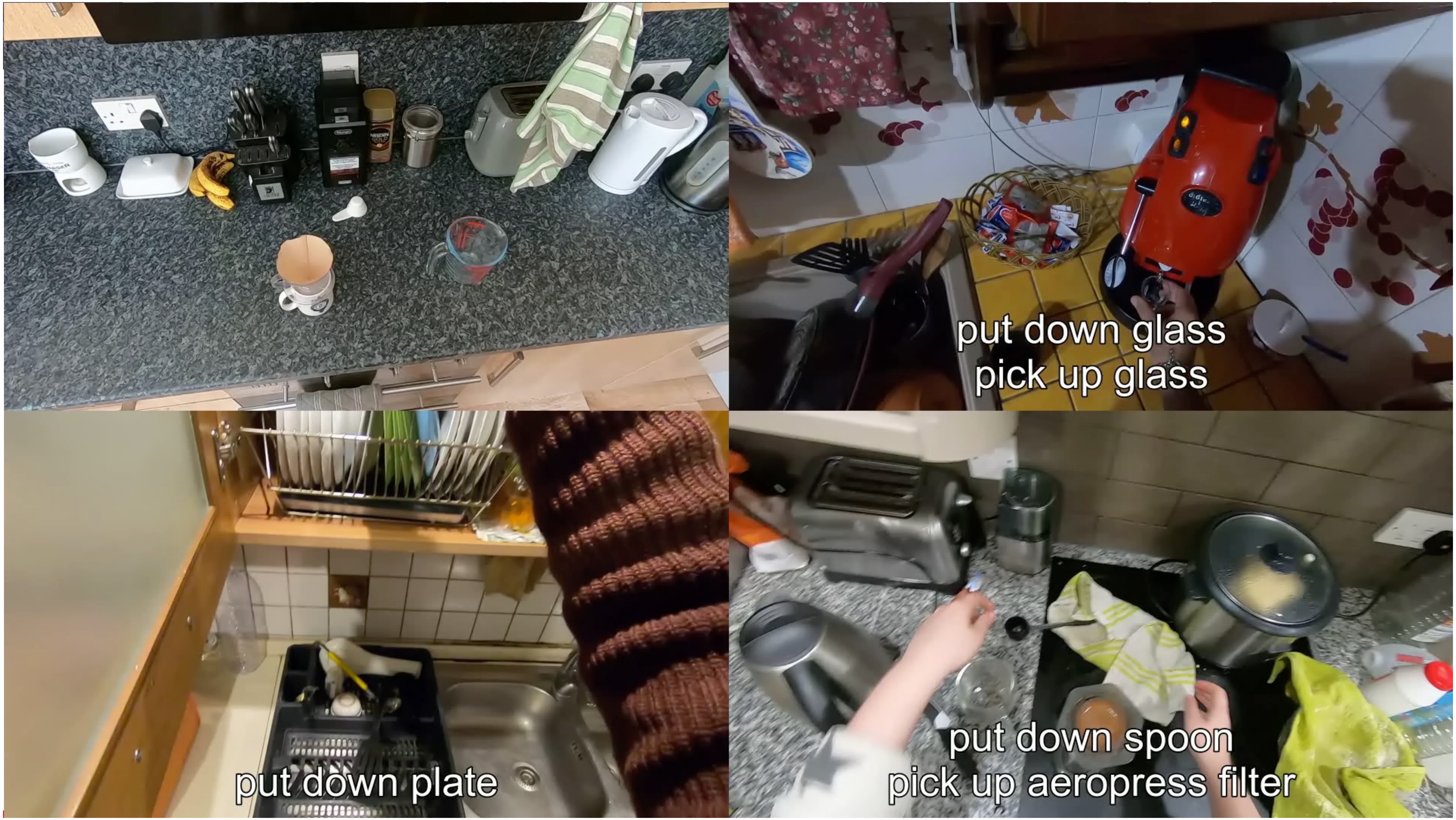
University of BRISTOL

Dima Damen
LOVEU @CVPR2024

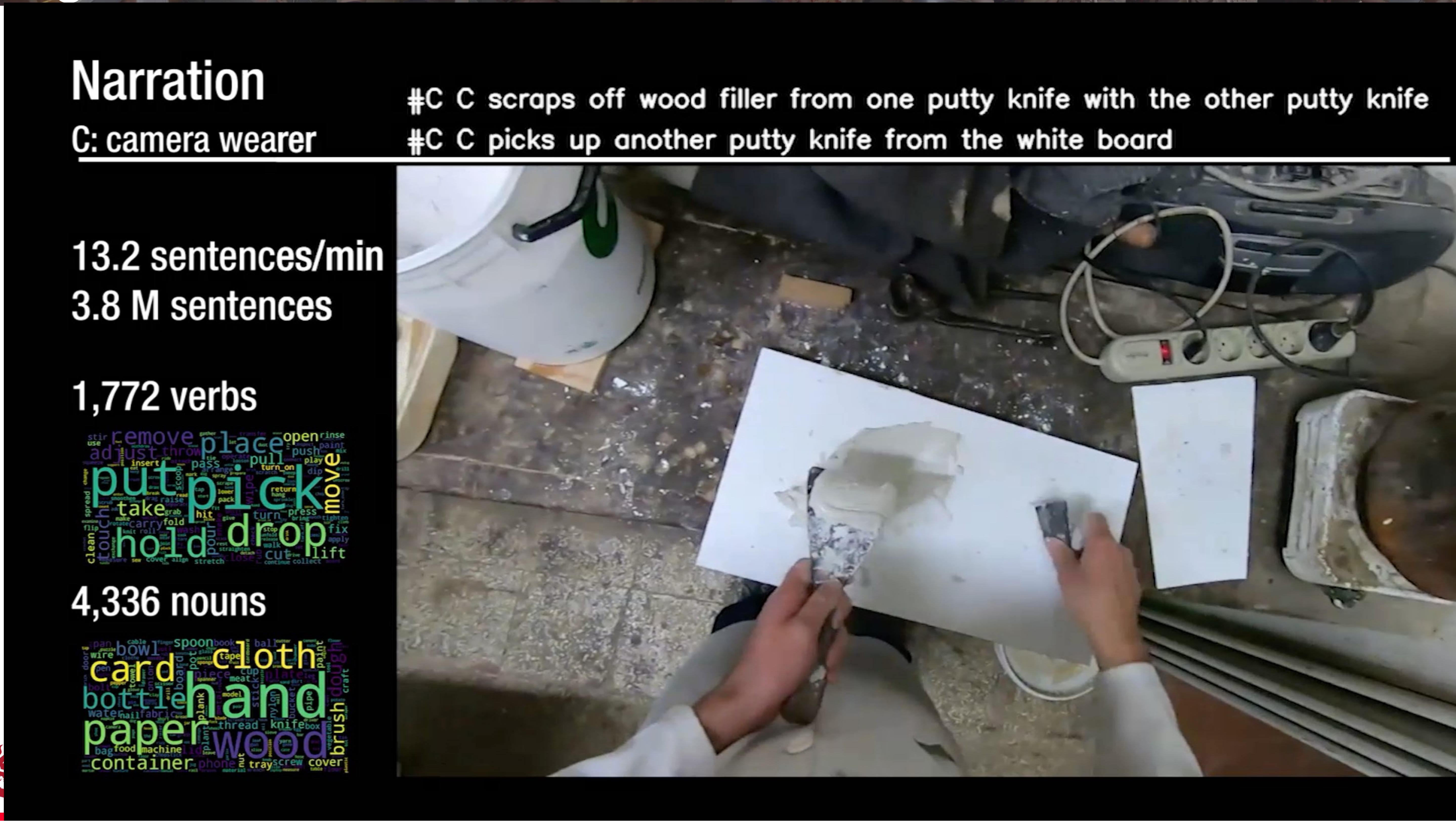
EPIC-KITCHENS Narrating Egocentric Videos




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

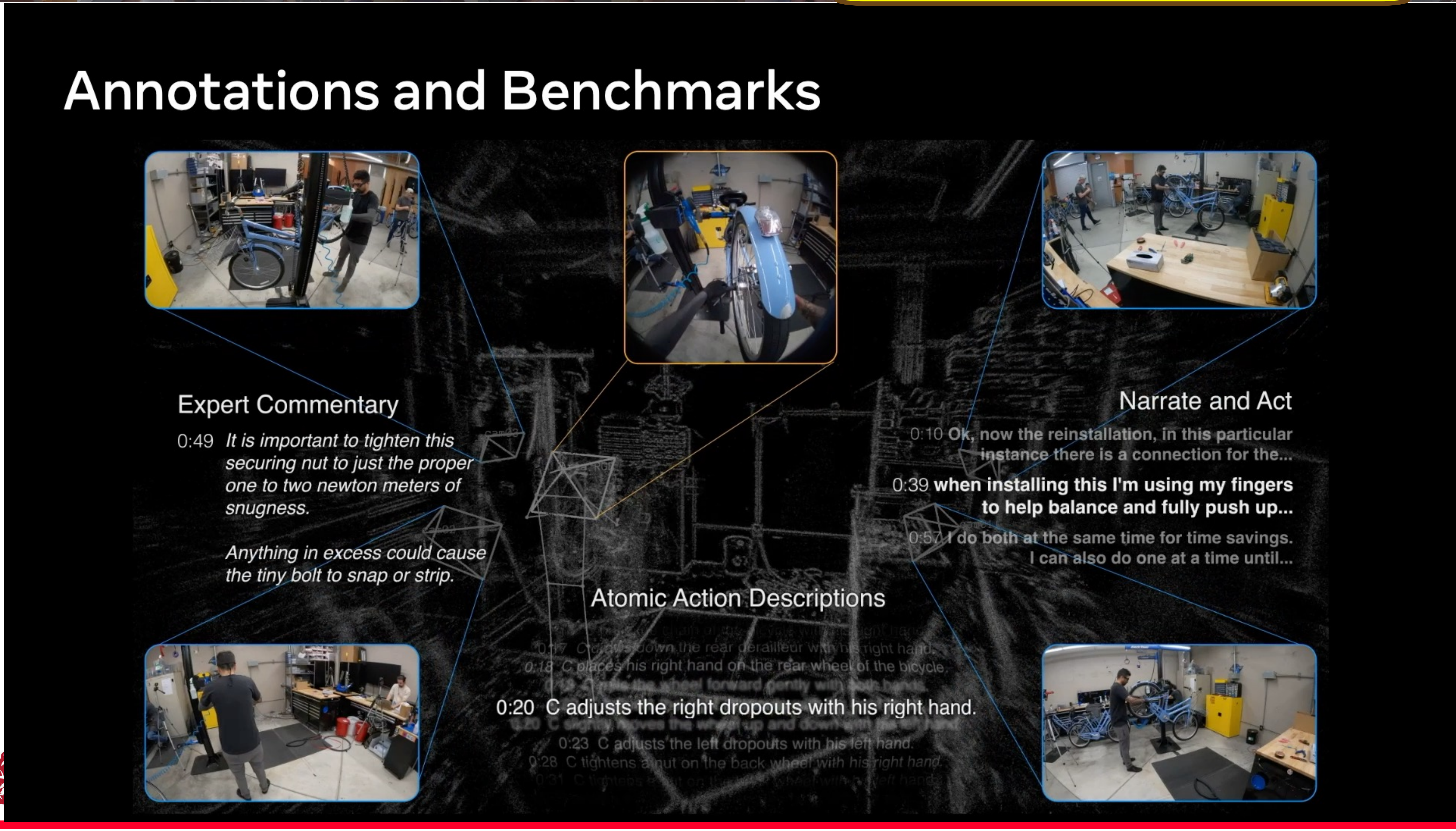

put down glass
pick up glass
put down plate
put down spoon
pick up aeropress filter

Ego4D Narrations



Ego-Exo4D Annotations

Fri Oral + Poster (Session 5 #292)

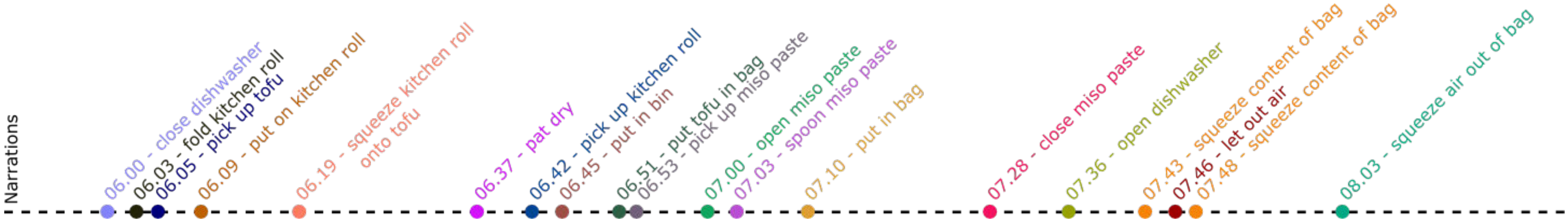


Long-Form Egocentric Video Understanding

- No Semantic Supervision
- No Shots - Temporal Alignment
- Audio-Visual Semantic Gap
- Quick View Changes
- Repeating Actions
- Long Continuous Streams

Learning from a Single Timestamp

with: Davide Moltisanti
Sanja Fidler



D Moltisanti et al (2019). Action Recognition from Single Timestamp Supervision in Untrimmed Videos. *Computer Vision and Pattern Recognition (CVPR)*

Learning from a Single Timestamp

with: Davide Moltisanti
Sanja Fidler

D Moltisanti et al (2019). Action Recognition from Single Timestamp Supervision in Untrimmed Videos. *Computer Vision and Pattern Recognition (CVPR)*

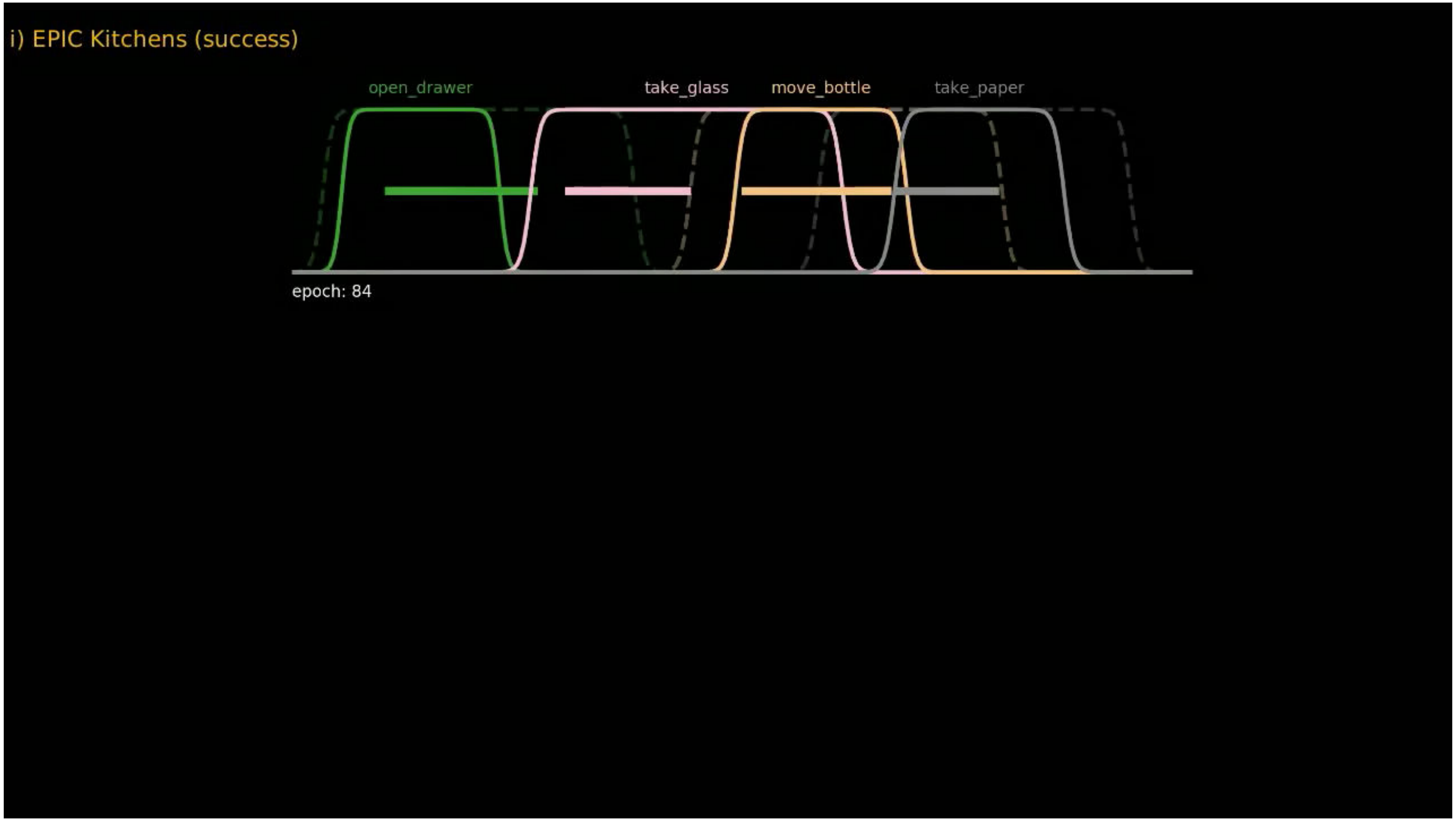
Learning from a Single Timestamp

with: Davide Moltisanti
Sanja Fidler

D Moltisanti et al (2019). Action Recognition from Single Timestamp Supervision in Untrimmed Videos. *Computer Vision and Pattern Recognition (CVPR)*

Learning from a Single Timestamp

with: Davide Moltisanti
Sanja Fidler



D Moltisanti et al (2019). Action Recognition from Single Timestamp Supervision in Untrimmed Videos. *Computer Vision and Pattern Recognition (CVPR)*

Long-Form Egocentric Video Understanding

- No Semantic Supervision
- No Shots - Temporal Alignment
- Audio-Visual Semantic Gap
- Quick View Changes
- Repeating Actions
- Long Continuous Streams

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

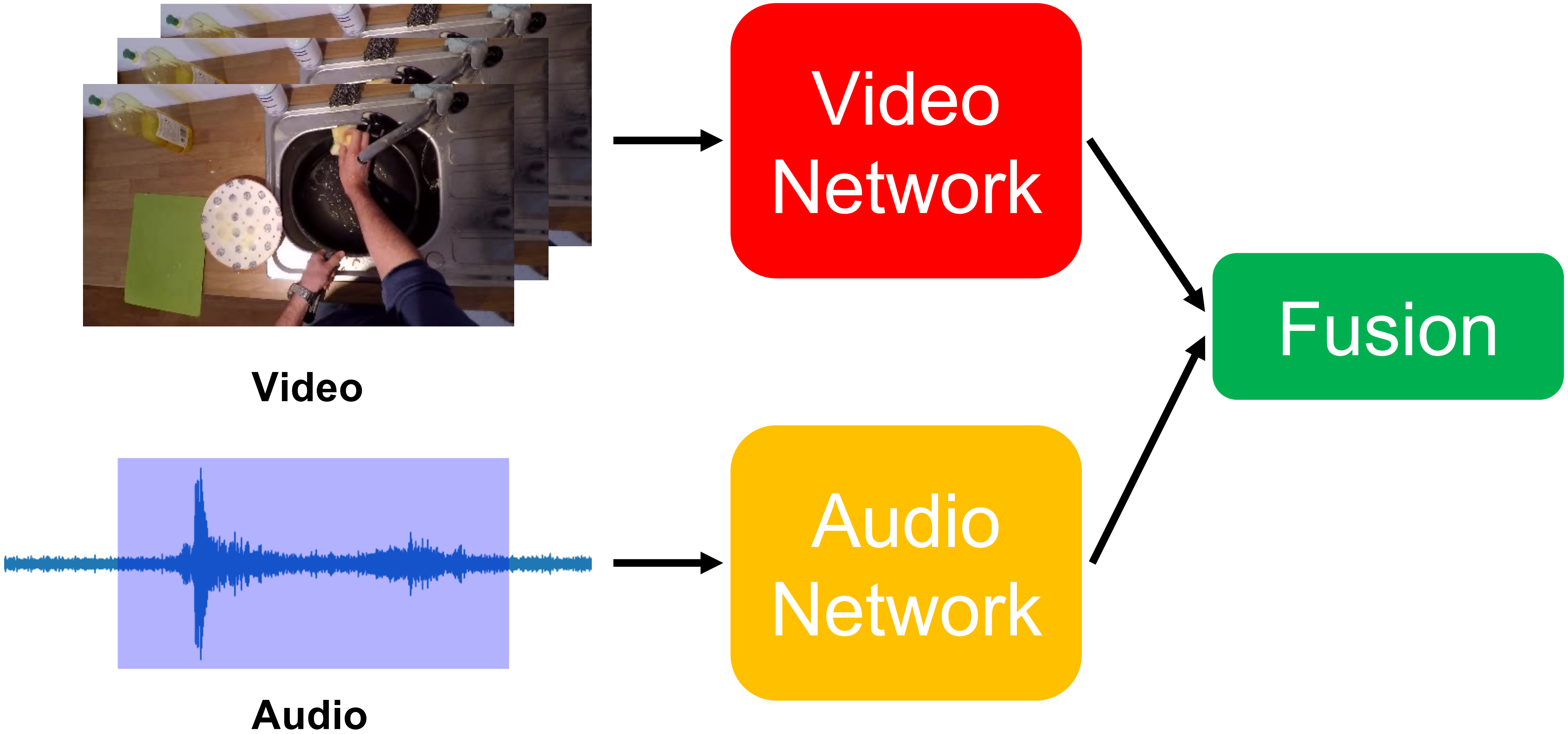
Jaesung Huh*, Jacob Chalk*, Evangelos Kazakos, Dima Damen, Andrew Zisserman
* : Equal contribution

VGG University of Oxford

University of Bristol

Current Audio-Visual Approaches

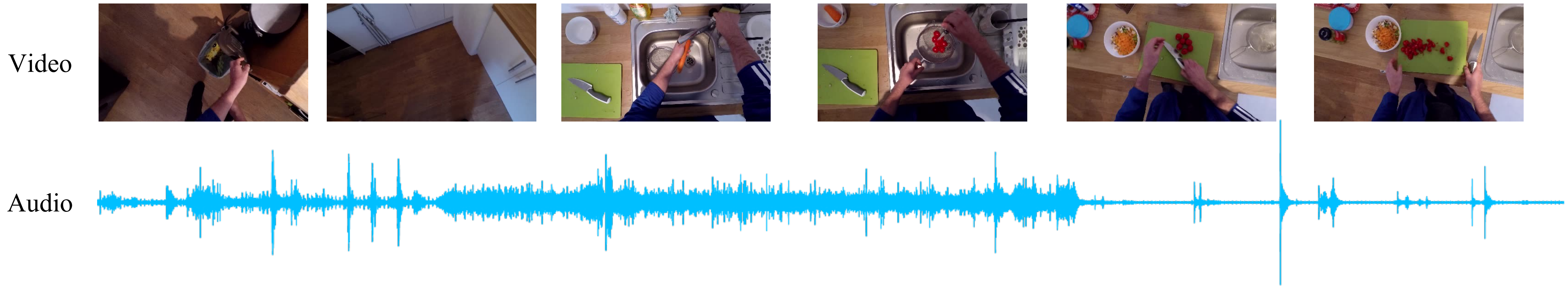
with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman



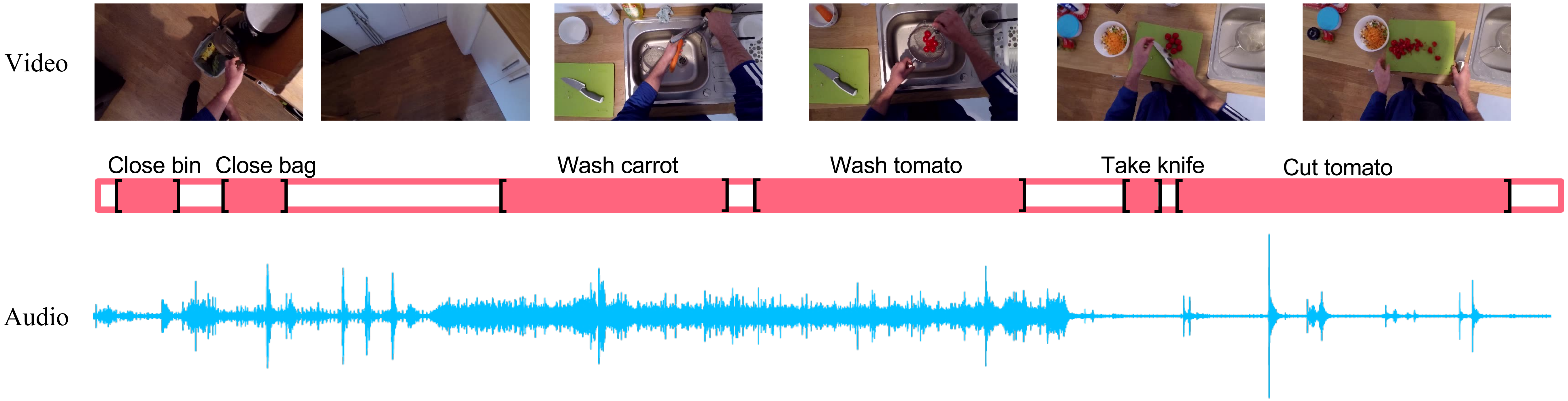
J Huh*, J Chalk* E Kazakos, D Damen, A Zisserman (2023). EPIC-Sounds: A Large-scale Dataset of Actions That Sound. ICASSP

Motivation

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



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



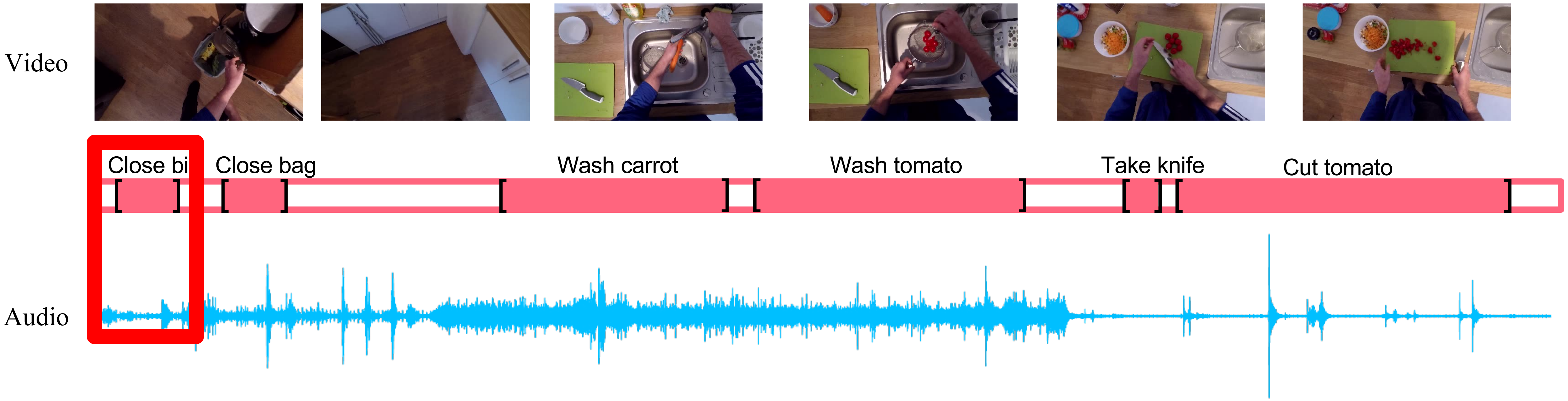
Video
Close bin
Close bag
Wash carrot
Wash tomato
Take knife
Cut tomato
Audio

Motivation

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

Motivation

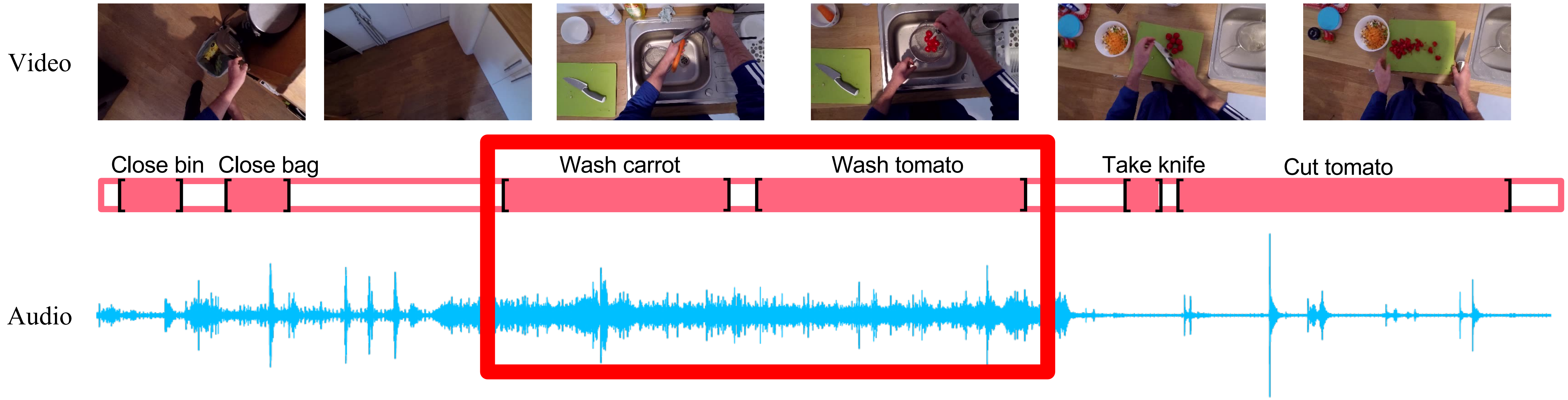
with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman



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

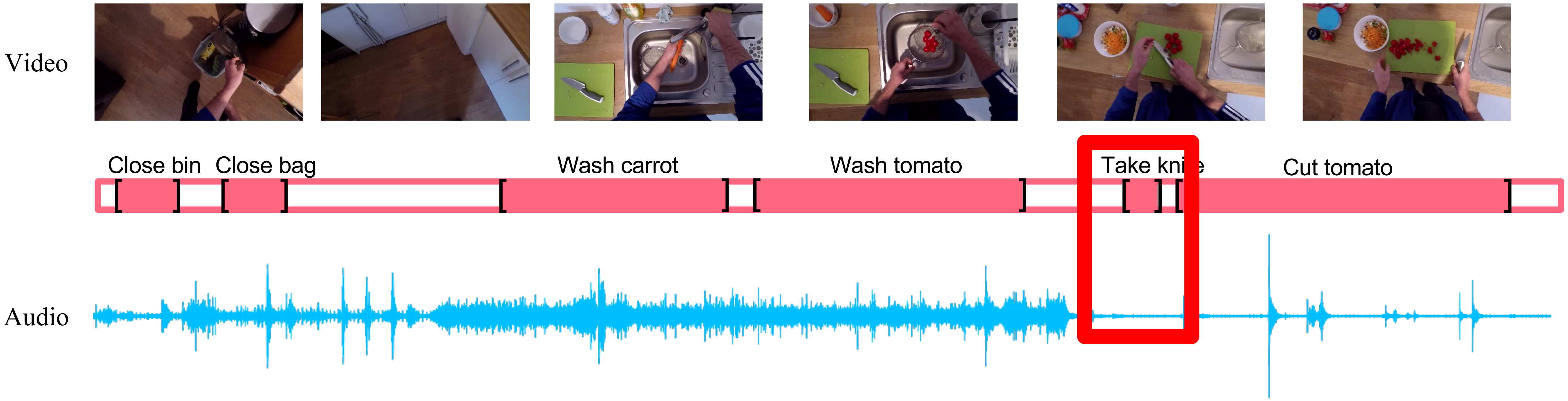
Motivation

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



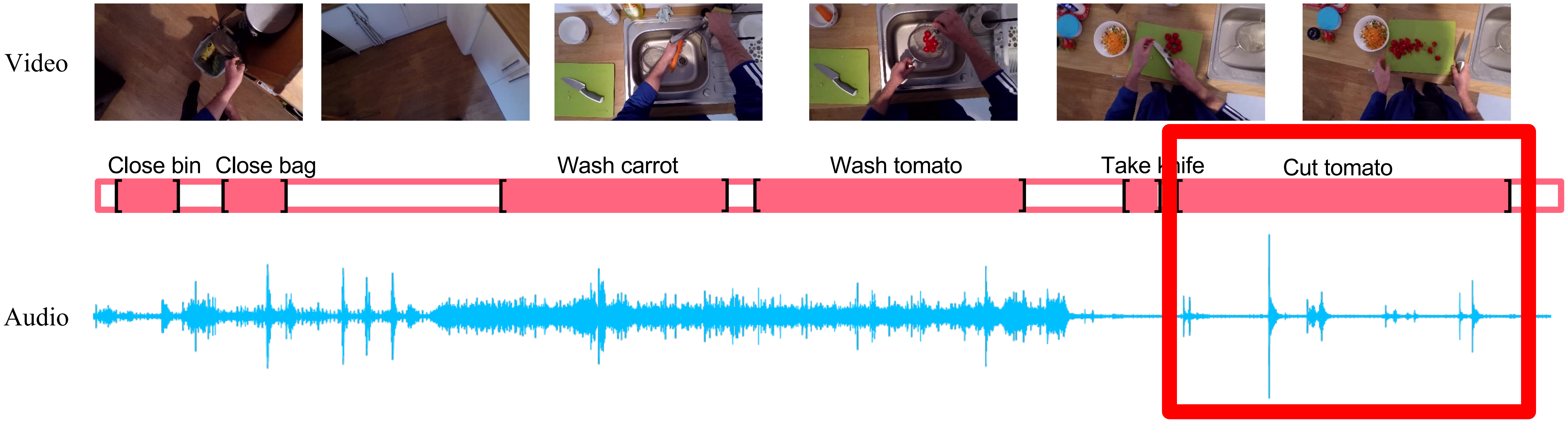
Motivation

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



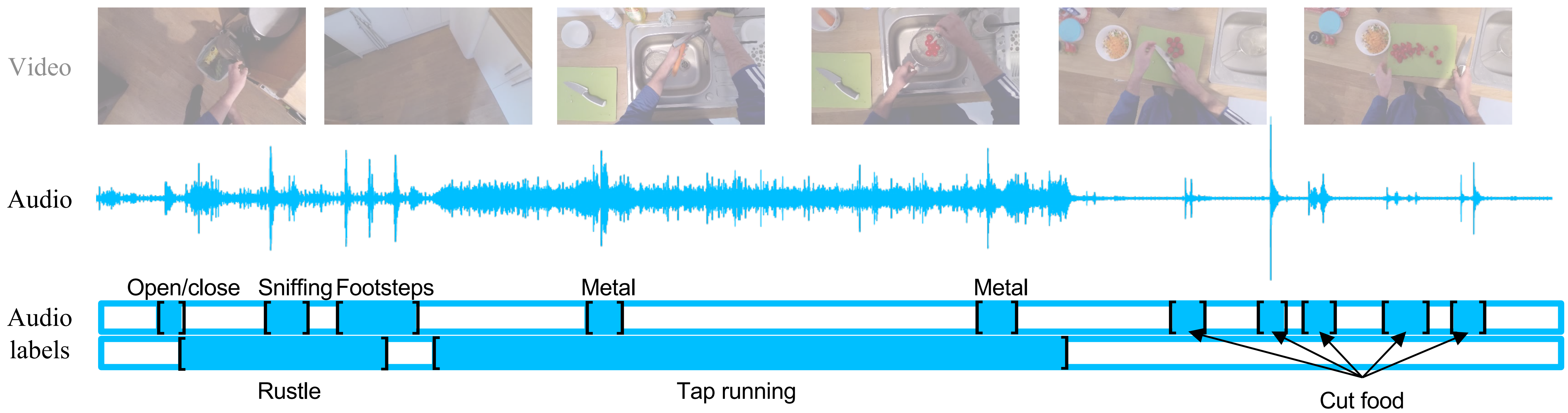
Motivation

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



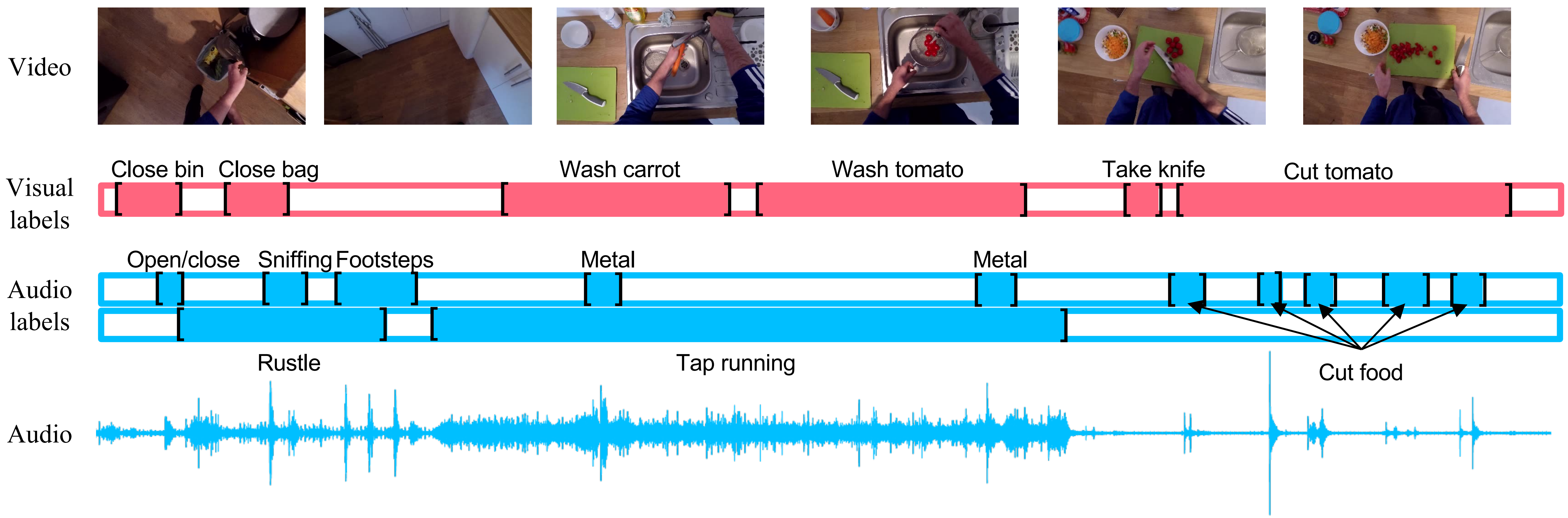
Motivation

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



Motivation

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



EPIC-SOUNDS

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

EPIC-KITCHENS VIDEOS
100 hours
45 kitchens

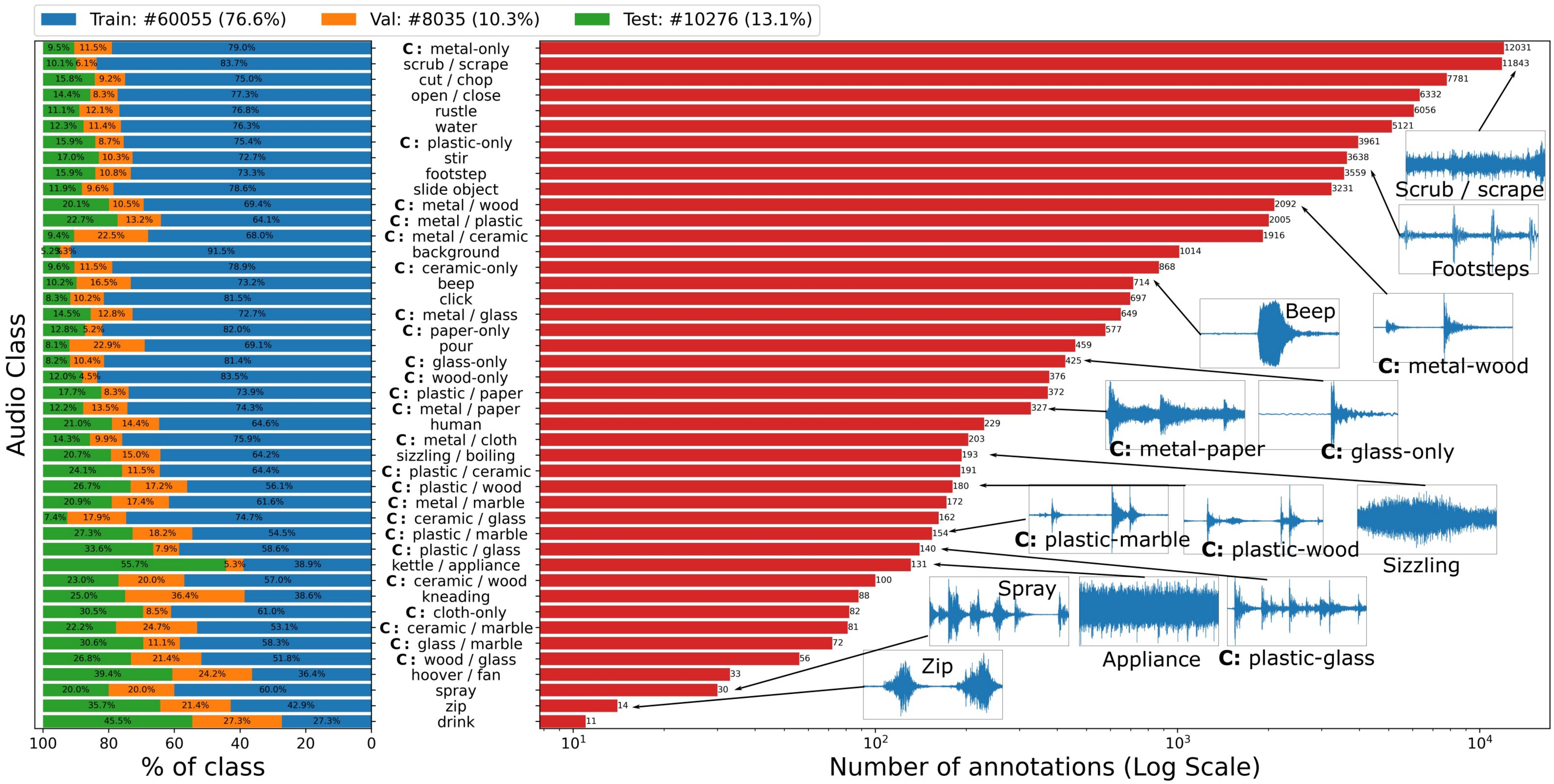
Visual Action Annotations
90K visual actions
97 verb classes
300 noun classes

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


spray

EPIC-SOUNDS

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

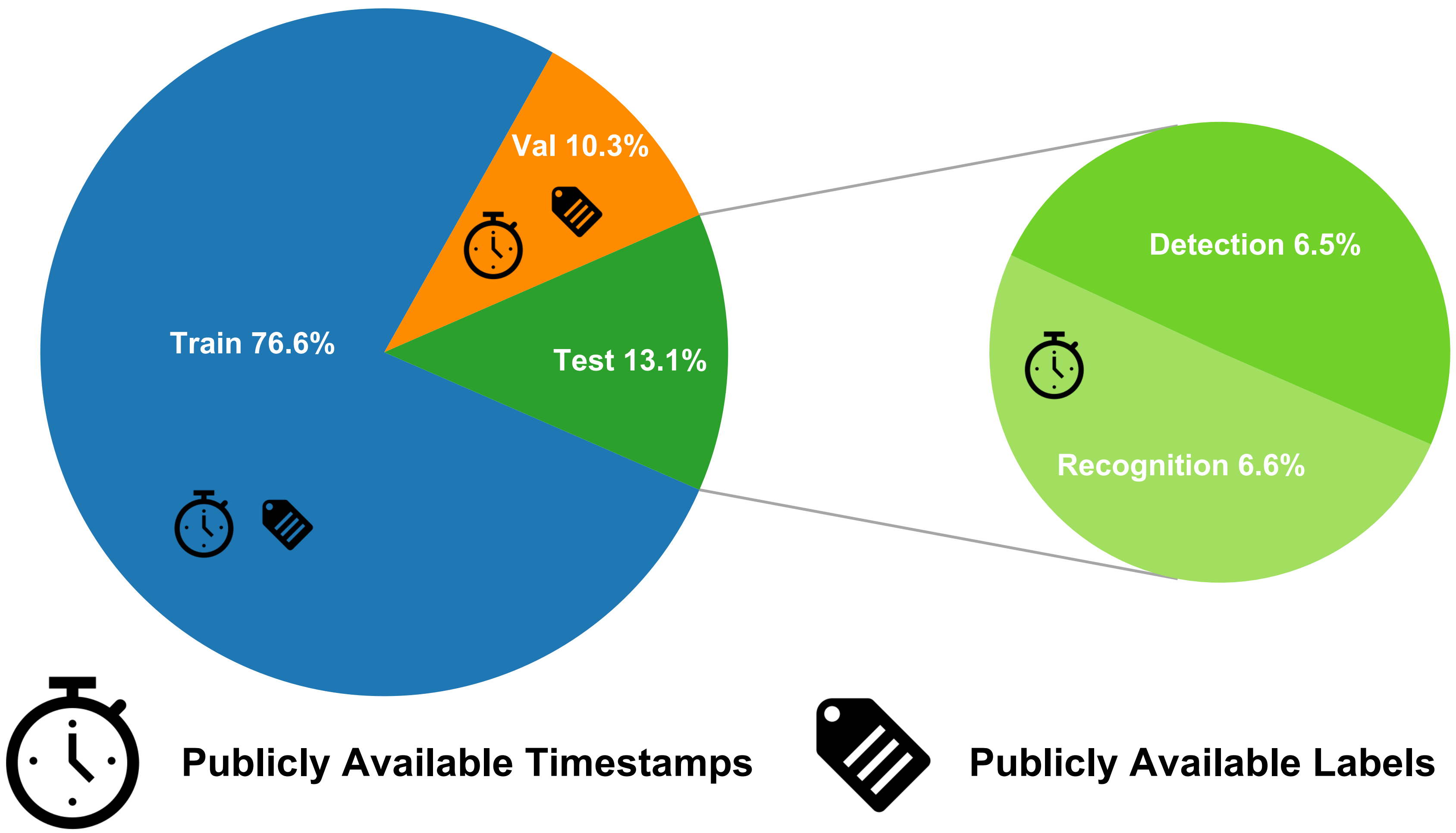


BRISTOL

EPIC-SOUNDS splits

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

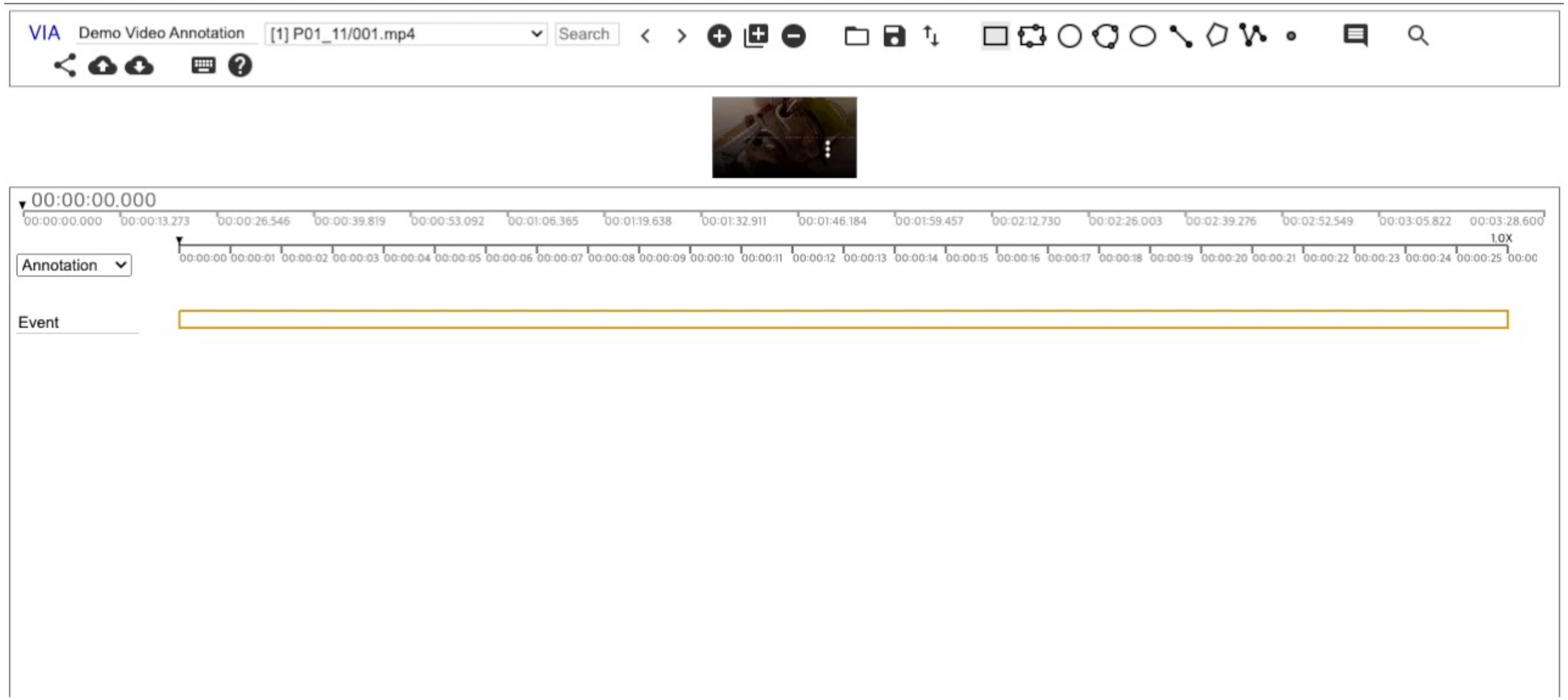
- We match the train/validation/test video splits from EPIC-KITCHENS-100
- We halve the test split into two challenge-specific subsets:
  - Recognition – with timestamps
  - Detection – without timestamps



Annotations Pipeline

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

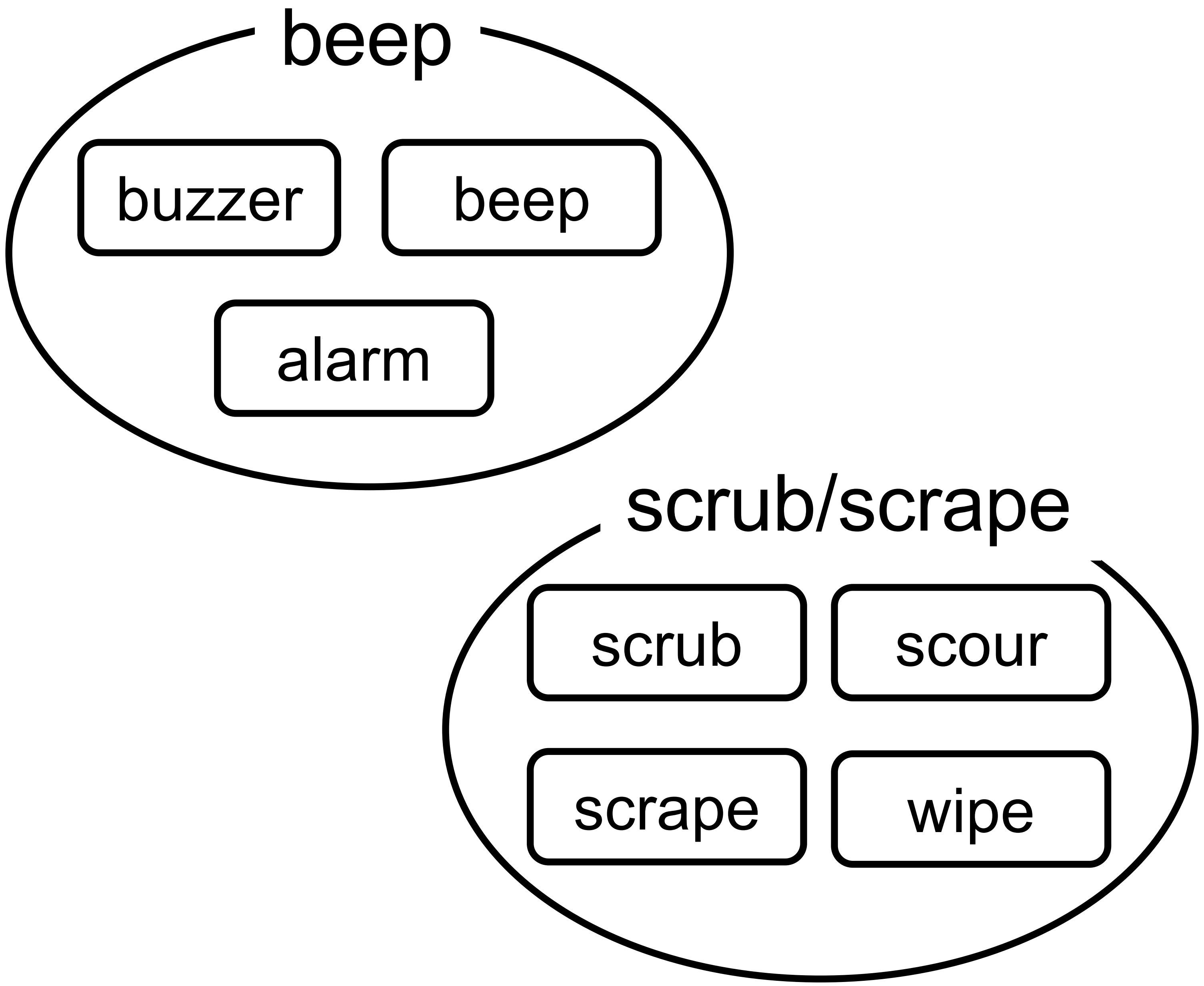
- We annotate all the distinctive sound events which consist of temporal intervals using free-form sound descriptions.
- Using VGG Image annotator tool



Post Processing

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

- From free-form descriptions to categories



Collision Sounds

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

- For collision sounds, we annotate the **materials** of the objects that colliding.
- Materials example

Ceramic

Cloth

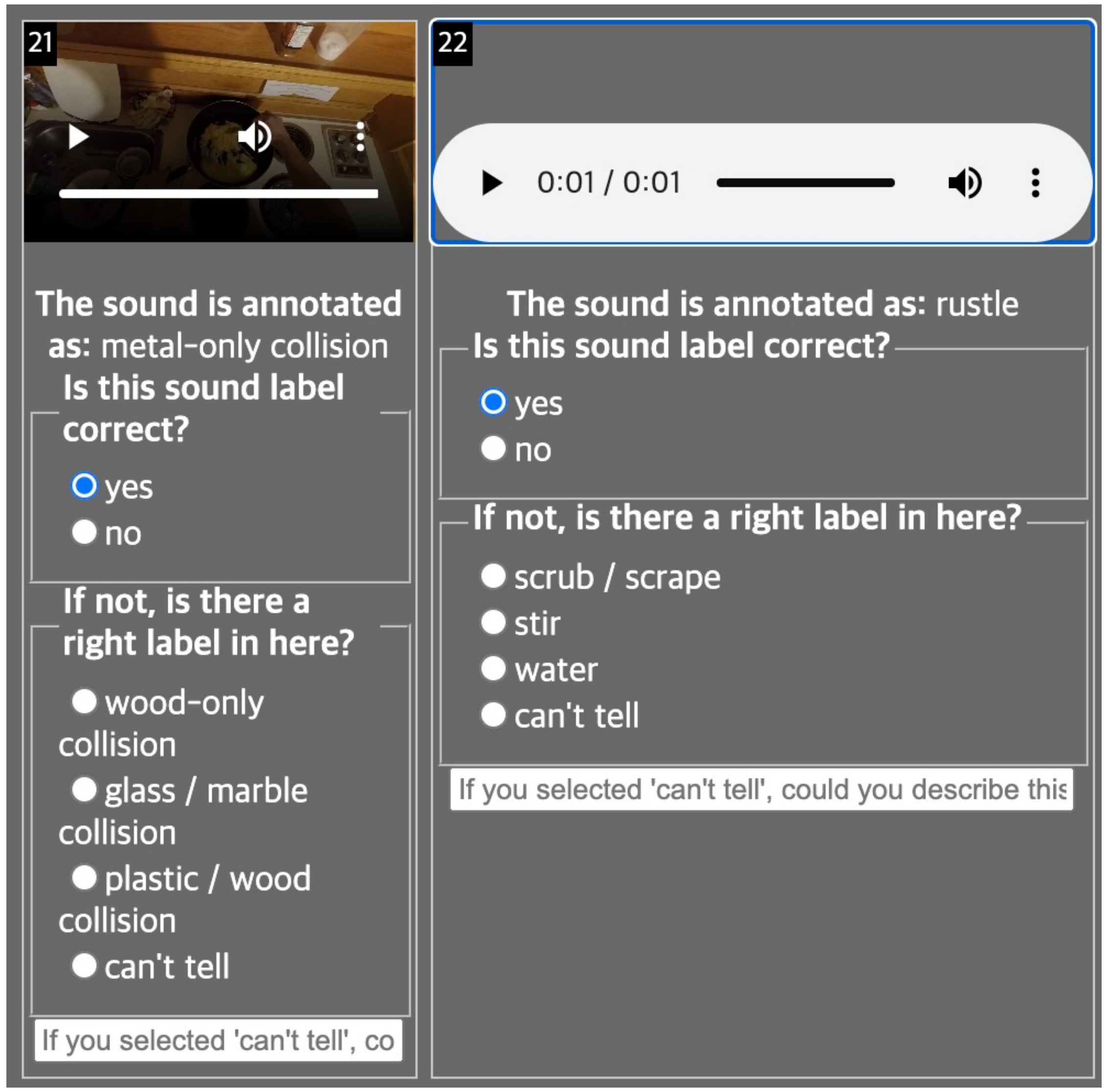
Metal

Plastic

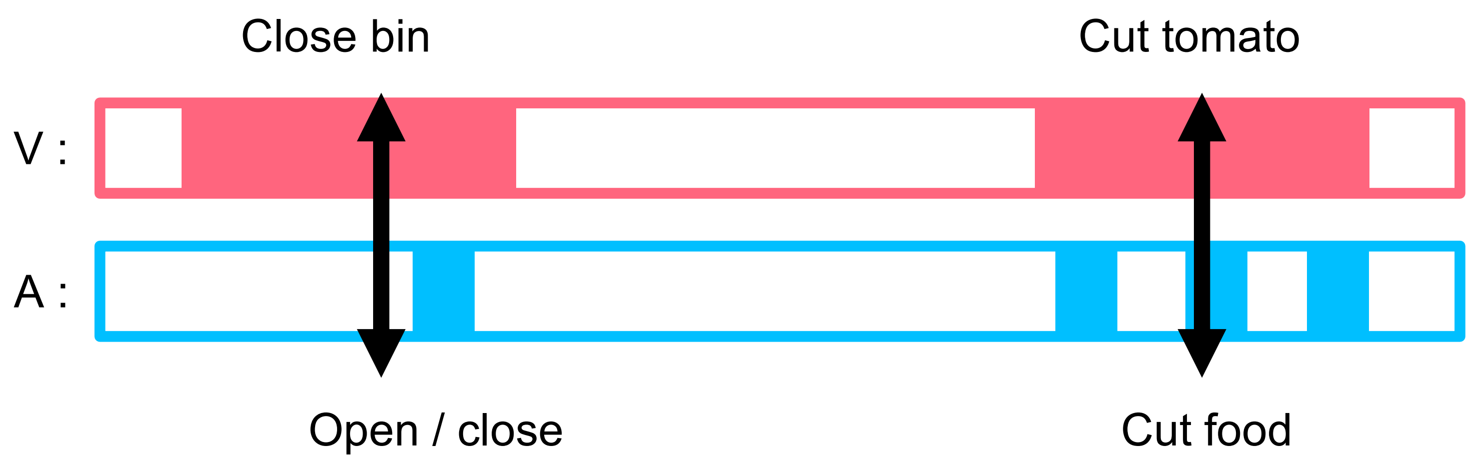
Glass

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

- Manual check on validation / test set



- We use the overlaps between audio and visual segments for reviewing train set.



Search or jump to... / Pull requests Issues Codespaces Marketplace Explore

epic-kitchens / epic-sounds-annotations Public

Edit Pins | Unwatch 5 | Fork 3 | Starred 47

Code | Issues 1 | Pull requests | Actions | Projects | Wiki | Security | Insights | Settings

111 lines (91 sloc) | 10.3 KB

Raw | Blame

EPIC-SOUNDS Dataset

We introduce EPIC-SOUNDS, a large scale dataset of audio annotations capturing temporal extents and class labels within the audio stream of the egocentric videos from EPIC-KITCHENS-100. EPIC-SOUNDS includes 78.4k categorised and 39.2k non-categorised segments of audible events and actions, distributed across 44 classes. In this repository, we provide labelled temporal timestamps for the train / val split, and just the timestamps for the recognition test split. We also provided the temporal timestamps for annotations that could not be clustered into one of our 44 classes, along with the free-form description used during the initial annotation. We train and evaluate two state-of-the-art audio recognition models on our dataset, which we also provide the code and pretrained models for.

Download the Data

A download script is provided for the videos here. You will have to extract the untrimmed audios from these videos. Instructions on how to extract and format the audio into a HDF5 dataset can be found on the Auditory SlowFast GitHub repo. Alternatively, you can email uob-epic-kitchens@bristol.ac.uk for access to an existing HDF5 file.

Contact: uob-epic-kitchens@bristol.ac.uk

Citing

When using the dataset, kindly reference our ICASSP 2023 Paper:

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

Thur (Session 4)
Poster # 344

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

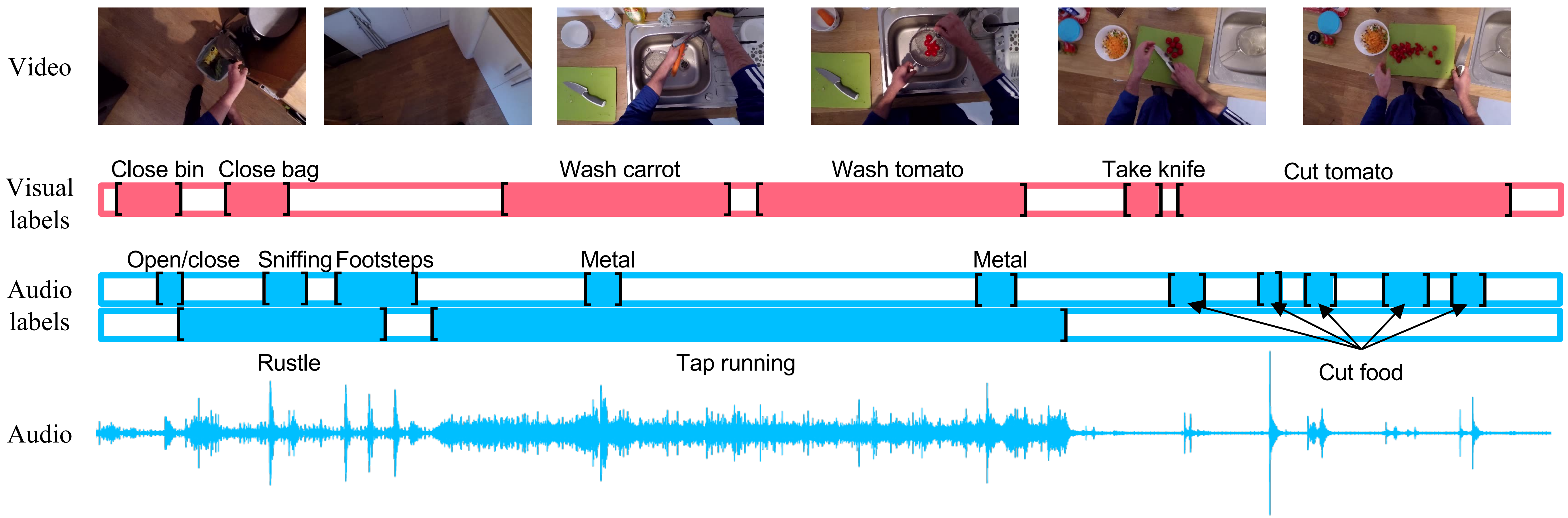
University of BRISTOL

VGG
UNIVERSITY OF OXFORD

CZECH TECHNICAL UNIVERSITY IN PRAGUE

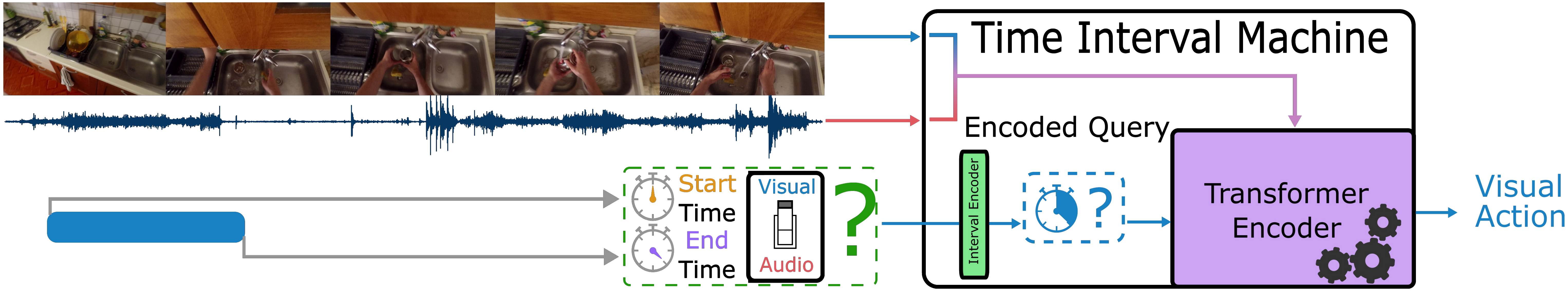
Multi-Modal Long-Form Dataset

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



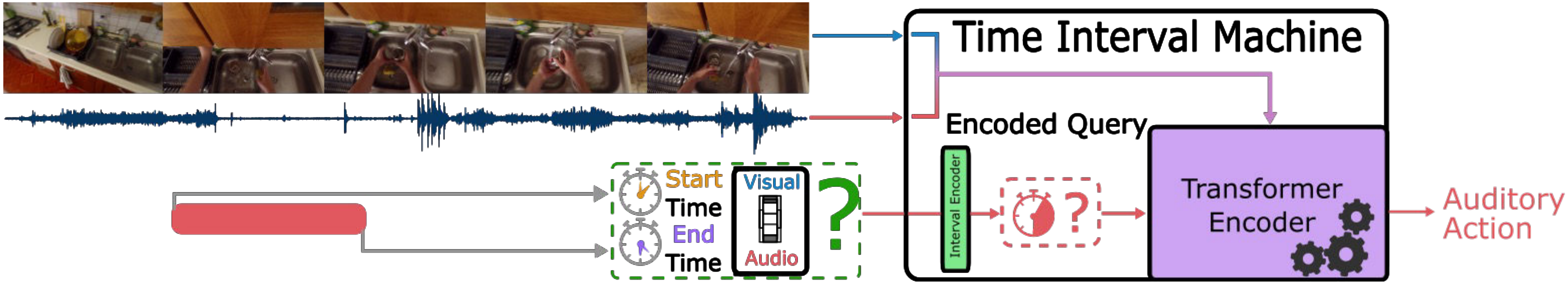
TIM: A Time-Interval Audio-Visual Machine

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



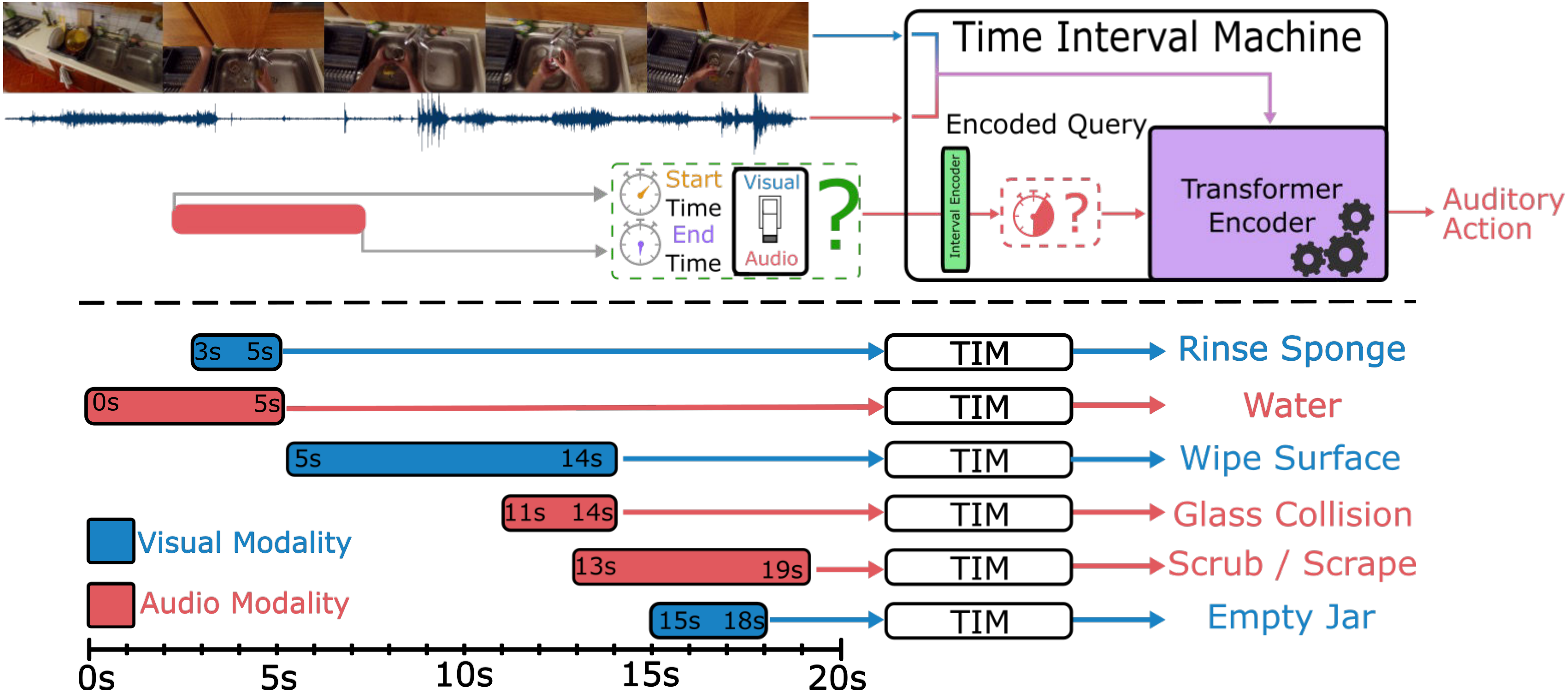
TIM: A Time-Interval Audio-Visual Machine

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



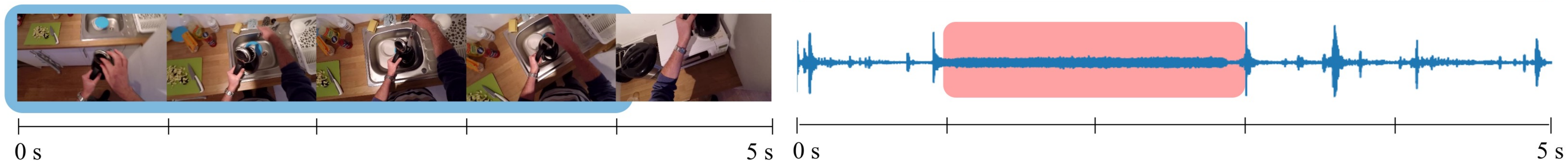
TIM: A Time-Interval Audio-Visual Machine

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

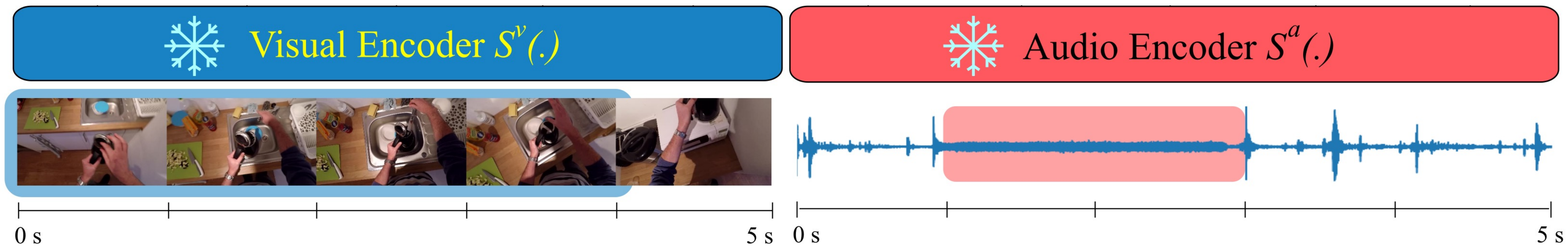


University of BRISTOL

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

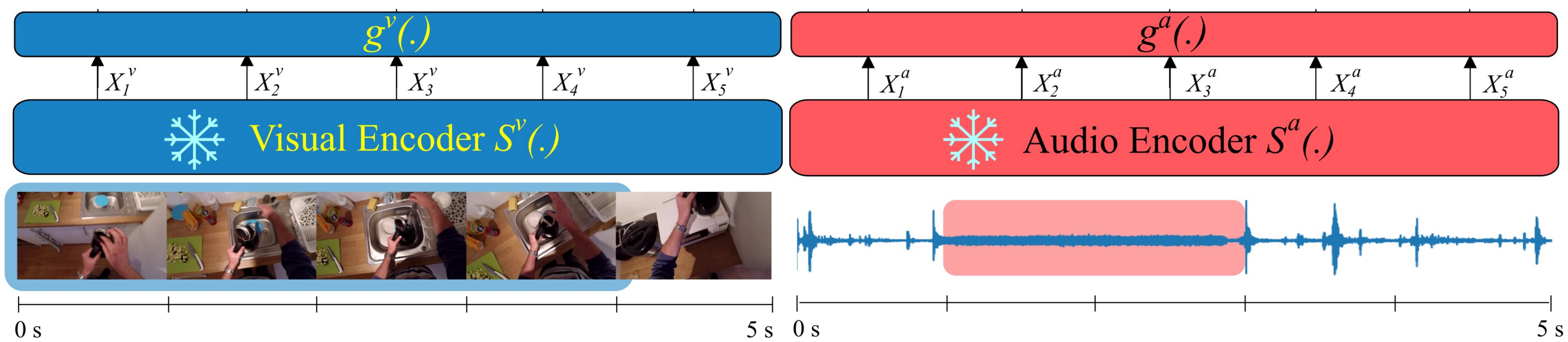
TIM: A Time-Interval Audio-Visual Machine

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



TIM: A Time-Interval Audio-Visual Machine

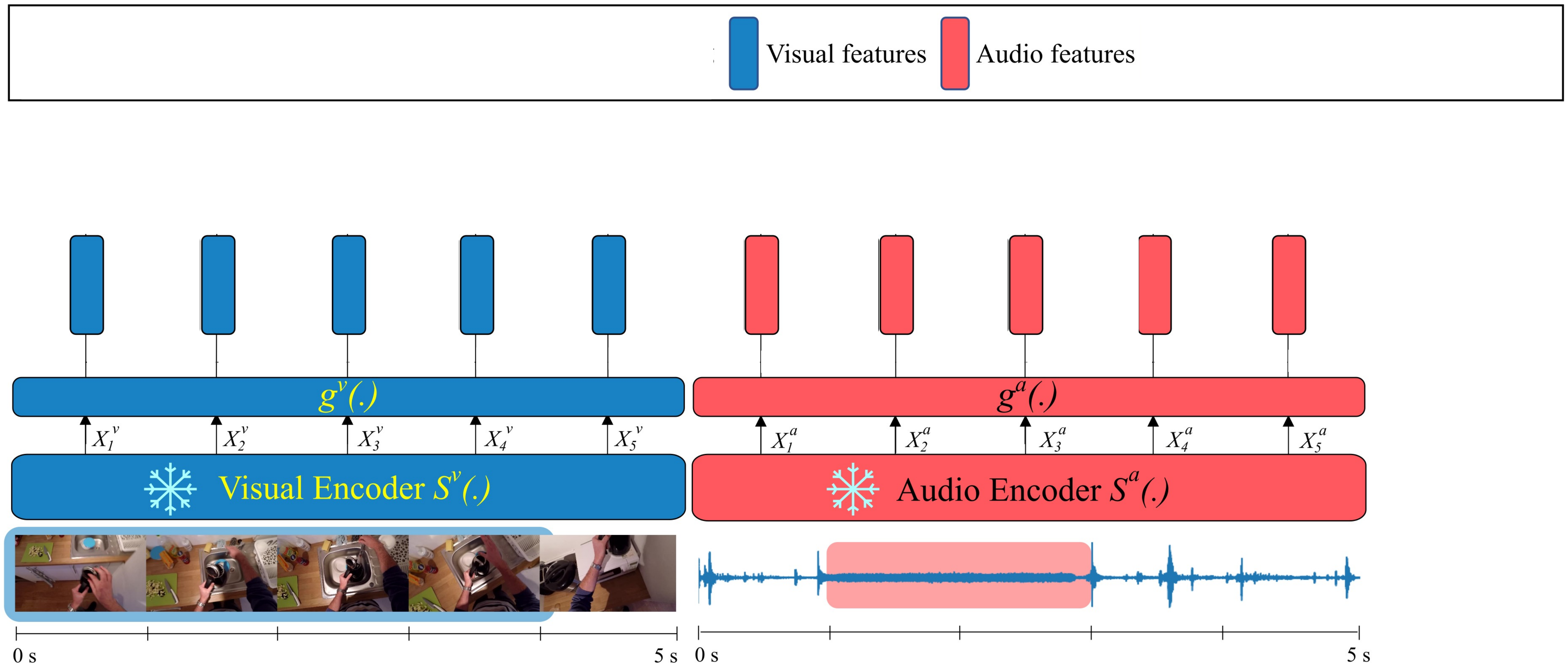
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



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

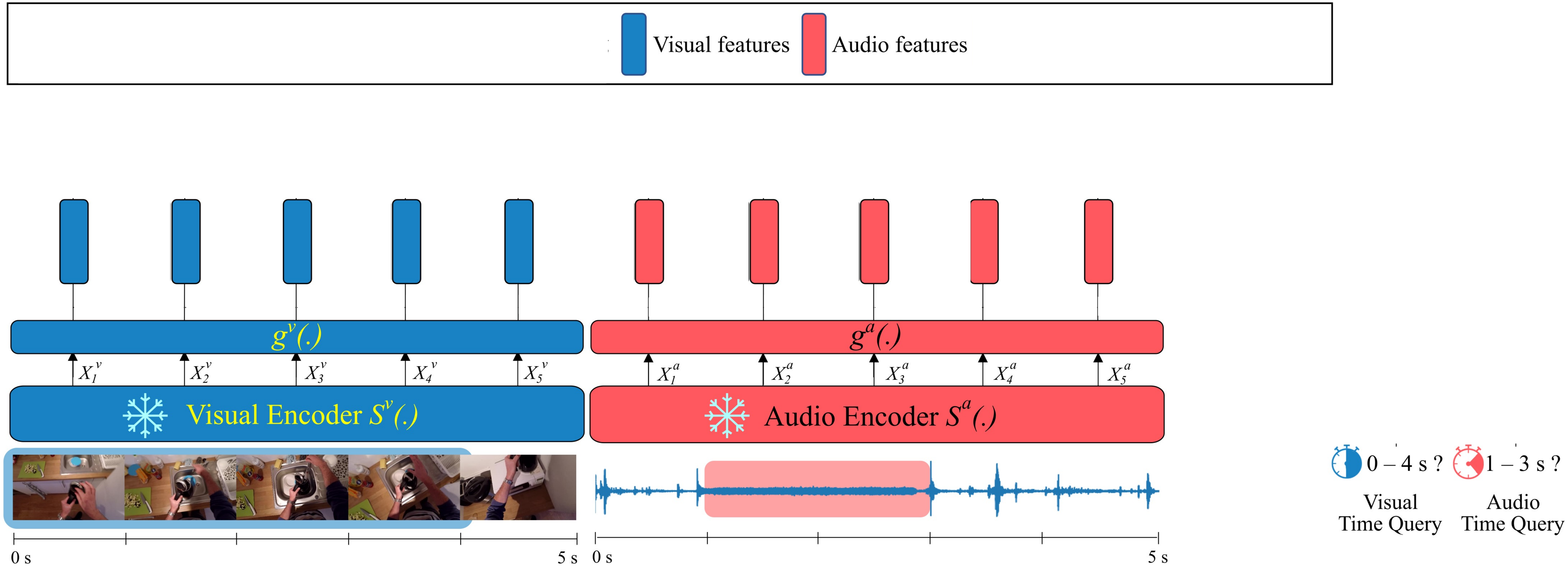
TIM: A Time-Interval Audio-Visual Machine

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



TIM: A Time-Interval Audio-Visual Machine

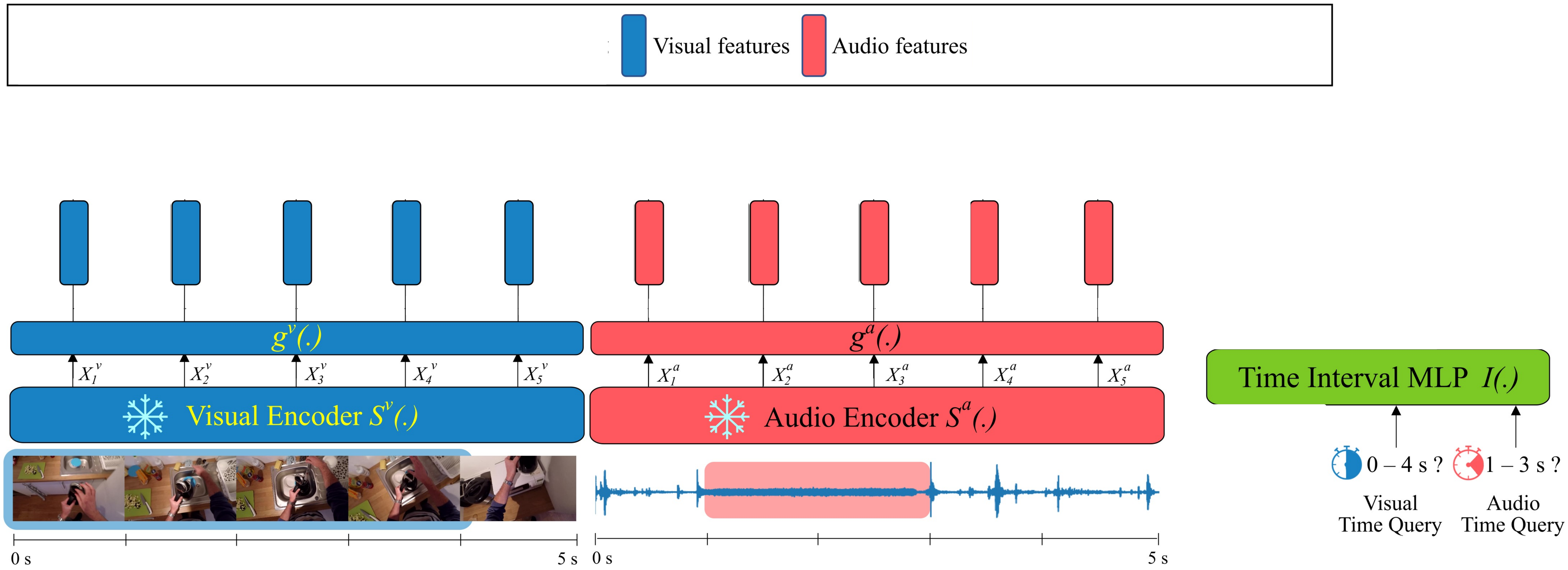
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



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

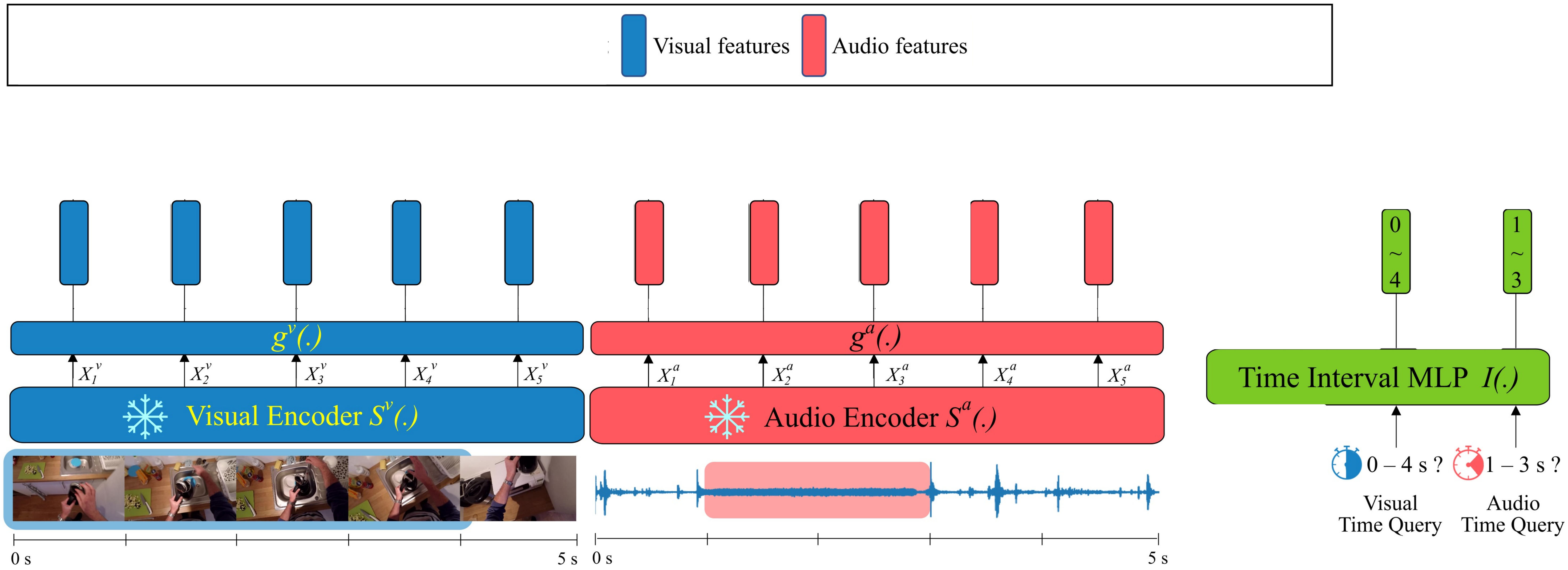
TIM: A Time-Interval Audio-Visual Machine

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



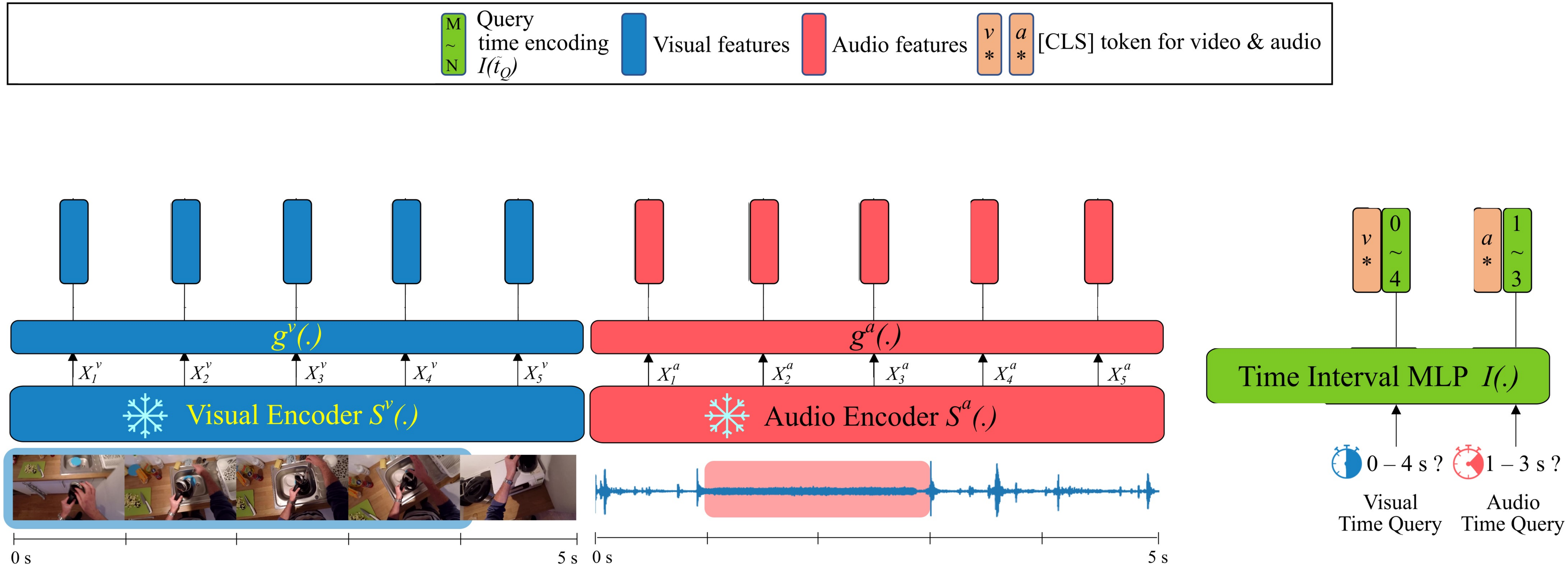
TIM: A Time-Interval Audio-Visual Machine

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



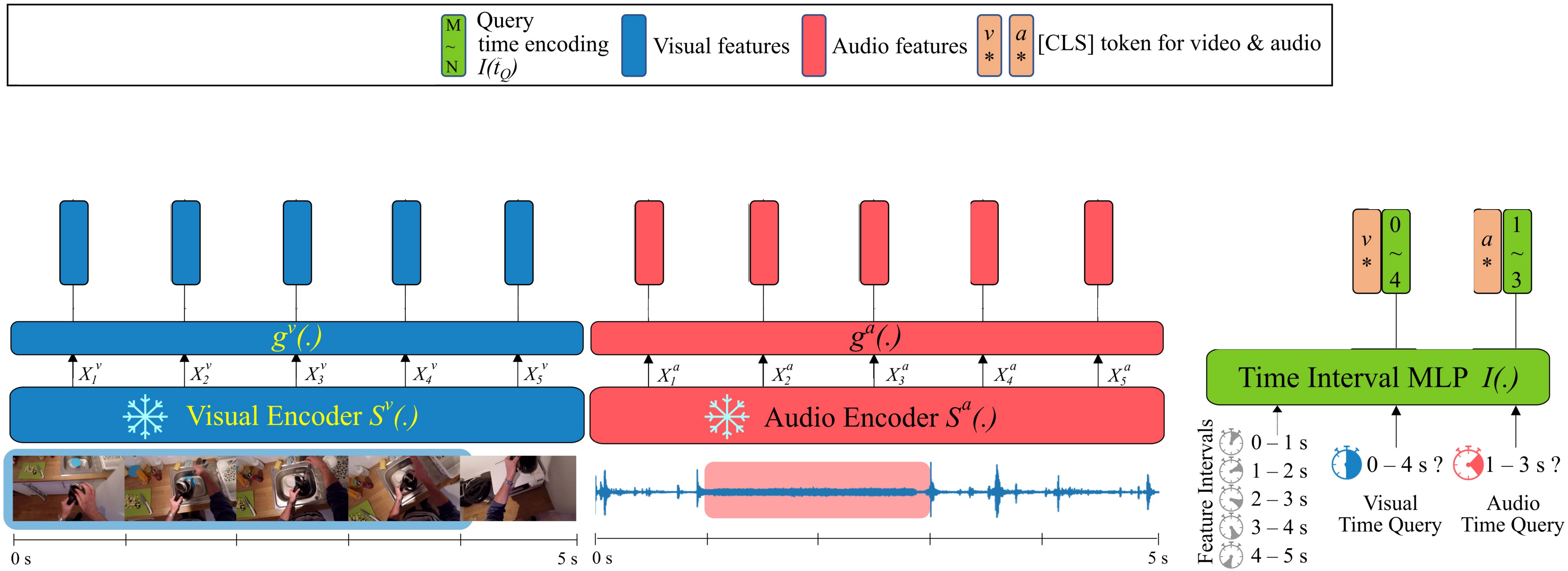
TIM: A Time-Interval Audio-Visual Machine

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

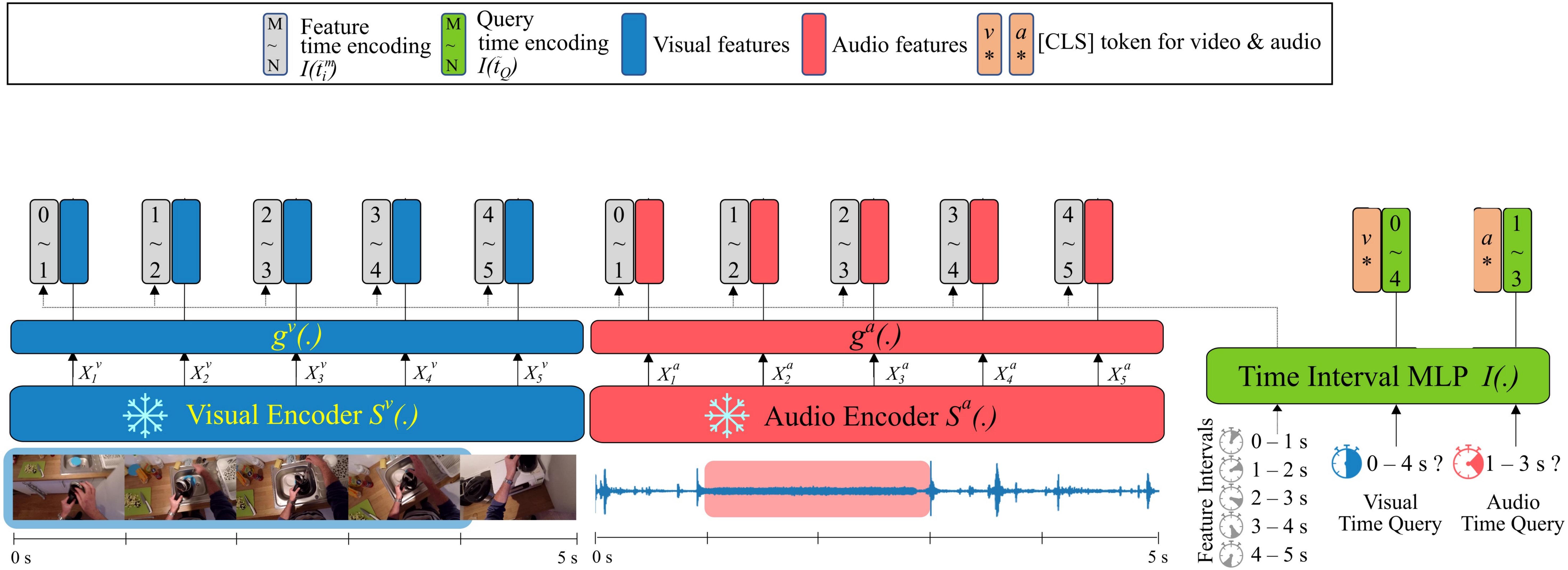


University of BRISTOL

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



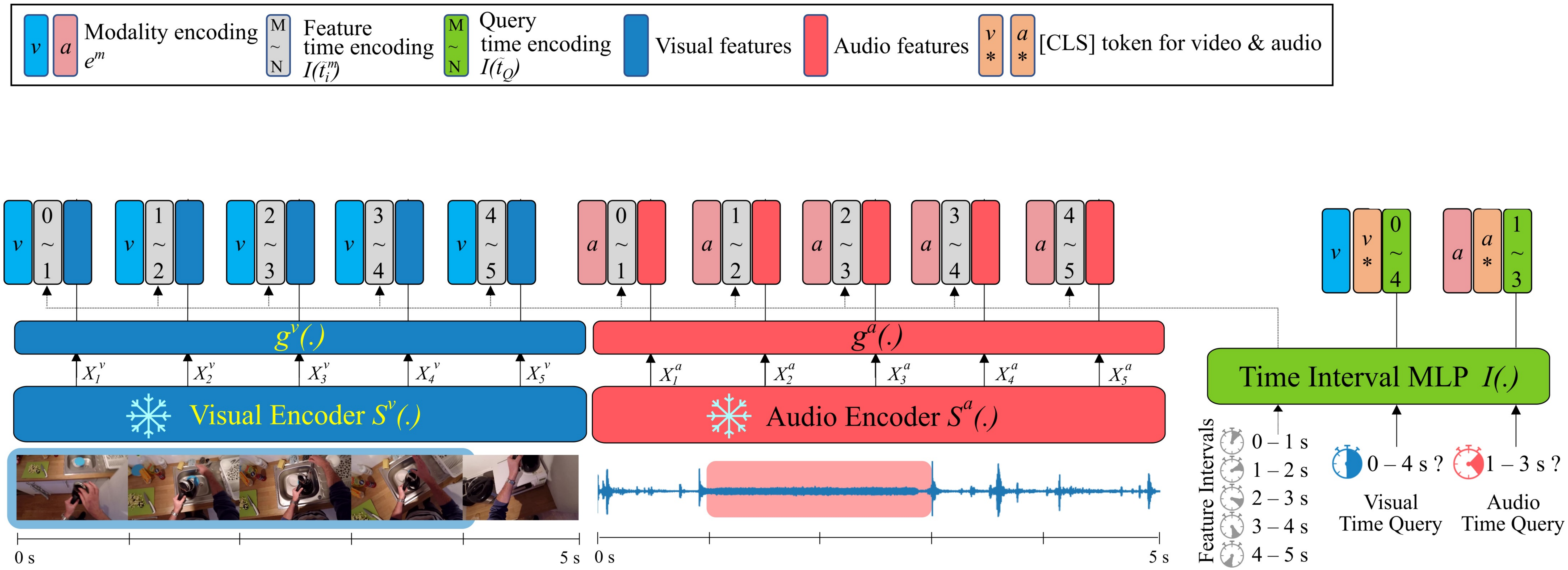
University of BRISTOL

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

Dima Damen
LOVEU @CVPR2024

TIM: A Time-Interval Audio-Visual Machine

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

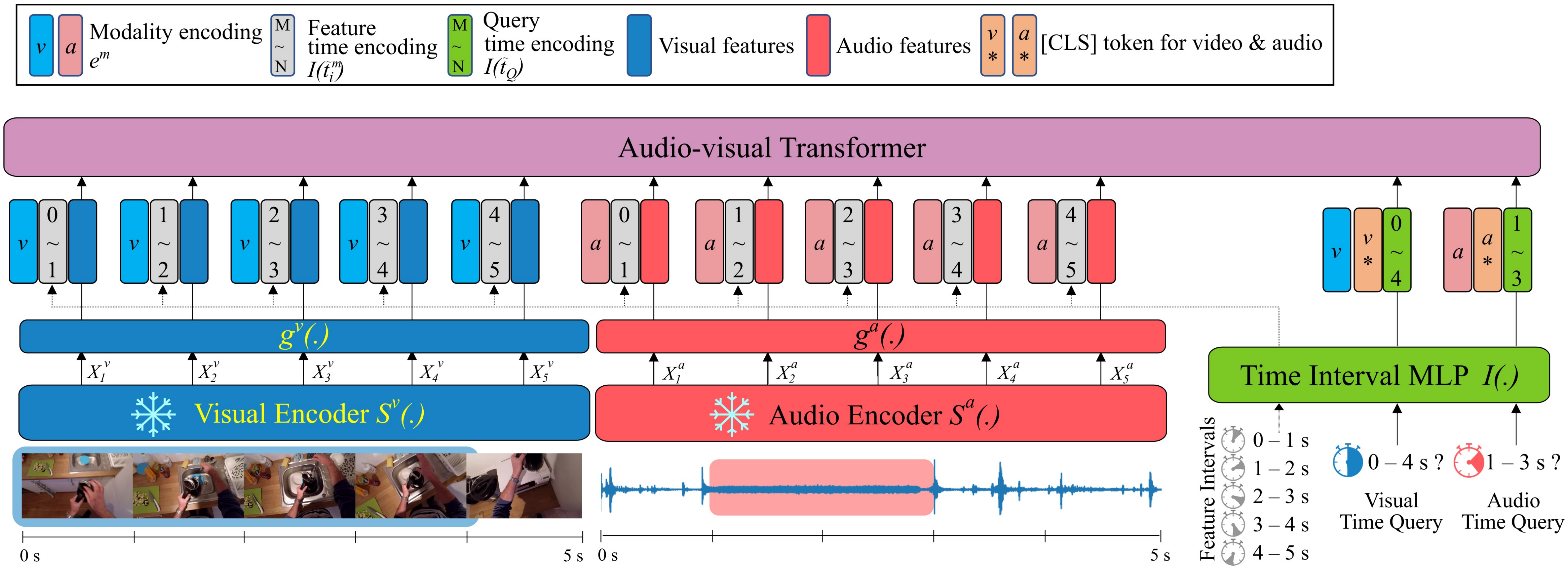


University of BRISTOL

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

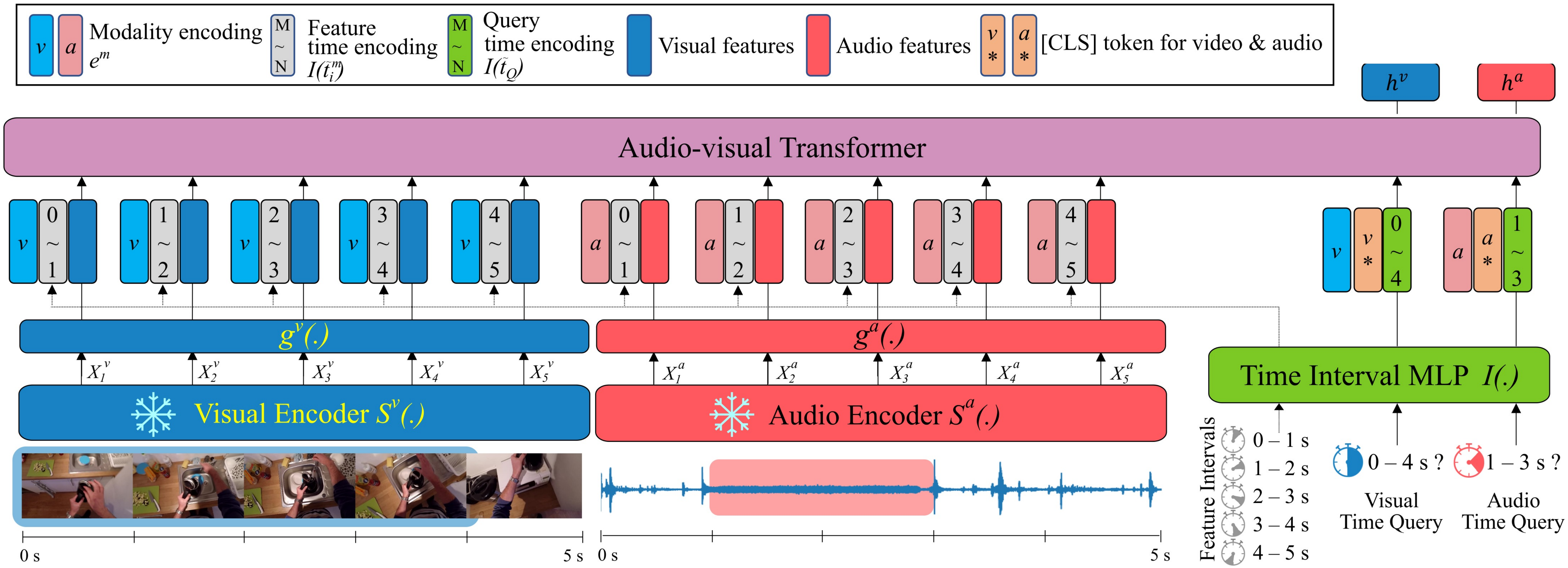


University of BRISTOL

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

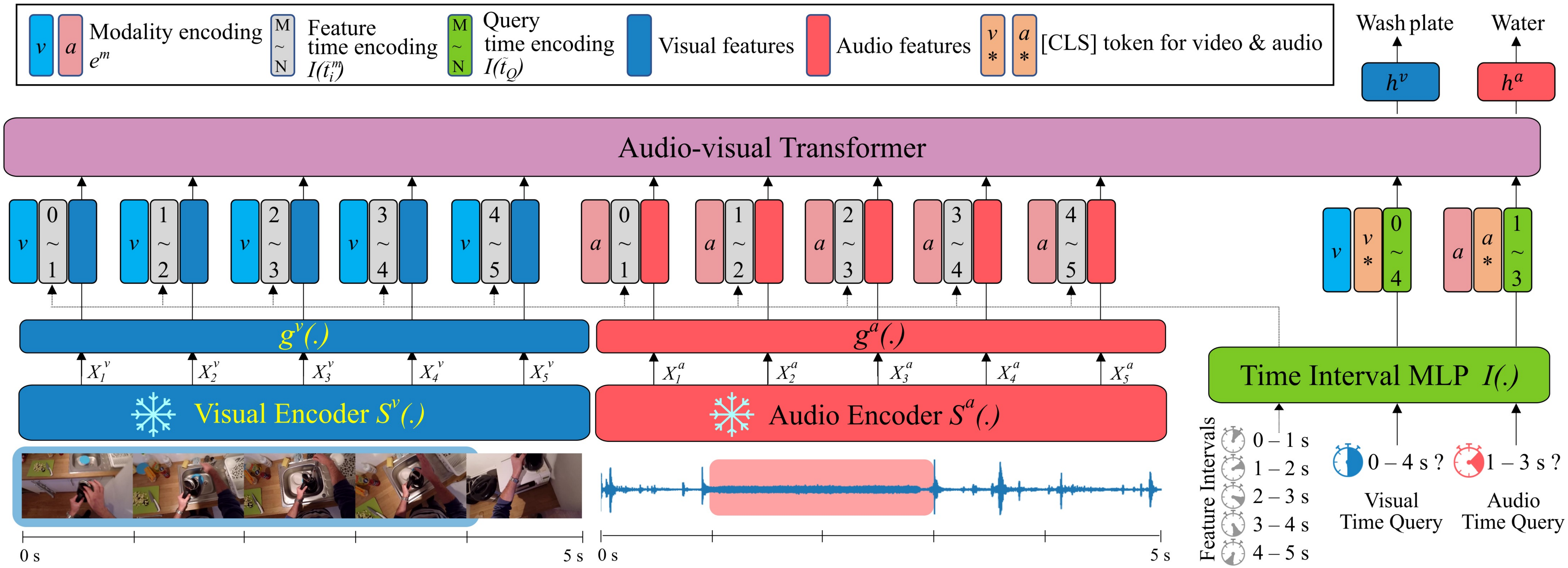
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



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



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

Dima Damen
LOVEU @CVPR2024

TIM: A Time-Interval Audio-Visual Machine

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

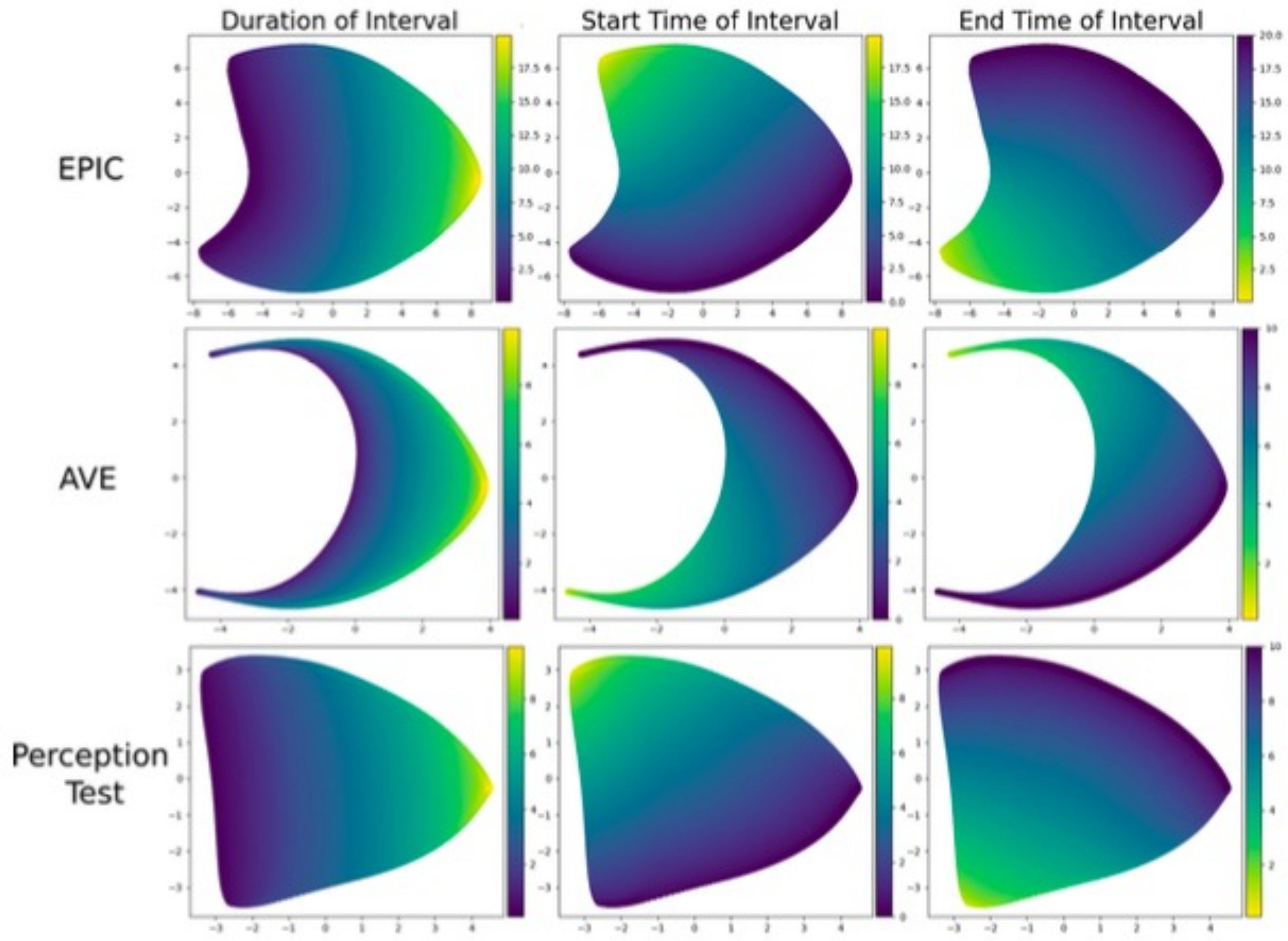
Model	*xp*	LLM	Verb	Noun	Action
Visual-only models					
MFormer-HR [37]	336p	✗	67.0	58.5	44.5
MoViNet-A6 [27]	320p	✗	72.2	57.3	47.7
MeMViT [55]	224p	✗	71.4	60.3	48.4
Omnivore [14]	224p	✗	69.5	61.7	49.9
MTV [59]	280p	✗	69.9	63.9	50.5
LaViLa (TSF-L) [63]	224p	✓	72.0	62.9	51.0
AVION (ViT-L) [62]	224p	✓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

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

Thur (Session 4)
Poster # 344

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

Long-Form Egocentric Video Understanding

- No Semantic Supervision
- No Shots - Temporal Alignment
- Audio-Visual Semantic Gap
- Quick View Changes
- Repeating Actions
- Long Continuous Streams

EPIC-KITCHENS

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
1859

Berkeley
UNIVERSITY OF CALIFORNIA

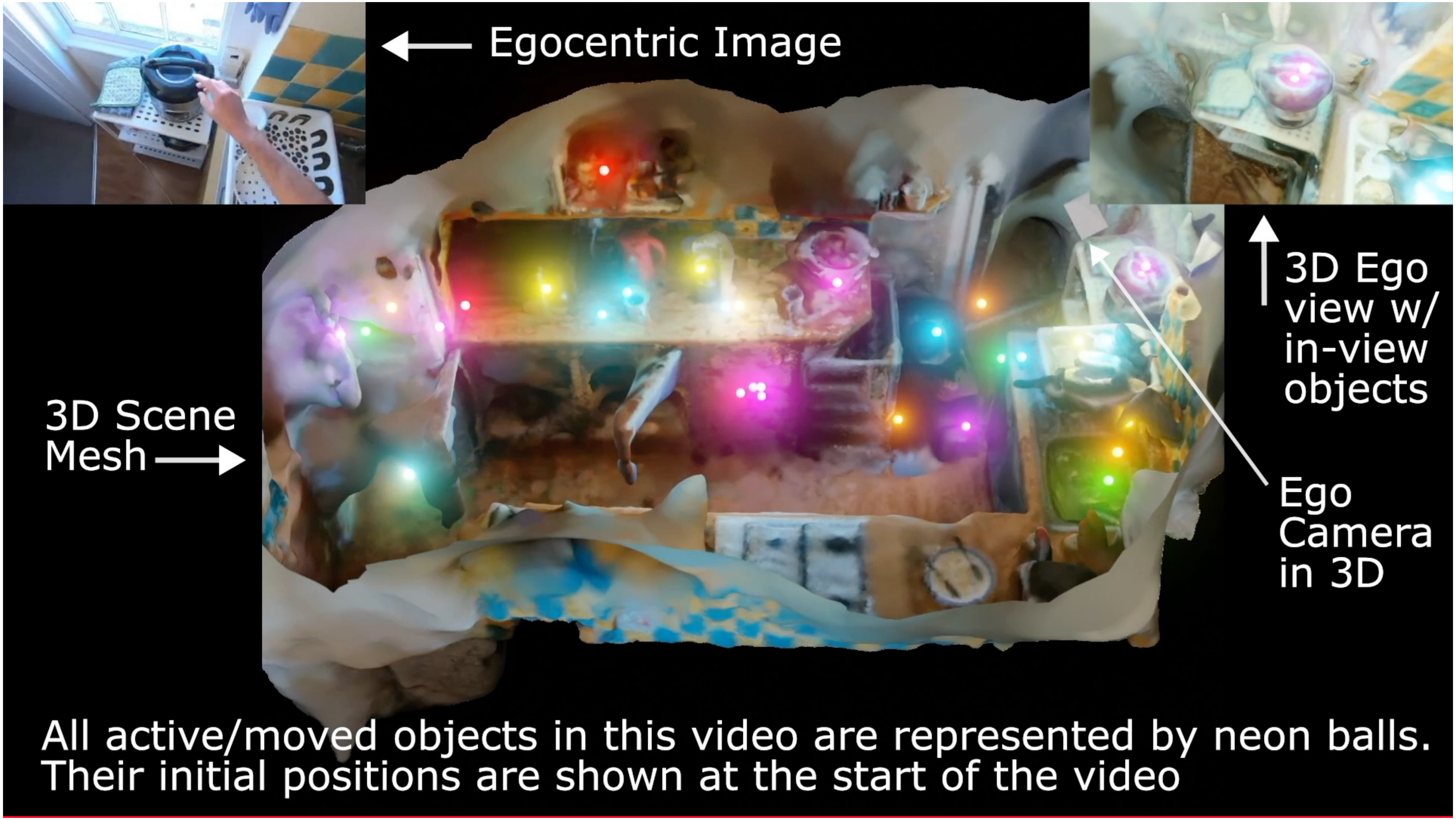
University of BRISTOL

Out of Sight, Not Out of Mind

with: Chiara Plizzari, Toby Perrett, Shubham Goel, Angjoo Kanazawa

University of BRISTOL

Plizzari et al (2024). Spatial Cognition from Egocentric Video: Out of Sight, Not Out of Mind. ArXiv

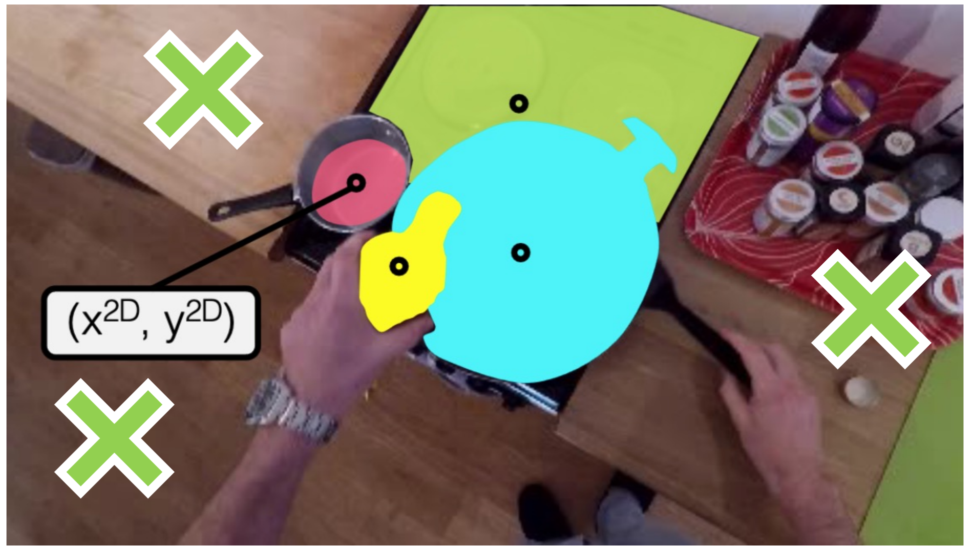


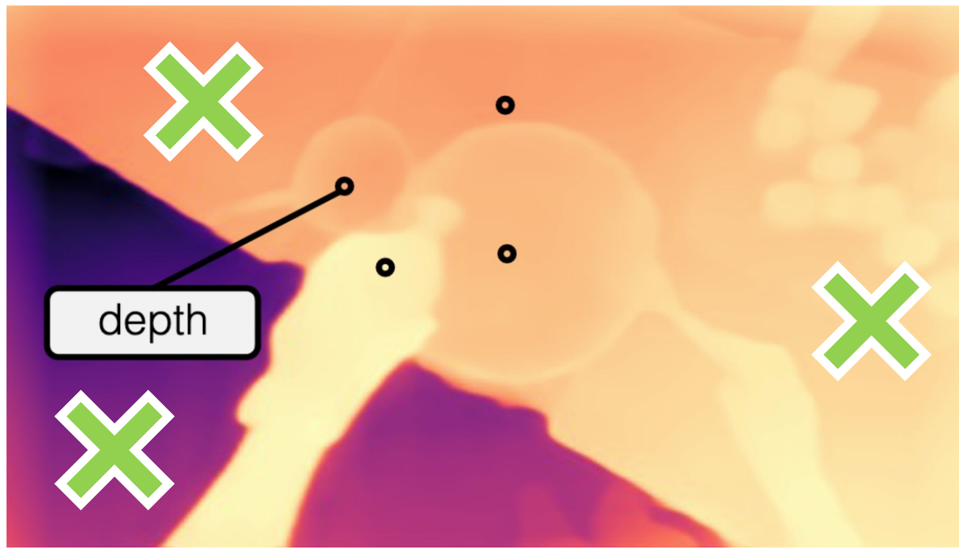
Egocentric Image
3D Ego view w/ in-view objects
3D Scene Mesh
Ego Camera in 3D
All active/moved objects in this video are represented by neon balls.
Their initial positions are shown at the start of the video

Out of Sight, not Out of Mind

with: Chiara Plizzari
Toby Perrett
Shubham Goel
Angjoo Kanazawa

Lift → Match → Keep



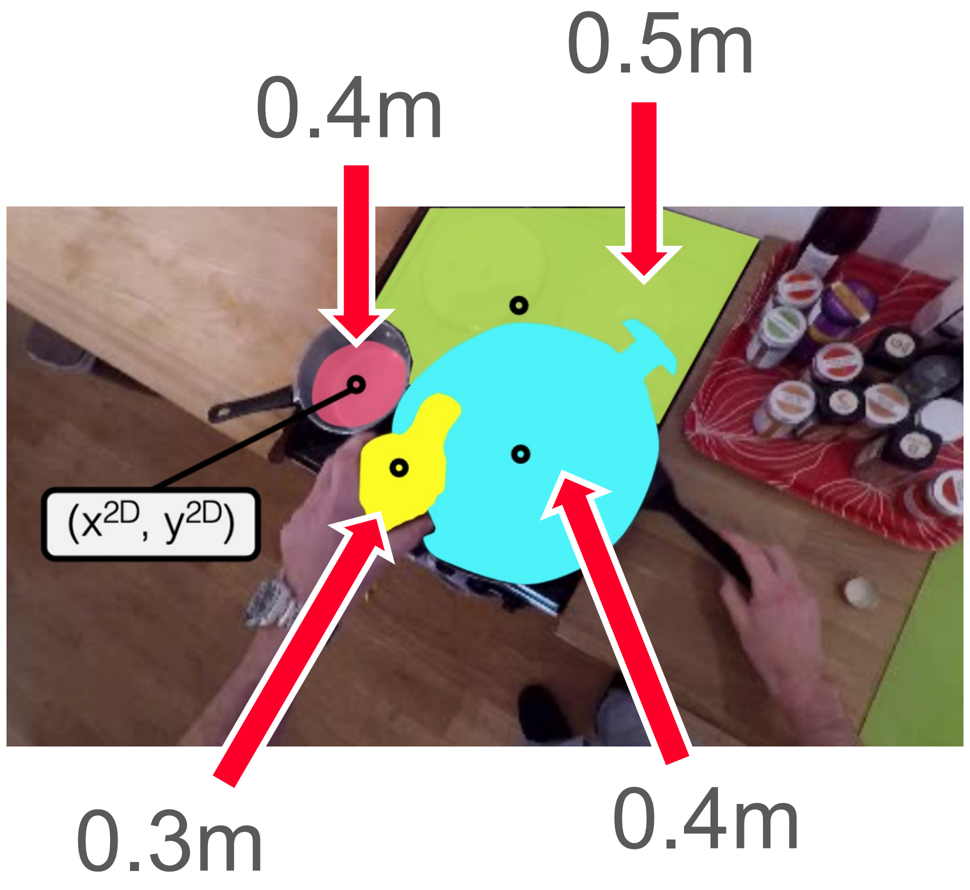


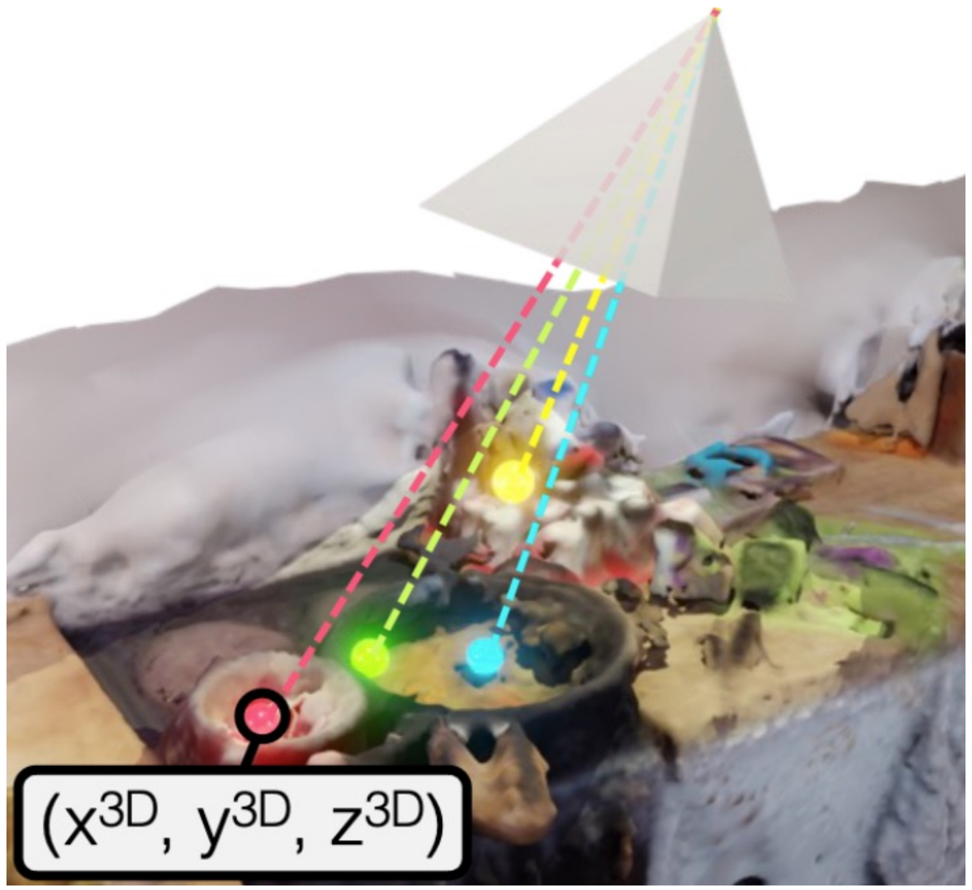
0.0 … 1.0

0.3m … 1.8m

with: Chiara Plizzari
Toby Perrett
Shubham Goel
Angjoo Kanazawa

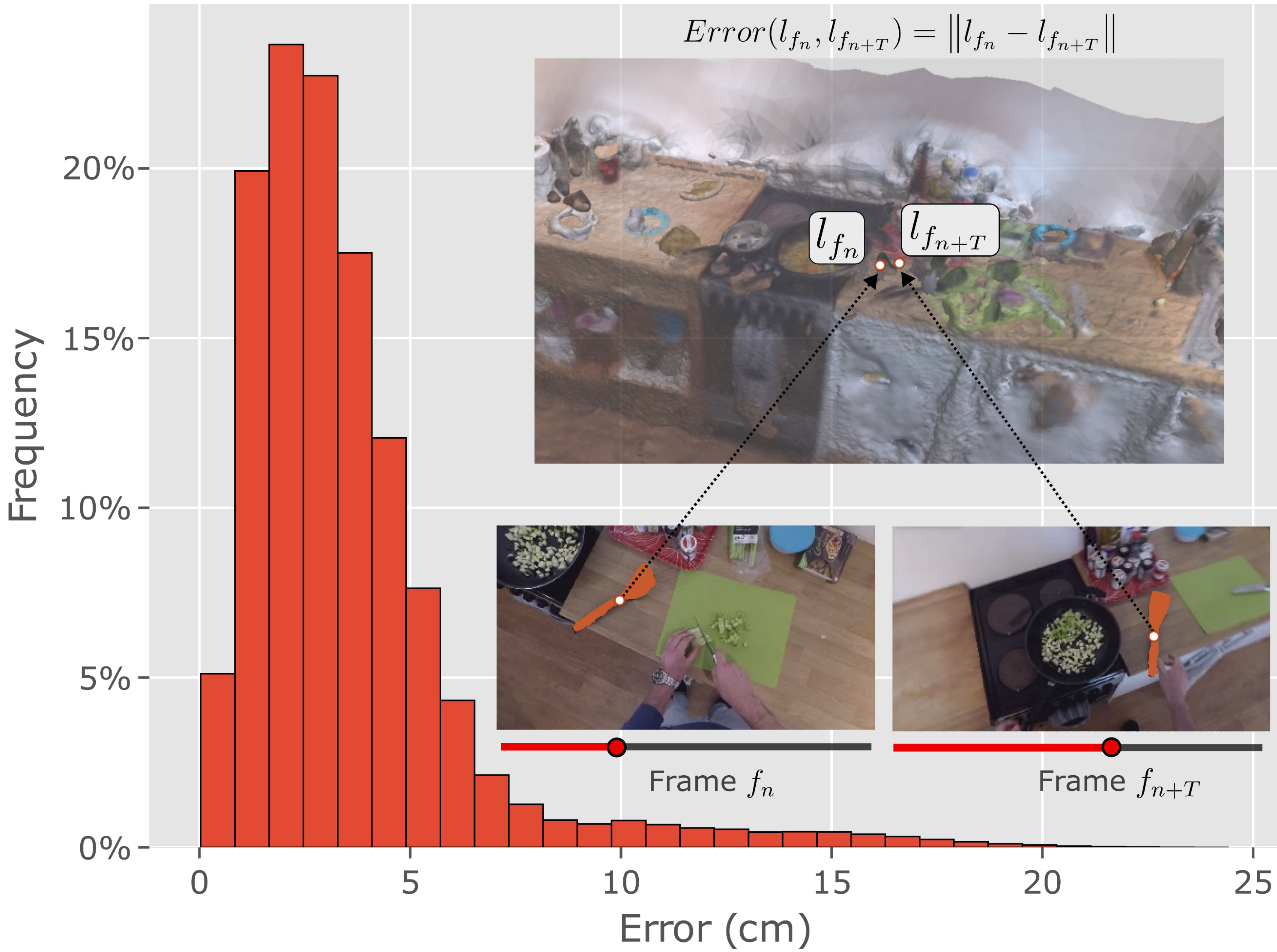
Lift | Match | Keep





Out of Sight, not Out of Mind

with: Chiara Plizzari Shubham Goel
Toby Perrett Angjoo Kanazawa



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, 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)

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)

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)

University of BRISTOL

Plizzari et al (2024). Spatial Cognition from Egocentric Video: Out of Sight, Not Out of Mind. ArXiv

Long-Form Egocentric Video Understanding

- No Semantic Supervision
- No Shots - Temporal Alignment
- Audio-Visual Semantic Gap
- Quick View Changes
- Repeating Actions
- Long Continuous Streams

Every Shot Counts:

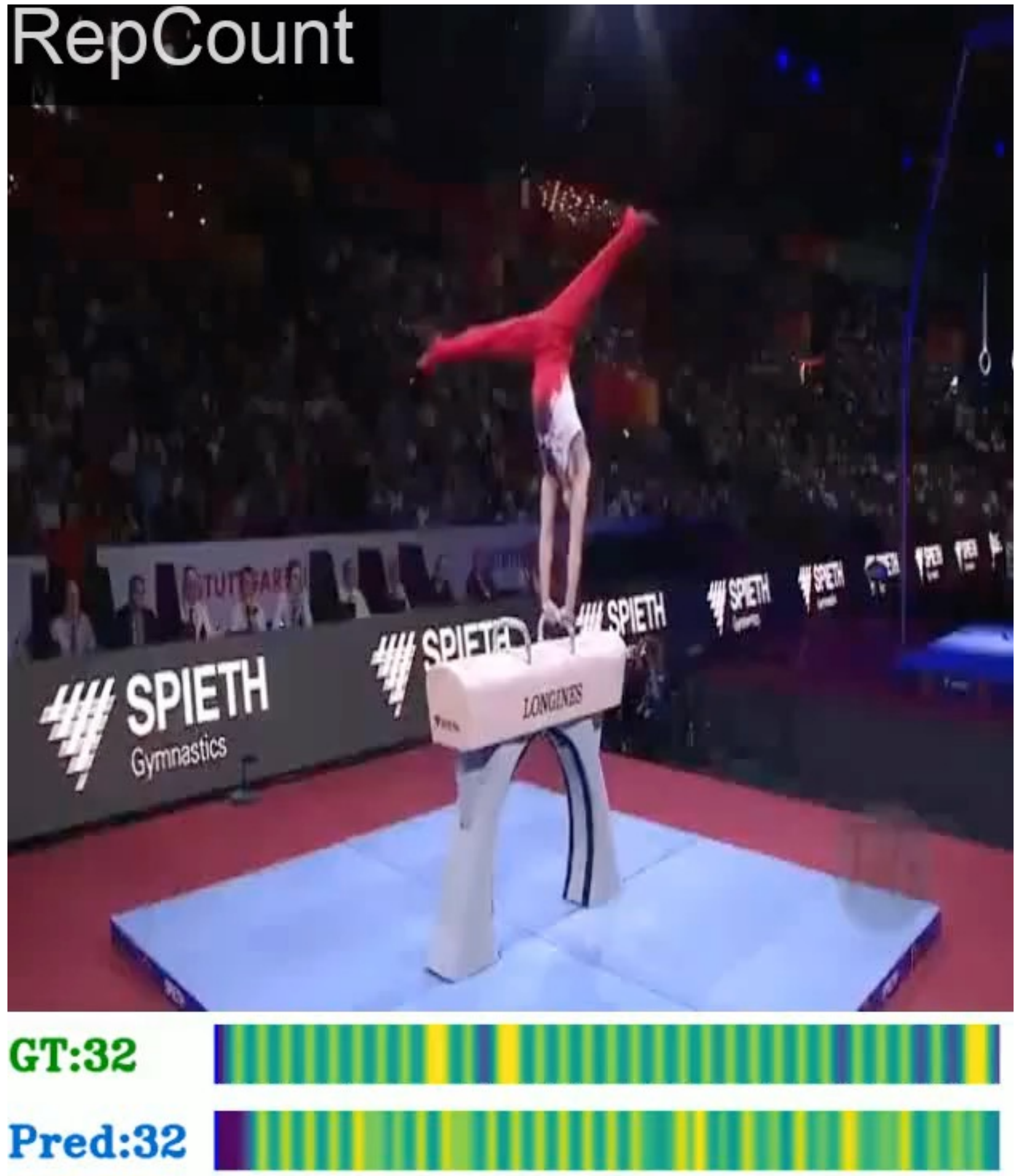
Using Exemplars for Repetition Counting in Videos

Saptarshi Sinha, Alexandros Stergiou, Dima Damen

University of BRISTOL

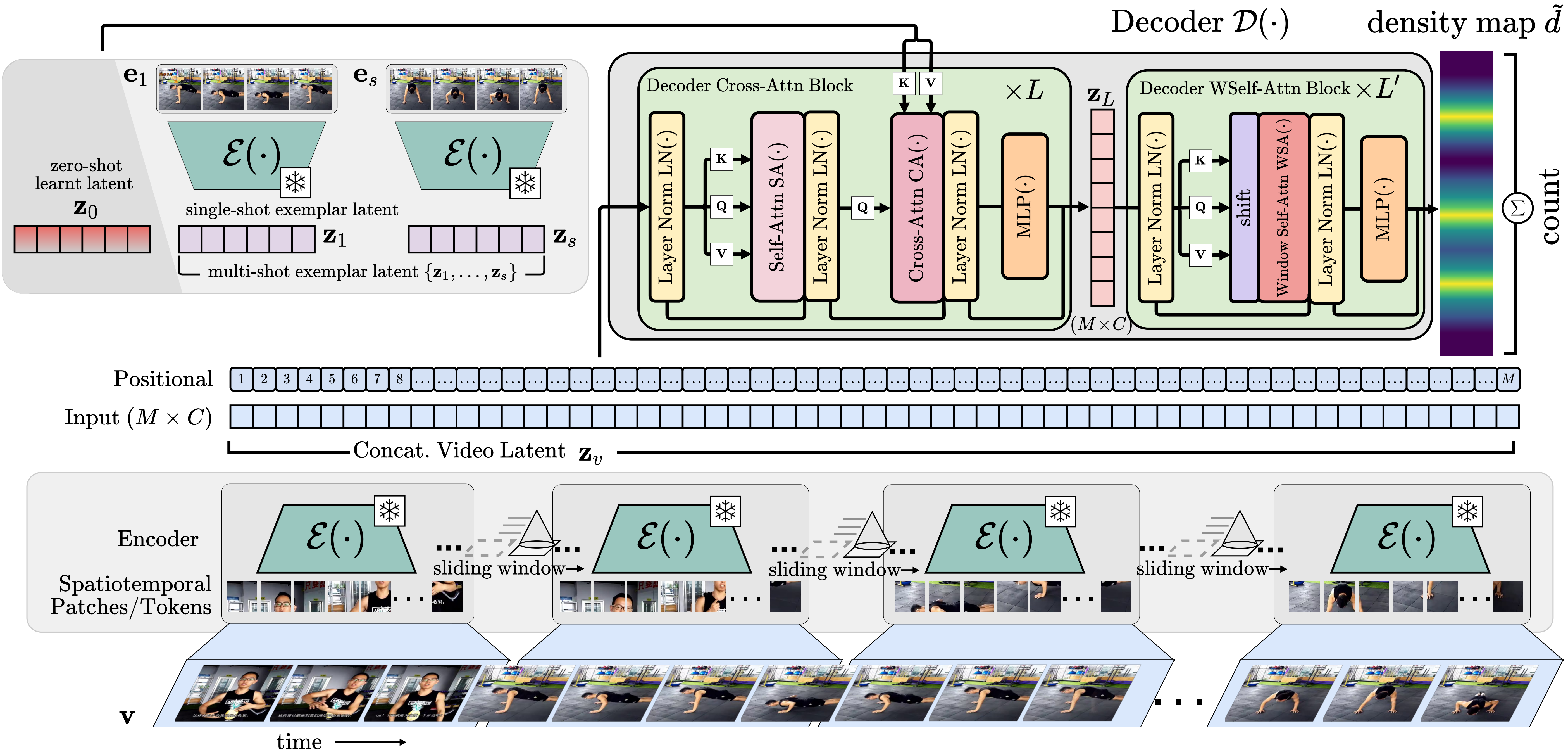
Every Shot Counts

with: Saptarshi Sinha
Alexandros Stergiou



Every Shot Counts

with: Saptarshi Sinha
Alexandros Stergiou



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

Every Shot Counts

with: Saptarshi Sinha
Alexandros Stergiou

(a) RepCount

Method	Encoder	RMSE↓	MAE↓	OBZ↑	OBO↑
RepNet [15]	R2D50	-	0.995	-	0.013
TransRAC [18]	VSwinT	9.130*	0.443	0.085*	0.291
MFL [27]†	VSwinT	-	0.384	-	0.386
ESCounts	VSwinT	6.905	0.298	0.183	0.403
ESCounts	VMAE	**4.455**	**0.213**	**0.245**	**0.563**

(c) UCFRep

Method	Encoder	RMSE↓	MAE↓	OBZ↑	OBO↑
Levy & Wolf [25]	RX3D101	-	0.286	-	0.680
RepNet [15]	R2D50	-	0.998	-	0.009
Context (F) [62]	RX3D101	5.761*	0.653*	0.143*	0.372*
TransRAC [18]	VSwinT	-	0.640	-	0.324
MFL [27]†	RX3D101	-	0.388	-	0.510
ESCounts	RX3D101	2.004	0.247	0.343	0.731
ESCounts	VMAE	**1.972**	0.216	0.381	0.704

(b) Countix

Method	Encoder	RMSE↓	MAE↓	OBZ↑	OBO↑
RepNet [15]	R2D50	-	0.364	-	0.697
Sight & Sound [64]†	R(2+1)D18	-	0.307	-	0.511
ESCounts	R(2+1)D18	3.536	0.293	0.286	**0.701**
ESCounts	VMAE	**3.029**	**0.276**	**0.319**	0.673

with: Saptarshi Sinha
Alexandros Stergiou

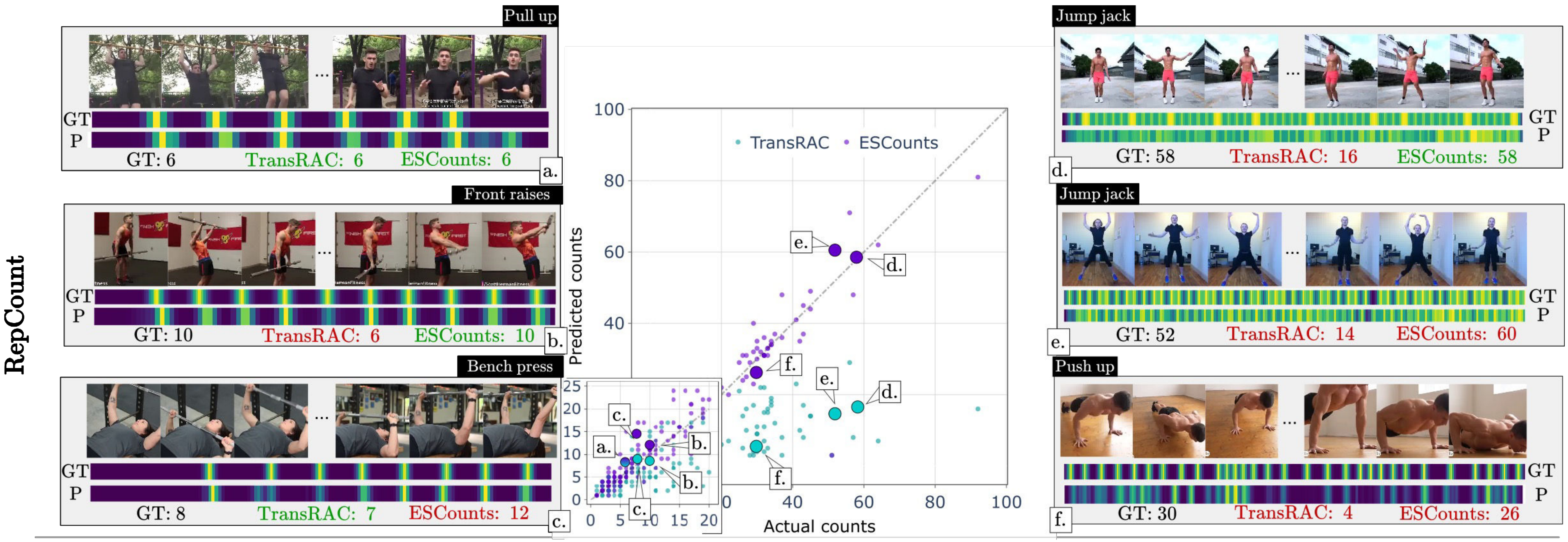
Table 4: Number of shots at inference. We test using exemplars from the same video or a different video of the same action class from the train set.

Shots	Same video	RepCount				UCFRep			
		RMSE↓	MAE↓	OBZ↑	OBO↑	RMSE↓	MAE↓	OBZ↑	OBO↑
0	N/A	4.455	0.213	0.245	0.563	1.972	0.216	0.381	0.704
1	✗	4.432	0.207	0.251	0.563	1.912	0.211	0.388	0.712
	✔	4.369	0.210	0.247	0.589	1.890	0.203	0.400	0.714
2	✗	4.384	**0.206**	0.251	0.572	1.885	0.208	0.391	0.720
	✔	4.360	0.209	0.247	0.592	1.857	0.199	0.419	0.718
3	✗	4.381	0.207	**0.252**	0.579	1.878	0.207	0.399	**0.730**
	✔	**4.351**	**0.206**	0.250	**0.596**	**1.855**	**0.198**	**0.420**	0.723

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

Every Shot Counts

with: Saptarshi Sinha
Alexandros Stergiou



University of BRISTOL

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

Dima Damen
LOVEU @CVPR2024

Every Shot Counts - Generalisation

with: Saptarshi Sinha
Alexandros Stergiou

Table 2: Cross-dataset generalisation scores. Arrows $X \rightarrow Y$ denote train dataset X and test dataset Y. Results obtained using provided checkpoints are denoted with *.

	RepCount → UCFRep				RepCount → Countix			
	RMSE↓	MAE↓	OBZ↑	OBO↑	RMSE↓	MAE↓	OBZ↑	OBO↑
RN [16]	-	0.998	-	0.009	-	-	-	-
TRAC [20]	6.701*	0.640	0.087*	0.324	6.867*	0.593*	0.132*	0.364*
MFL [30]	-	0.523	-	0.350	-	-	-	-
ESCounts	**3.536**	**0.317**	**0.219**	**0.571**	**4.429**	**0.374**	**0.185**	**0.521**

Table X2. **Close and open-set setting results on RepCount.**

Task	Method	benchmark		open-set	
		MAE↓	OBO↑	MAE↓	OBO↑
TAL	Huang *et al.*	0.527	0.159	1.000	0.000
VRC	TRAC	0.443	0.291	0.625	0.204
	ESCounts	**0.213**	**0.563**	**0.436**	**0.519**

Every Shot Counts - Generalisation

with: Saptarshi Sinha
Alexandros Stergiou

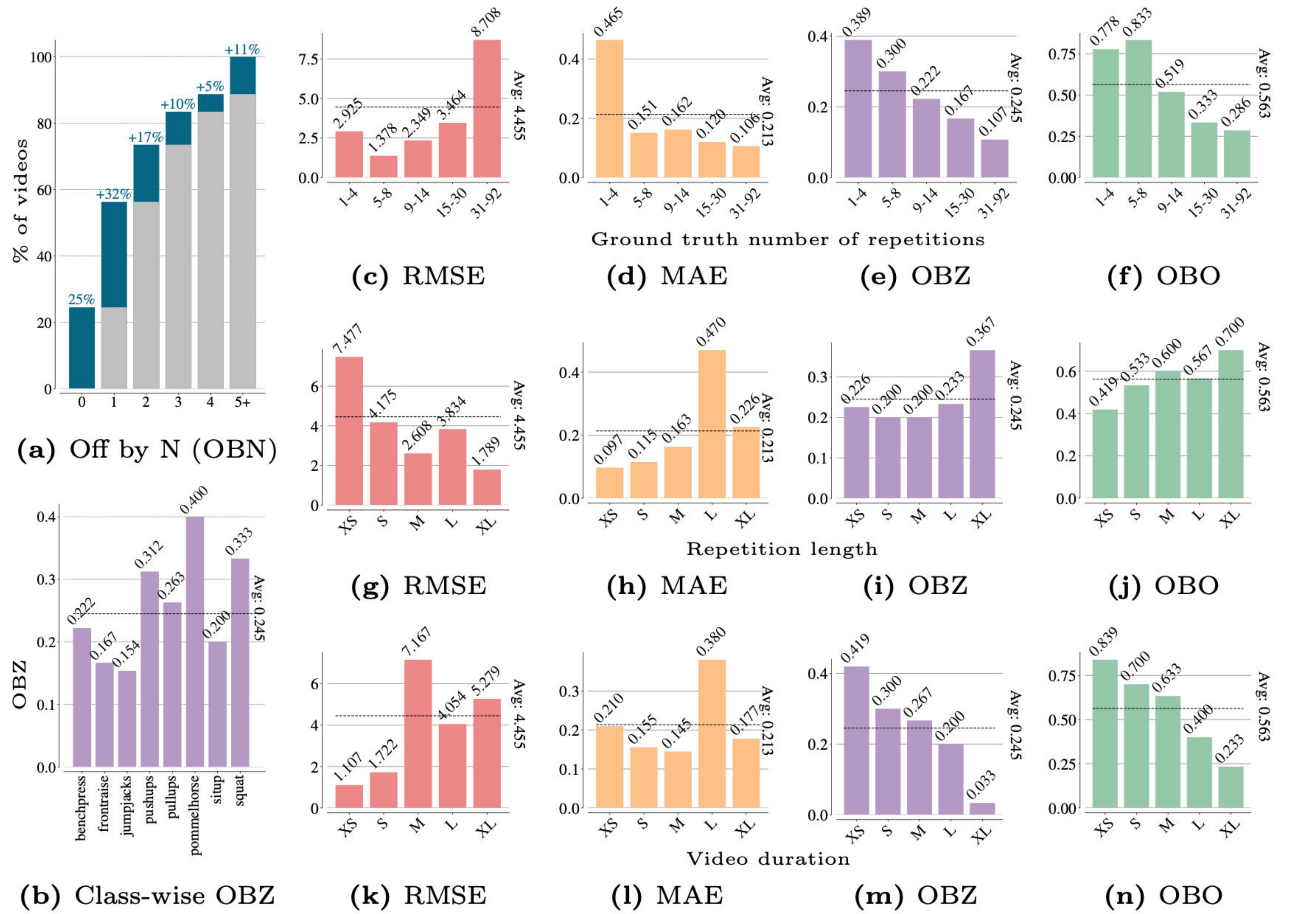
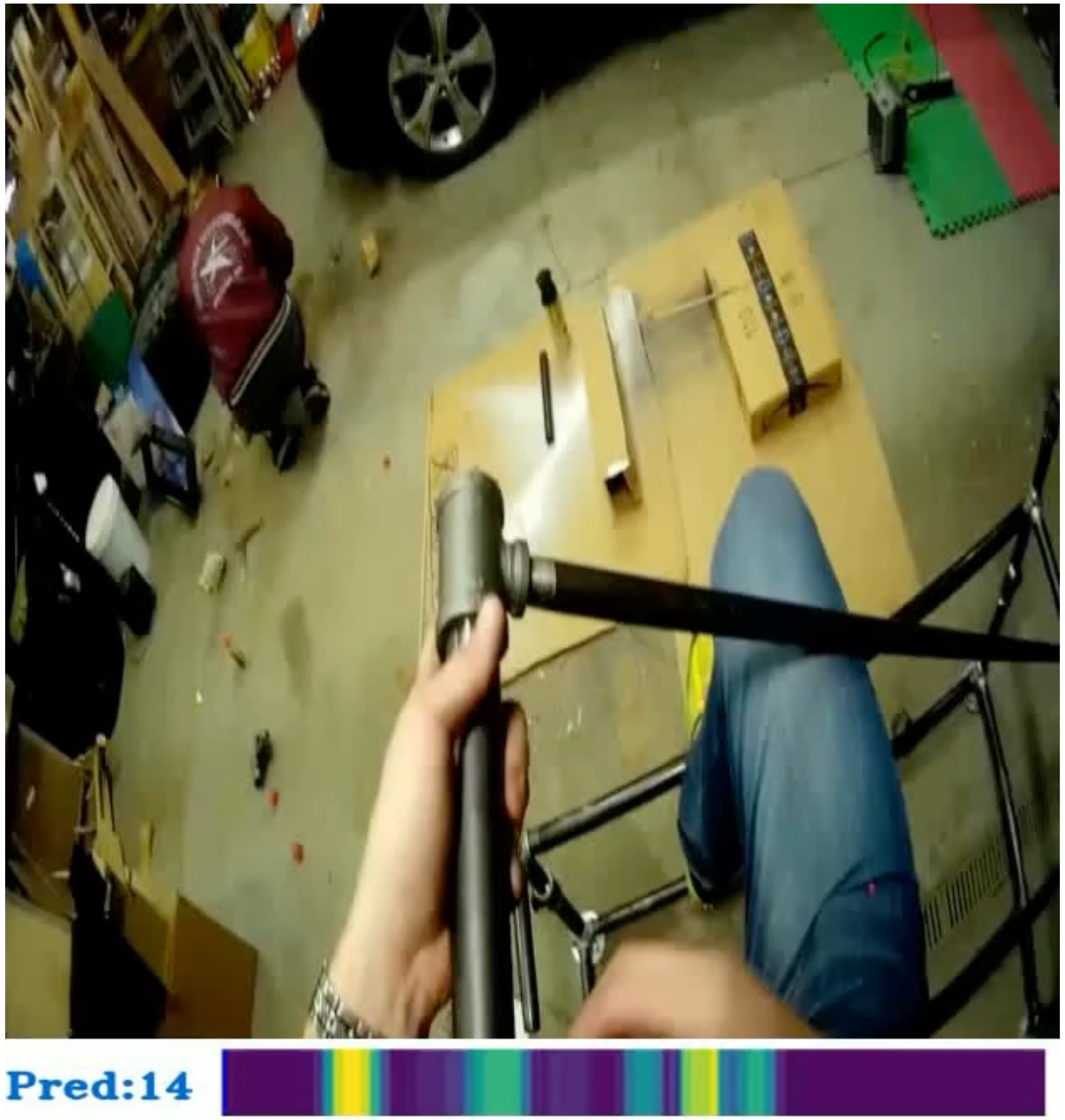
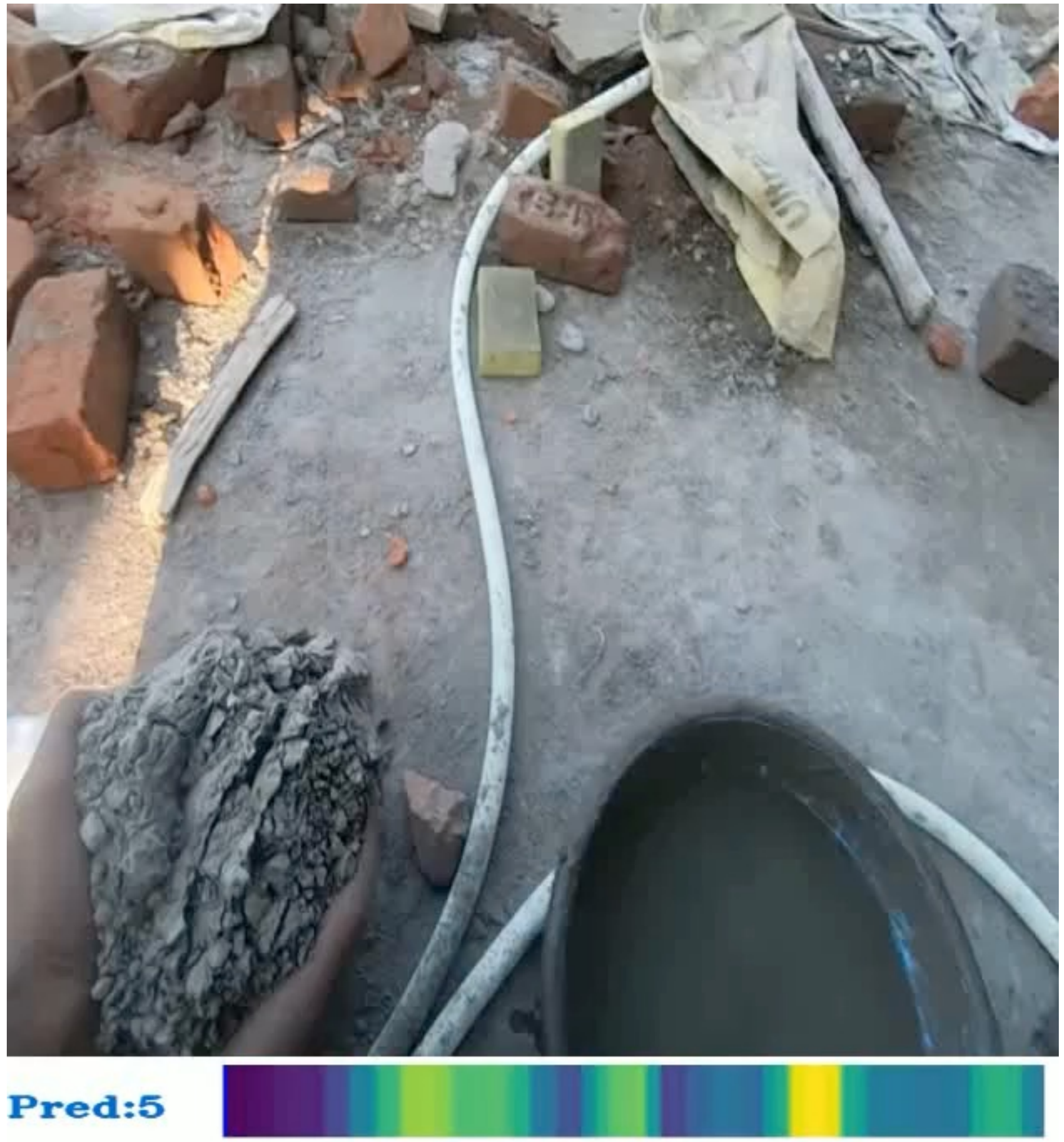


Fig. 6: Grouped VRC scores over different number of repetitions and lengths. (a) overviews the Off by N accuracy for increasing Ns. (b) shows OBZ by action class. The first row (c–f) reports results over different counts. (g–j) reports scores over groups by repetition durations. (k–n) reports metrics grouped by video duration.

Every Shot Counts - Ego4D

with: Saptarshi Sinha
Alexandros Stergiou





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

Long-Form Egocentric Video Understanding

No Semantic Supervision

No Shots - Temporal Alignment

Audio-Visual Semantic Gap

Quick View Changes

Repeating Actions

Long Continuous Streams

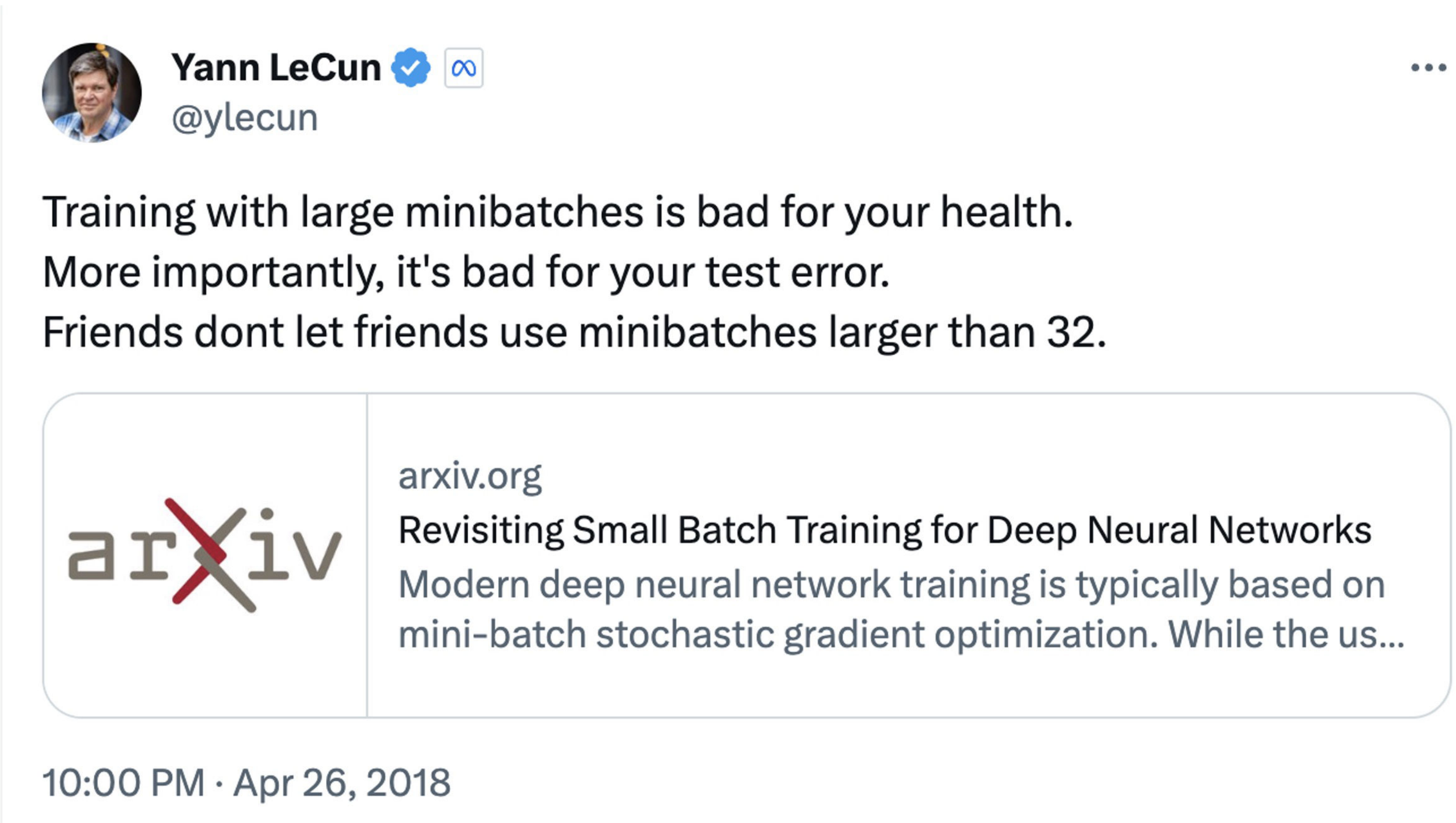
Fri (Session 6)
Poster # 443

Learning from One Continuous Video Stream

João Carreira, Michael King, Viorica Patraucean, Dilara Gokay, Catalin Ionescu, Yi Yang, Daniel Zoran, Joseph Heyward, Carl Doersch, Yusuf Aytar, Dima Damen, Andrew Zisserman

Learning from One Continuous Video Stream

- Original SGD = batch size 1 — this works fine

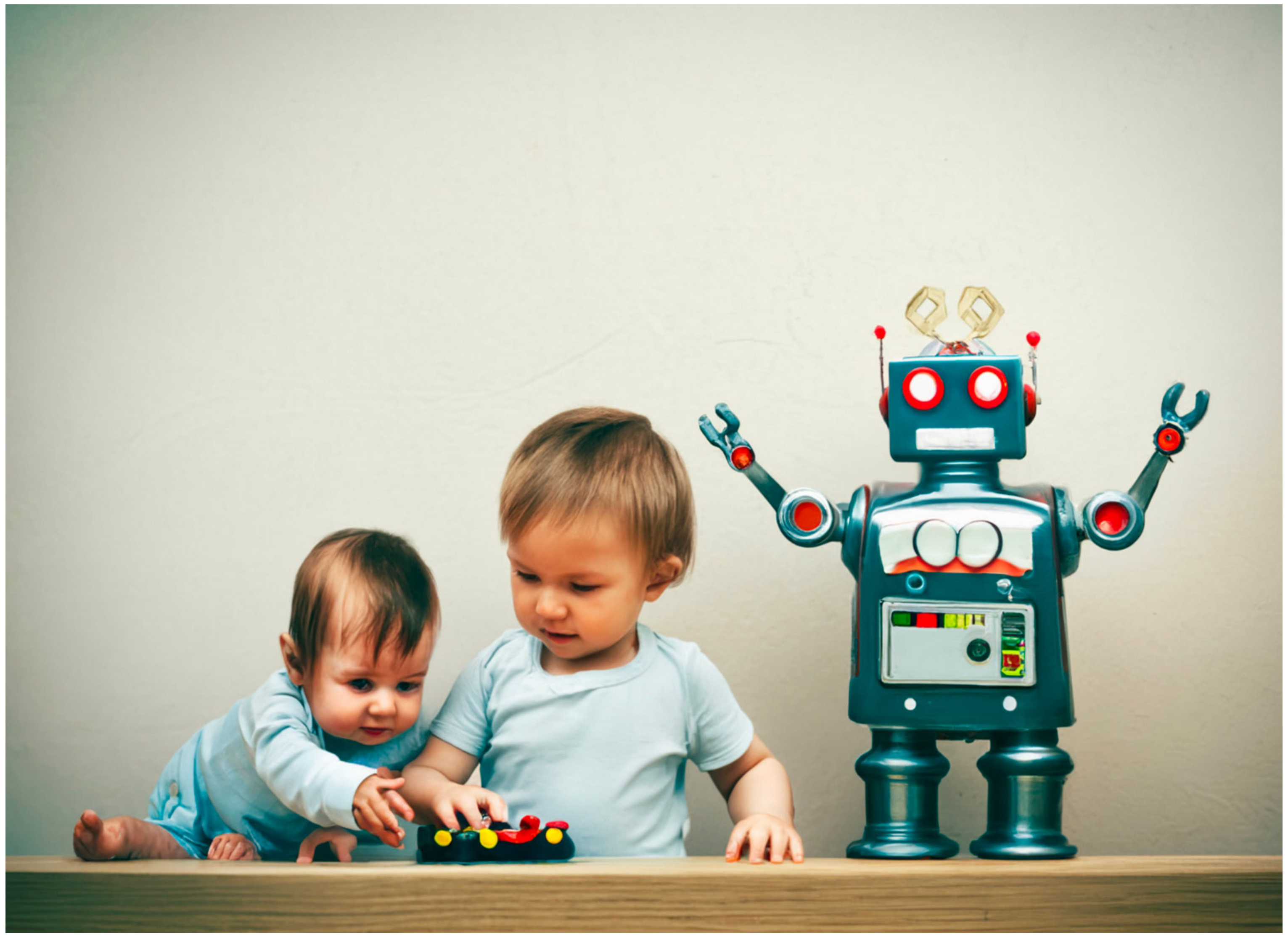


Large-scale machine learning with stochastic gradient descent. Bottou et al

Revisiting Small Batch Training for Deep Neural Networks, Masters et al

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

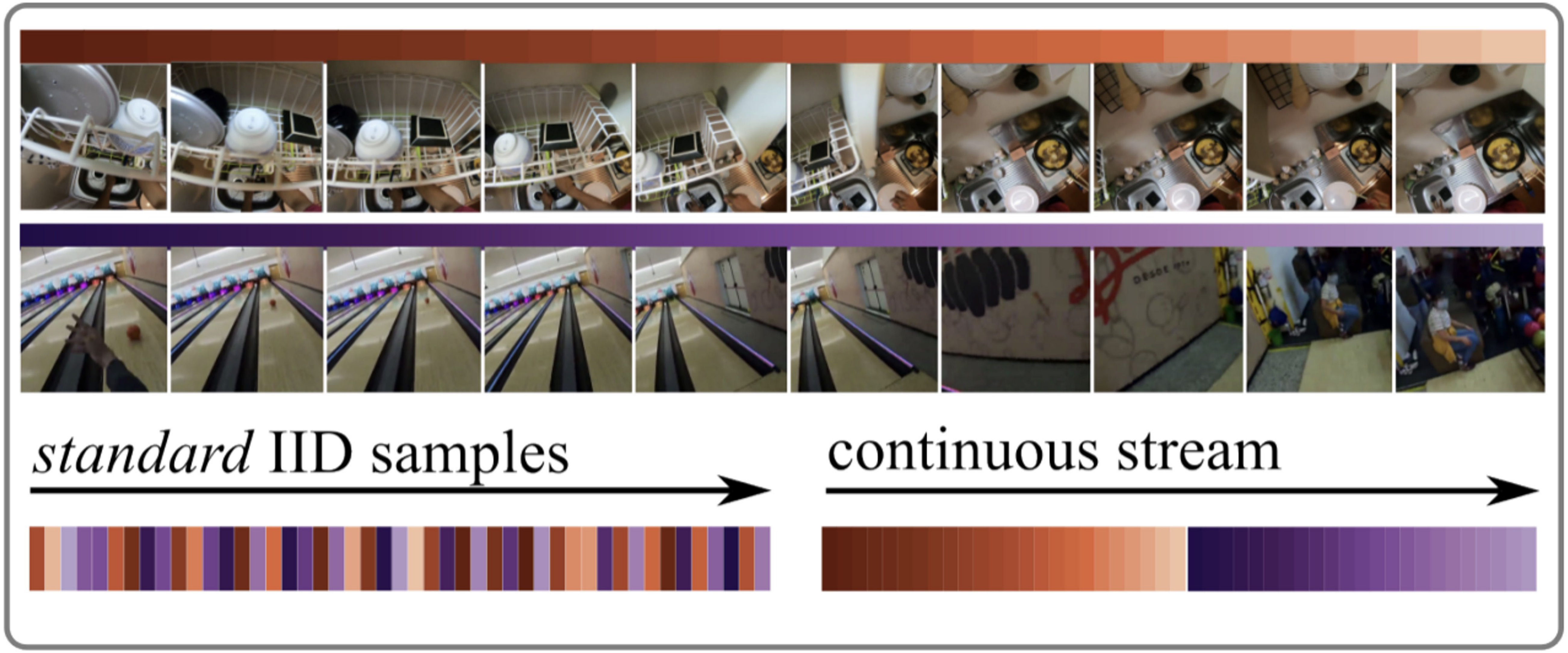
Learning from One Continuous Video Stream



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

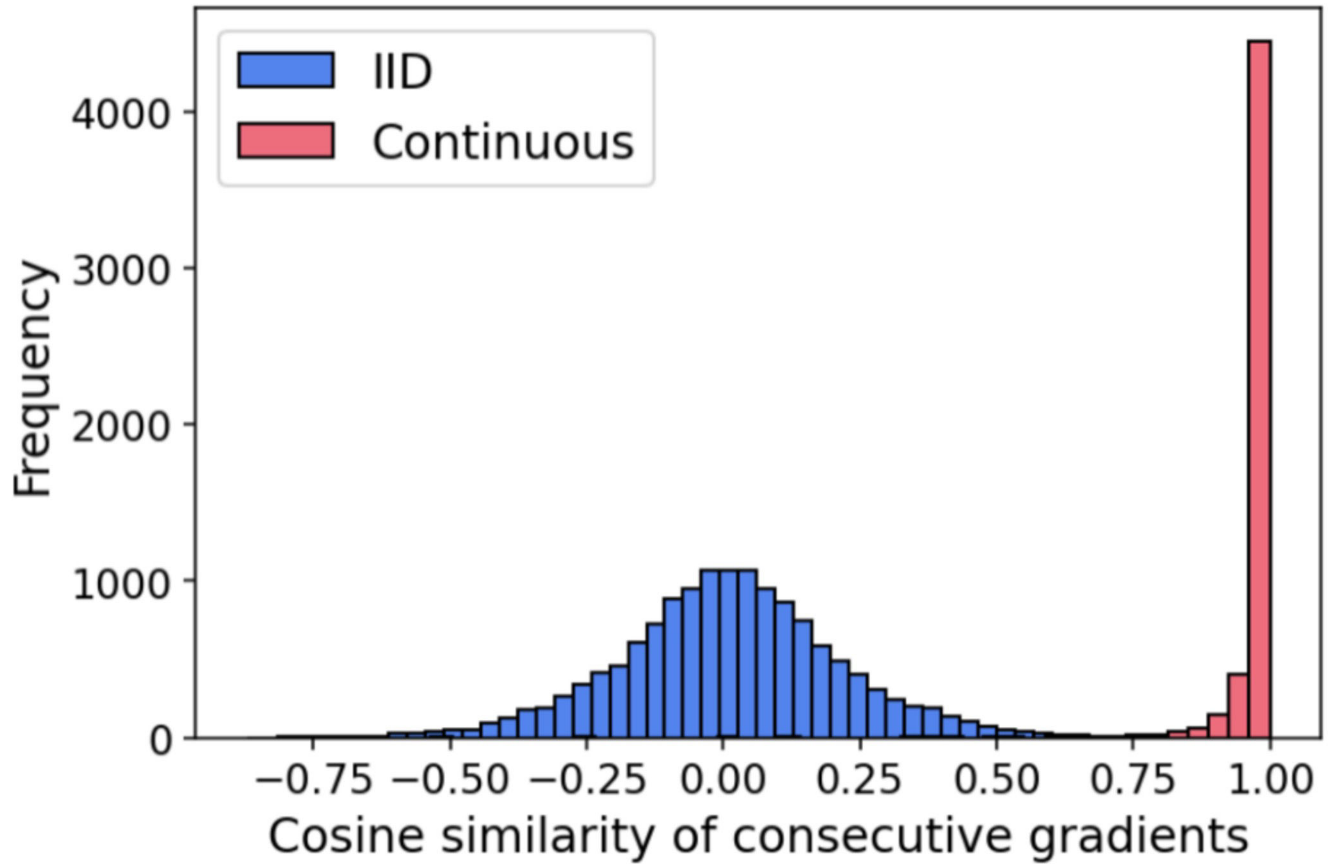
Dima Damen
LOVEU @CVPR2024

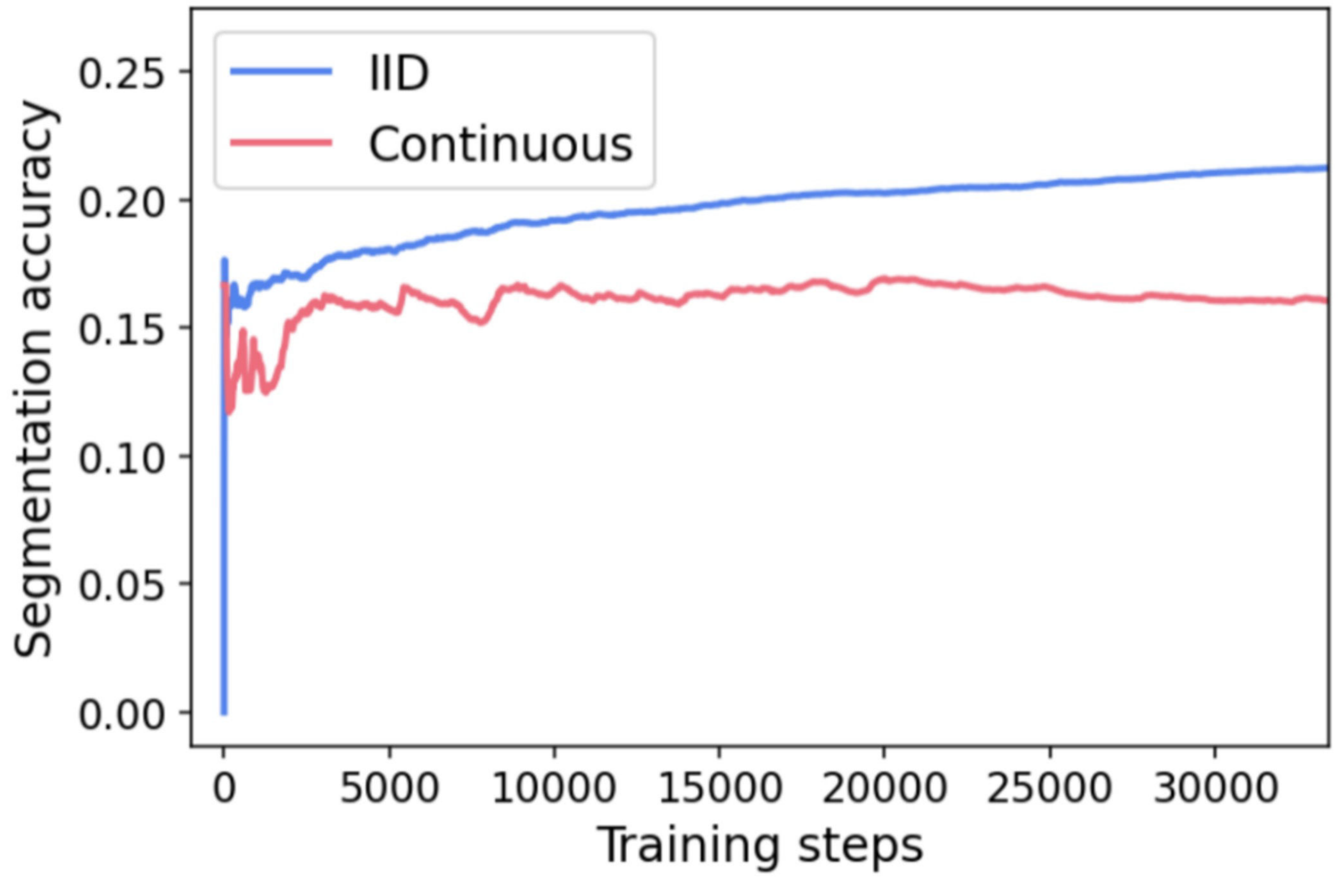
Learning from One Continuous Video Stream



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

Learning from One Continuous Video Stream





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

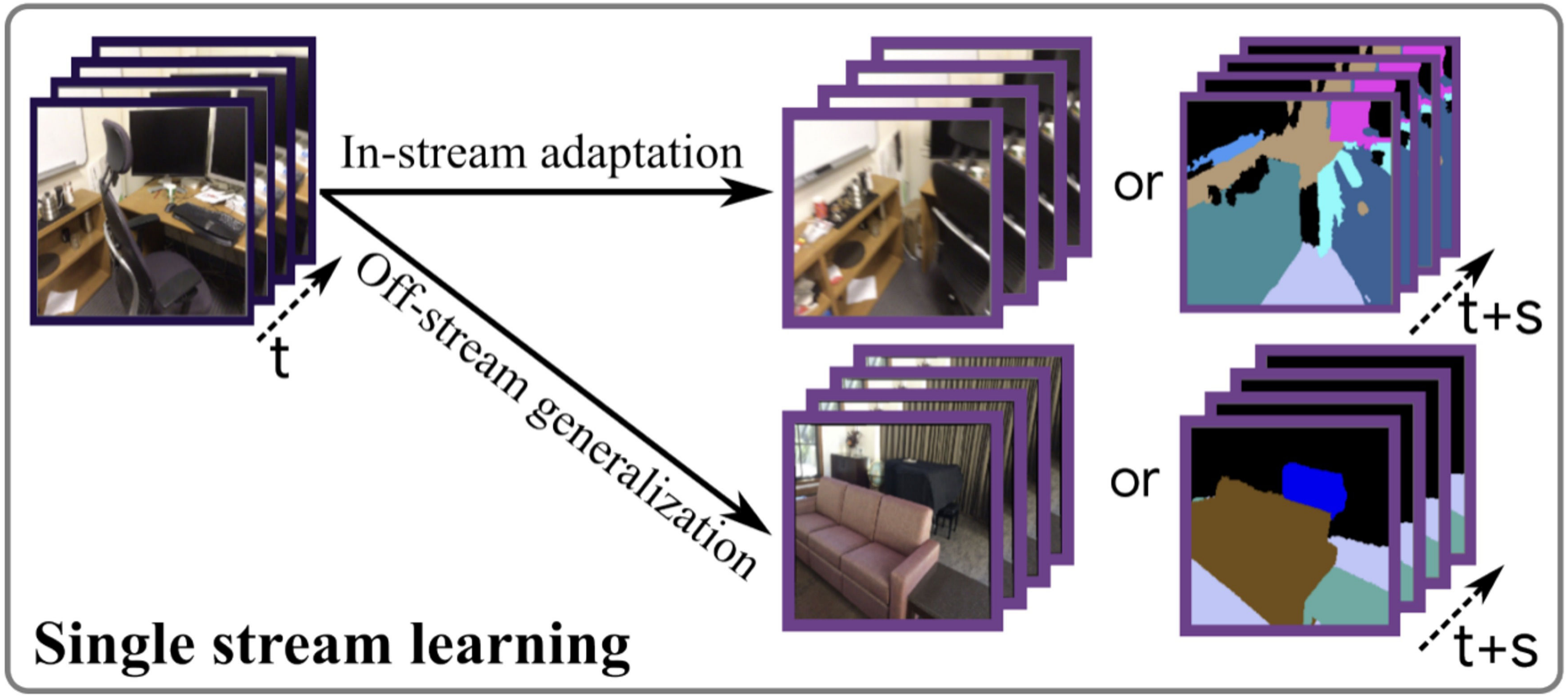
Learning from One Continuous Video Stream

Stream name	# videos train	# frames train	# videos val	# frames val	Max. length	Median length
Ego4D-stream	21,704	294M (3,265h)	2302	31M (348h)	1.95h	8.8 minutes
ScanNet-stream	1,199	1.8M (20h)	312	0.5M (5.7h)	5.5 minutes	1 minute

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

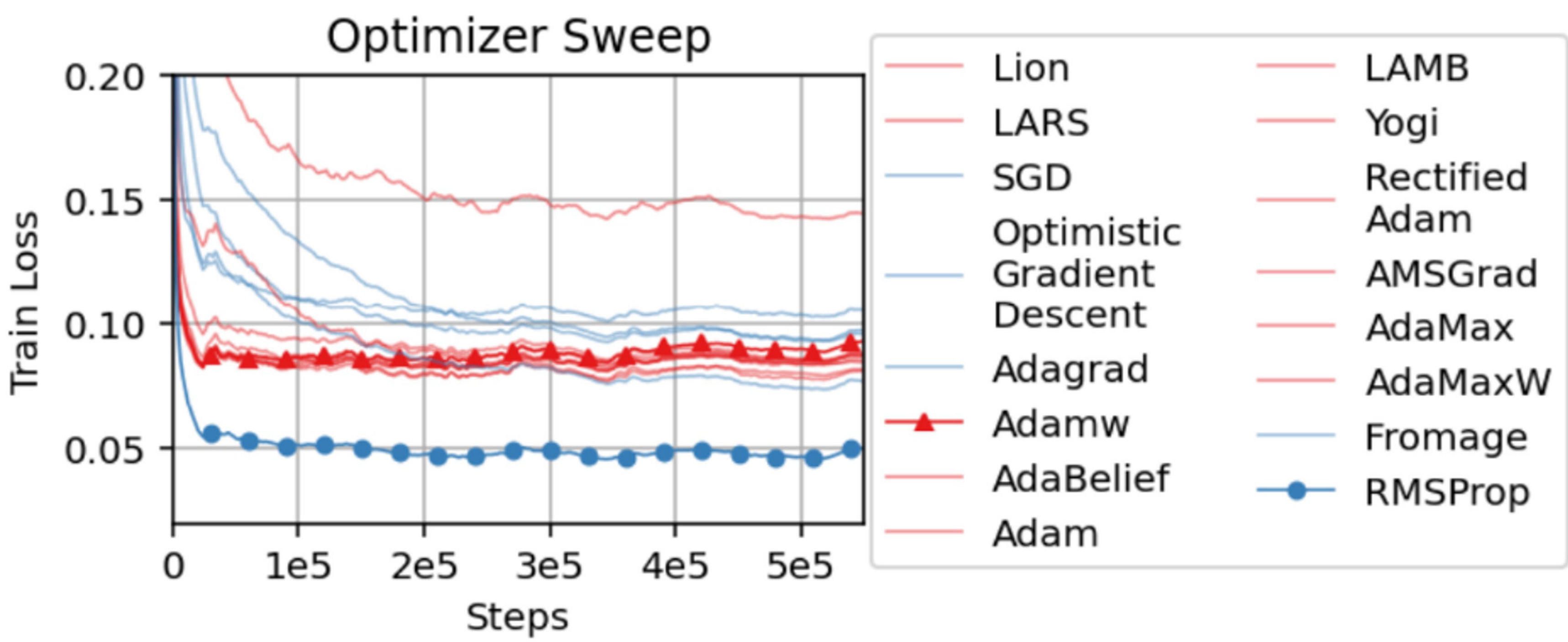
Learning from One Continuous Video Stream

- Future-frame pixel prediction (or supervised semantic seg)
- 2 models: VIT-L and UNet with self-attention
- Input 4 frames, output 4 frames (stacked along channel dimension)



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

Learning from One Continuous Video Stream



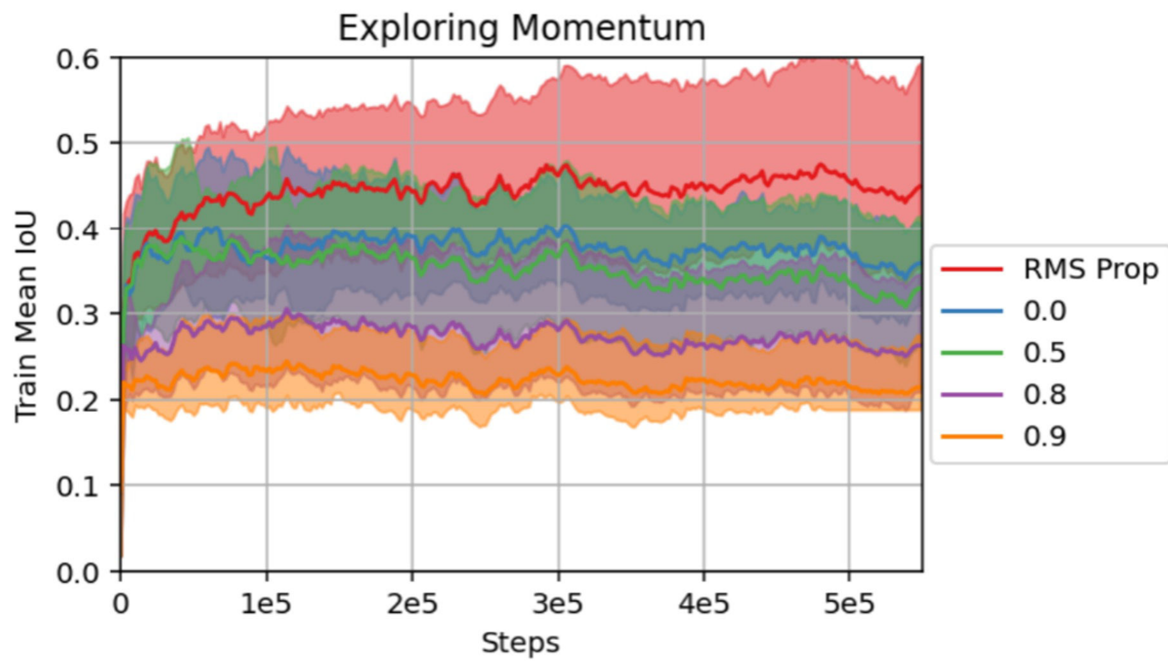


Figure 5. Reducing momentum with the AdamW optimizer helps to recover some of the performance of RMS Prop.

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

Dima Damen
LOVEU @CVPR2024

Fri (Session 6)
Poster # 443

Learning from One Continuous Video Stream

João Carreira, Michael King, Viorica Patraucean, Dilara Gokay, Catalin Ionescu, Yi Yang, Daniel Zoran, Joseph Heyward, Carl Doersch, Yusuf Aytar, Dima Damen, Andrew Zisserman

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

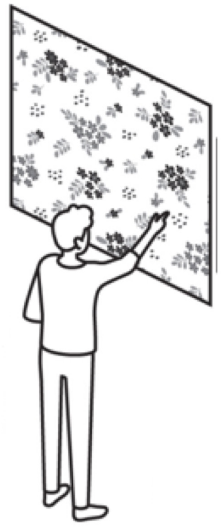
How did we do this?

University of BRISTOL

C Plizzari et al (2024). An Outlook into the Future of Egocentric Vision. IJCV

Dima Damen
LOVEU @CVPR2024

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

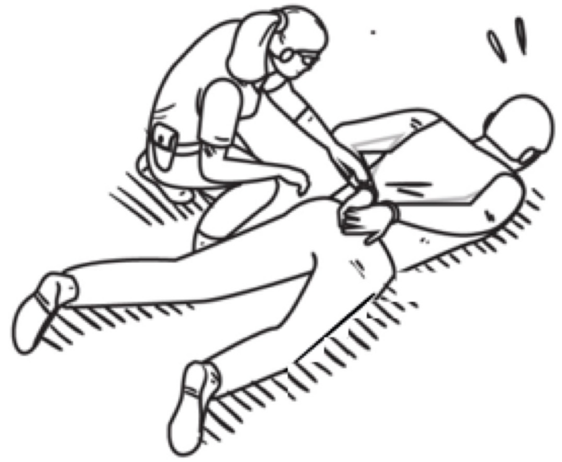


EGO-Designer

EGO-Worker

EGO-Tourist

EGO-Home



Ego-Police

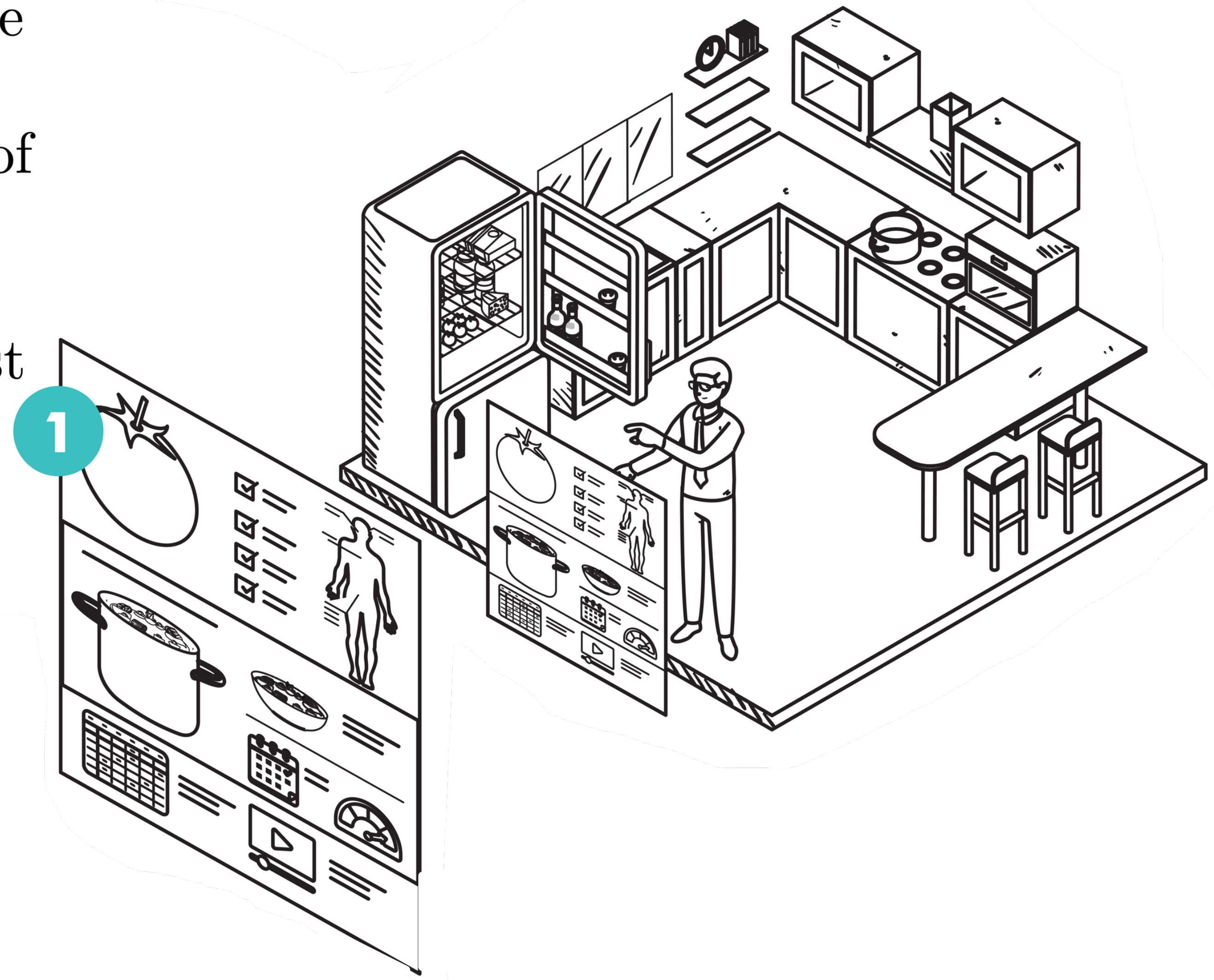
University of BRISTOL

C Plizzari et al (2024). An Outlook into the Future of Egocentric Vision. IJCV

EGO-Home

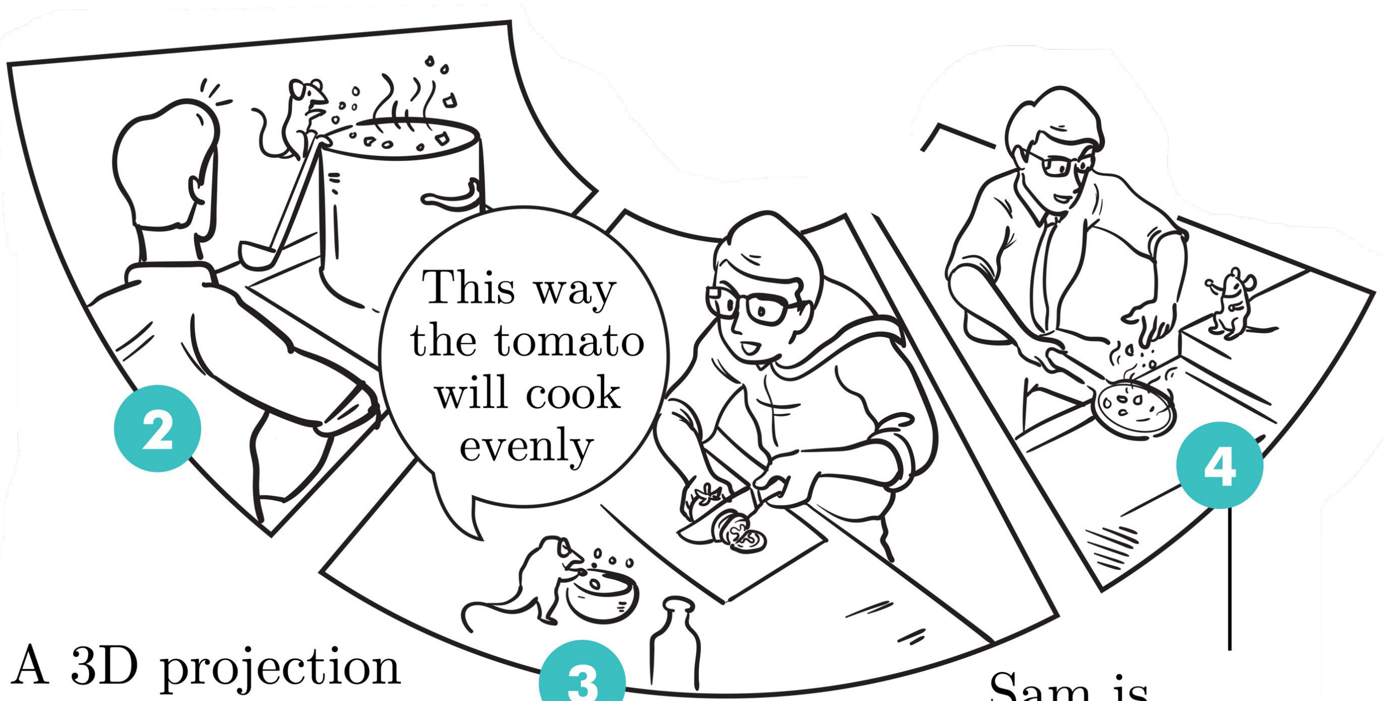
with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi

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



EGO-Home

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

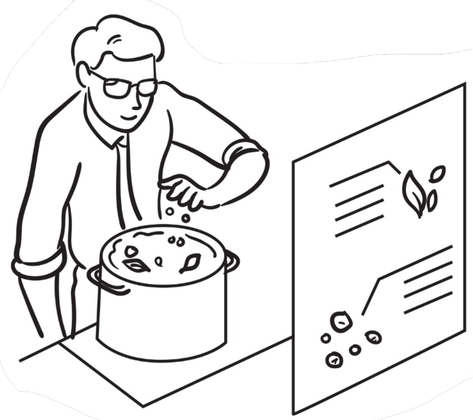


2 A 3D projection of Remy helps with cooking

3 Audible 3D projection

4 Sam is impressed by how fun it is to cook with his 3D friend

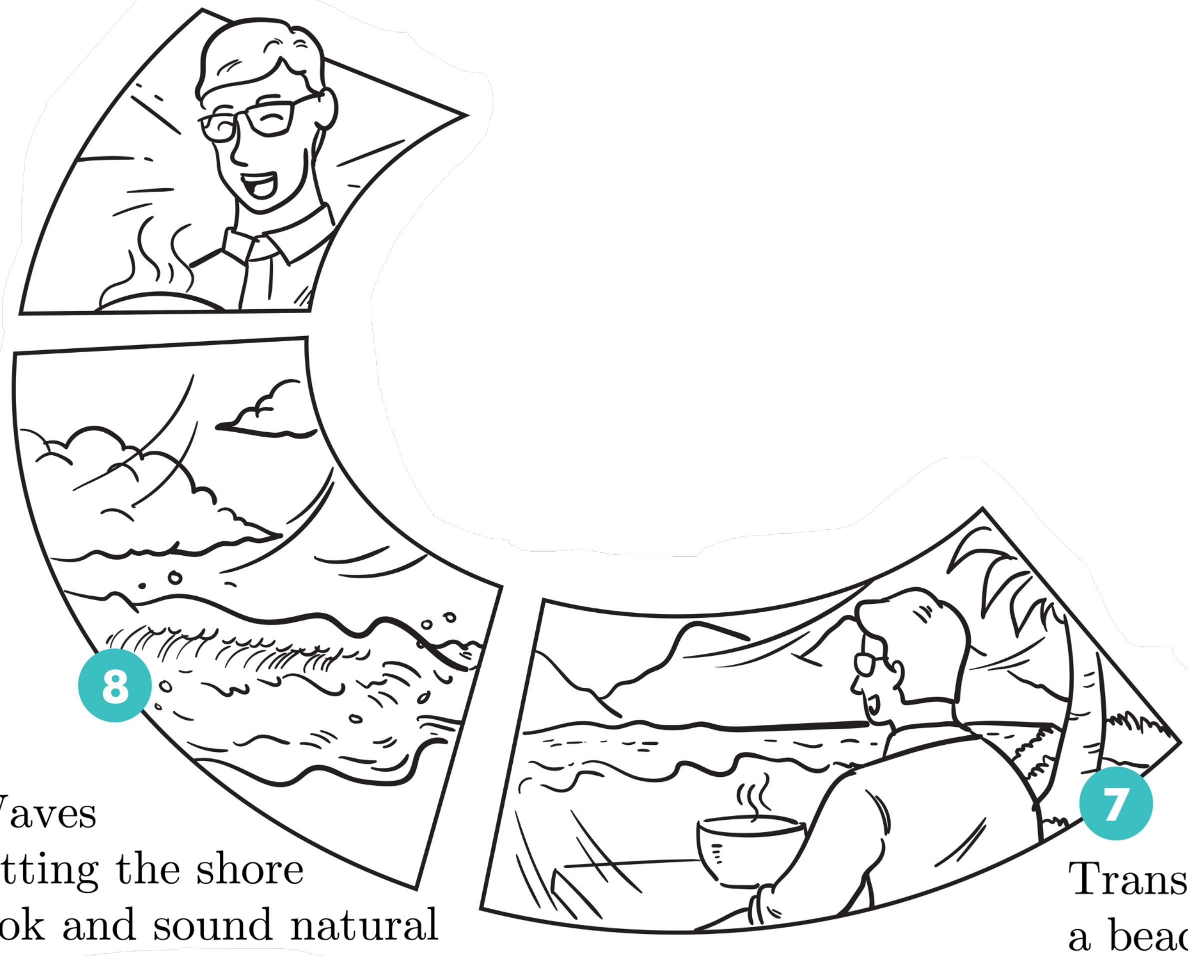
Toaster reminder



6 EgoAI recommends some more spice

EGO-Home

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



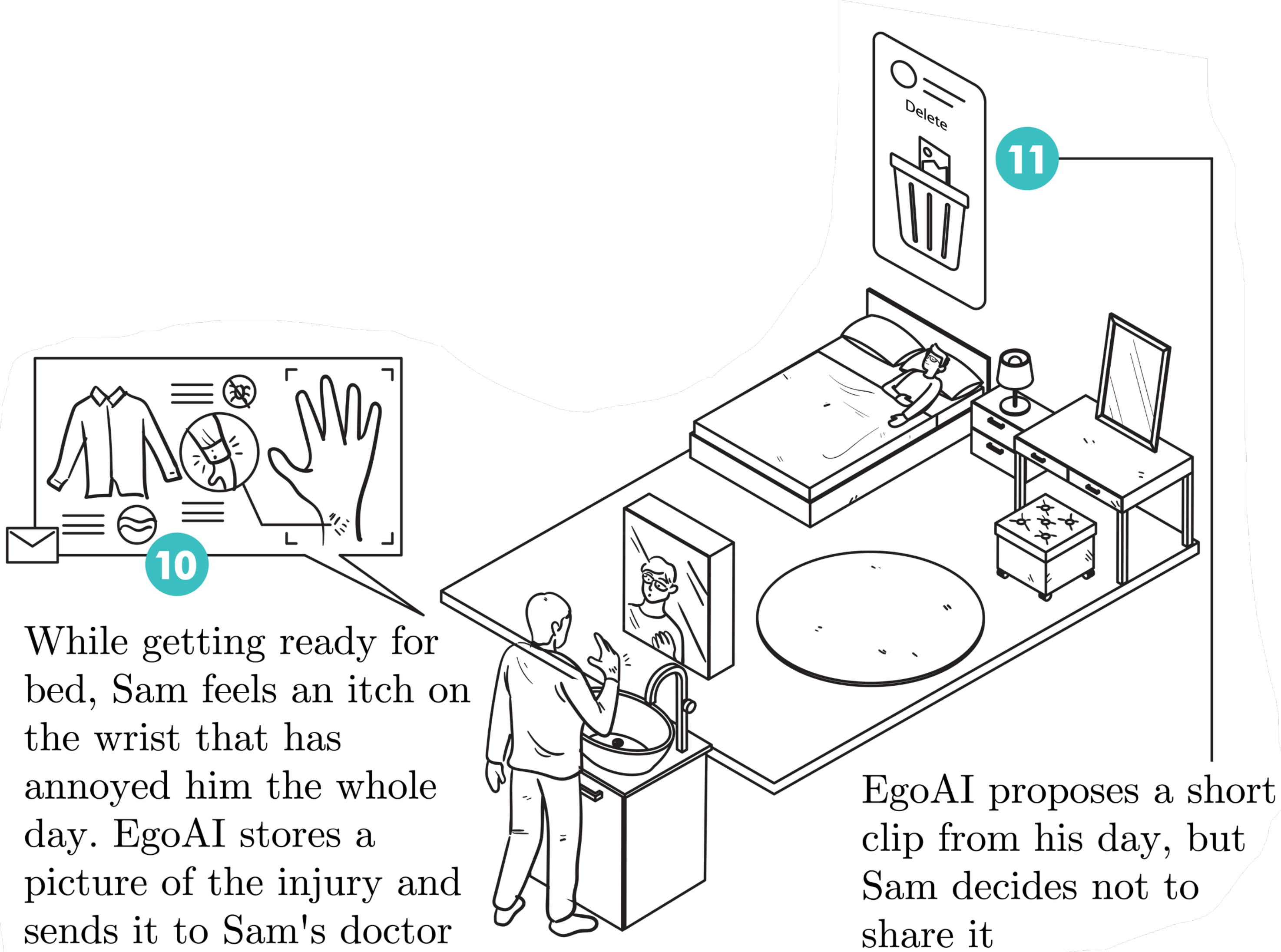
Waves hitting the shore look and sound natural

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

EGO-Home

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



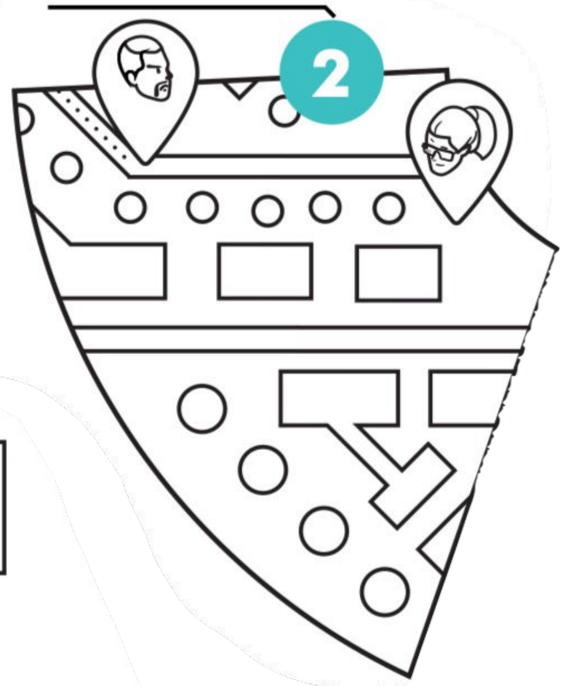
While getting ready for bed, Sam feels an itch on the wrist that has annoyed him the whole day. EgoAI stores a picture of the injury and sends it to Sam's doctor for advice

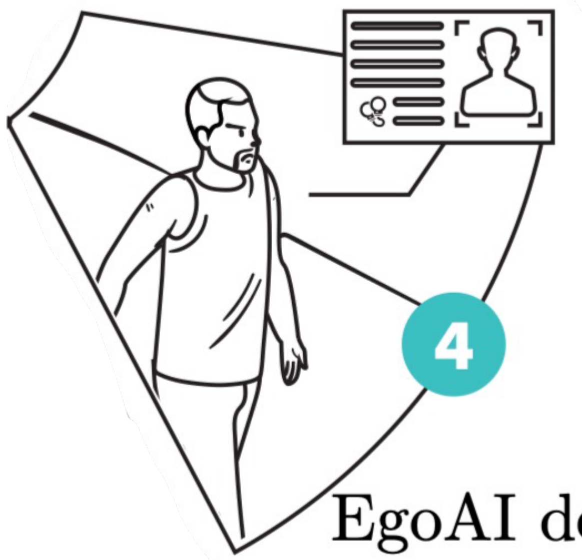
EgoAI proposes a short clip from his day, but Sam decides not to share it

From Stories to Tasks

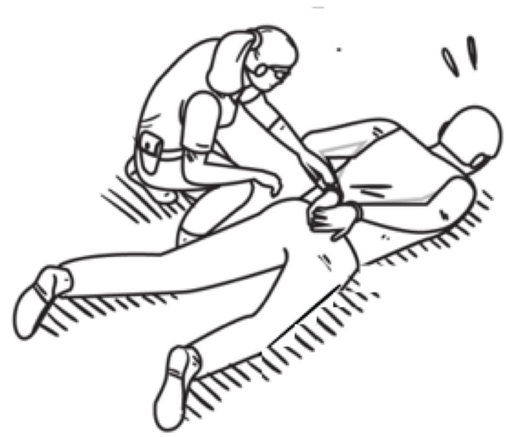
with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi

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





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



EGO-Police

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

The Survey Part

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

Seminal Works

State-of-the-art

Datasets

For the Future

Localisation and Navigation 1 2

Messaging 1 3 11

Action Recognition 2 13

The Survey Part

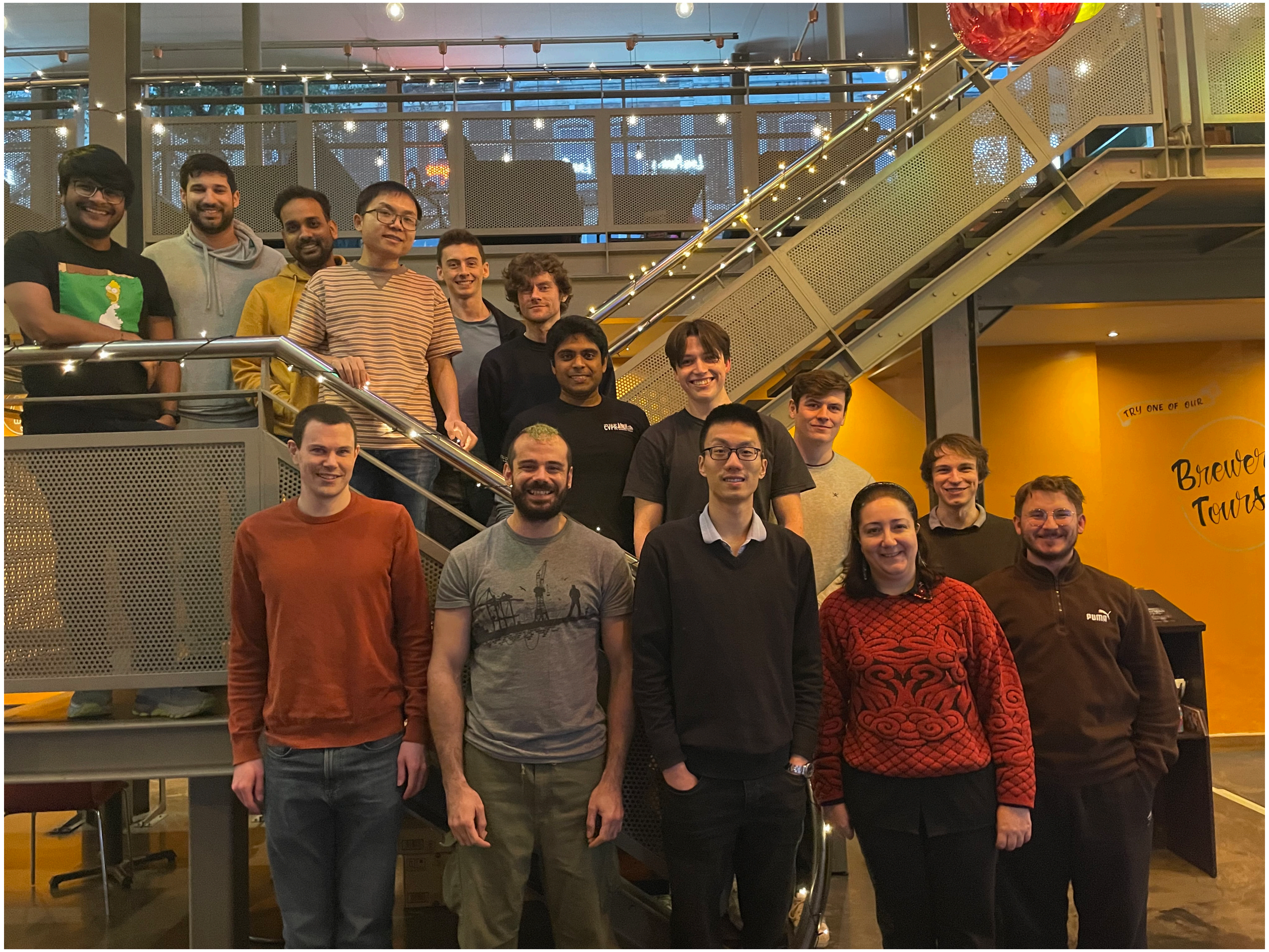
with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi

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

Long-Form Egocentric Video Understanding

- No Semantic Supervision
- No Shots - Temporal Alignment
- Audio-Visual Semantic Gap
- Quick View Changes
- Repeating Actions
- Long Continuous Streams

The Team



University of BRISTOL

Thank you

For further info, datasets, code, publications…

http://dimadamen.github.io

@dimadamen

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

Q&A