On Video, Audio and Language Multi-Modality in Egocentric Vision

University of BRISTOL

Dima Damen
MULA@CVPR2024

Multi-Modality in Egocentric Data

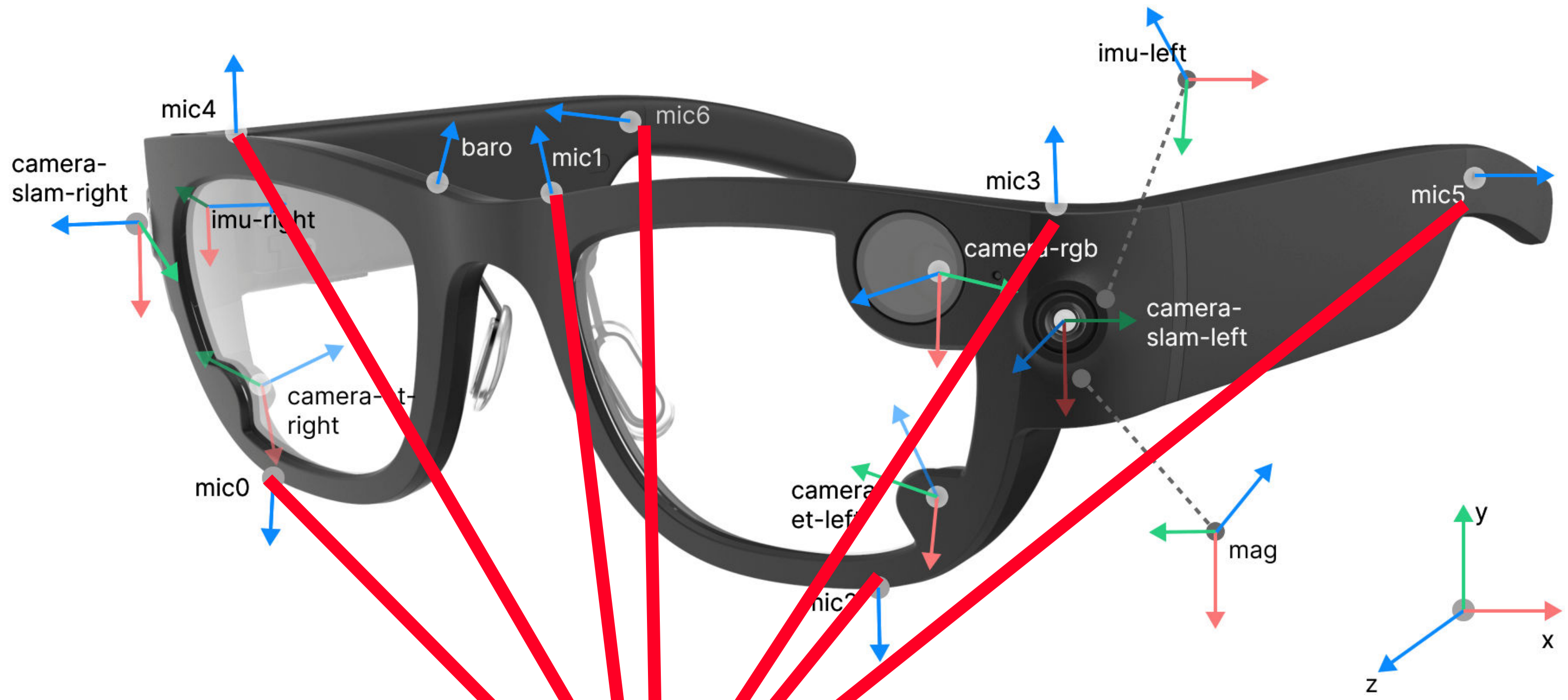
V — High frame-rate RGB footage from the camera wearer's perspective

A — One or many microphones, on the wearable device, best positioned to capture the sounds of actions and interactions

L — Speech in the video... or
Narrations/Captions added to index the videos

Egocentric Data Collection

microphone

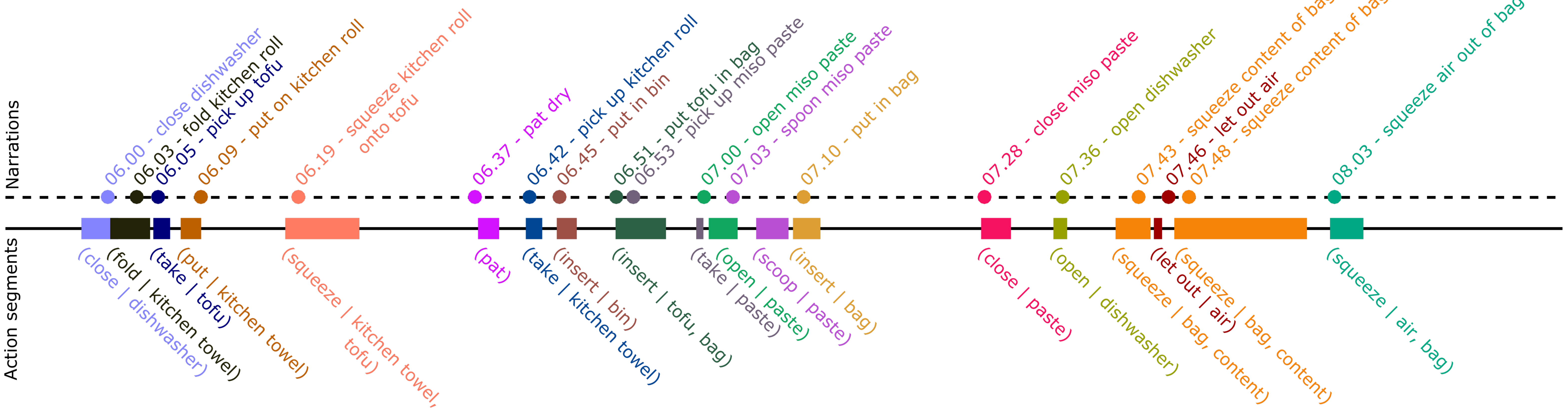


array of microphones

University of BRISTOL

Dima Damen

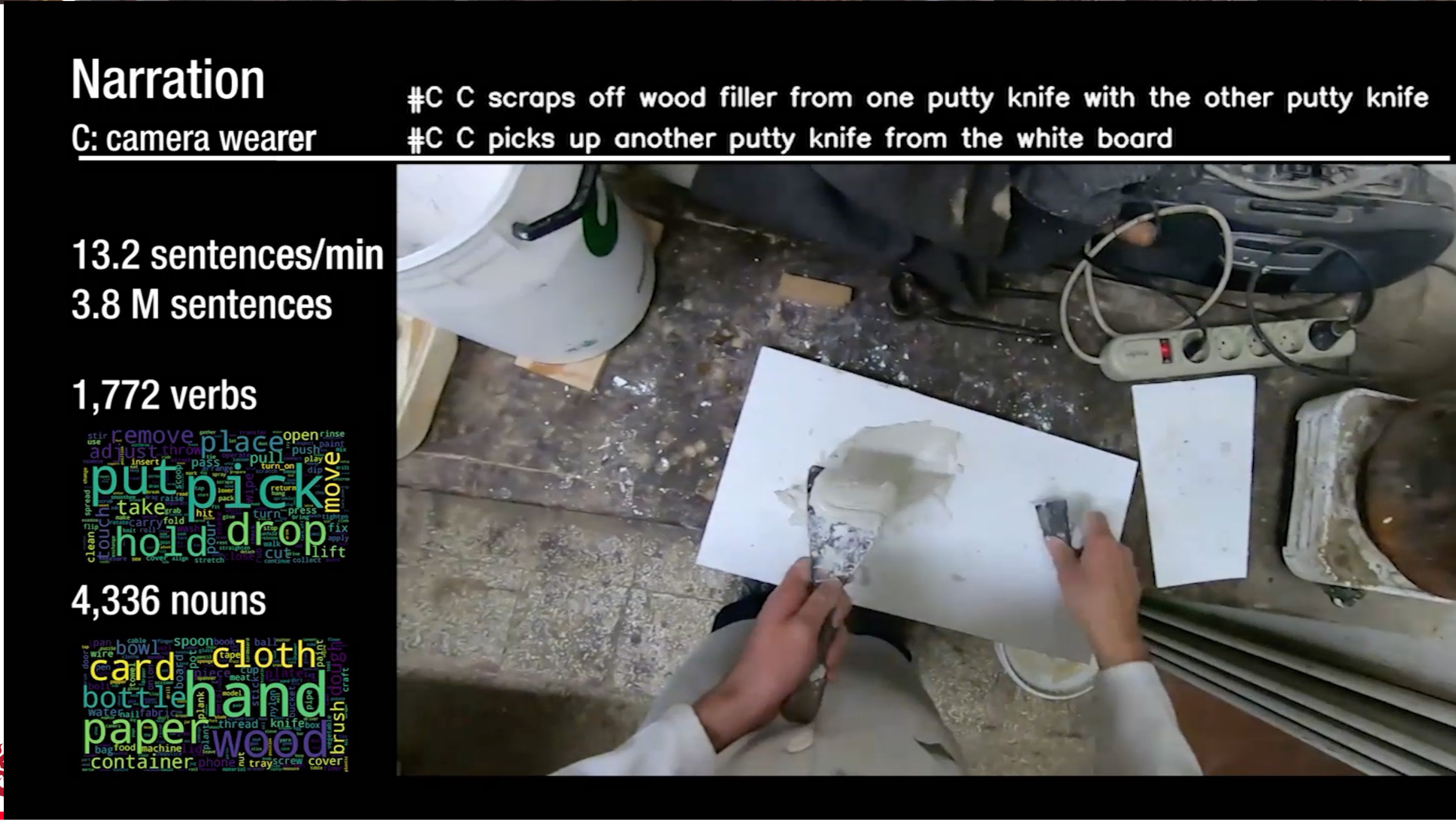
From Narrations to Start-End Times



University of BRISTOL

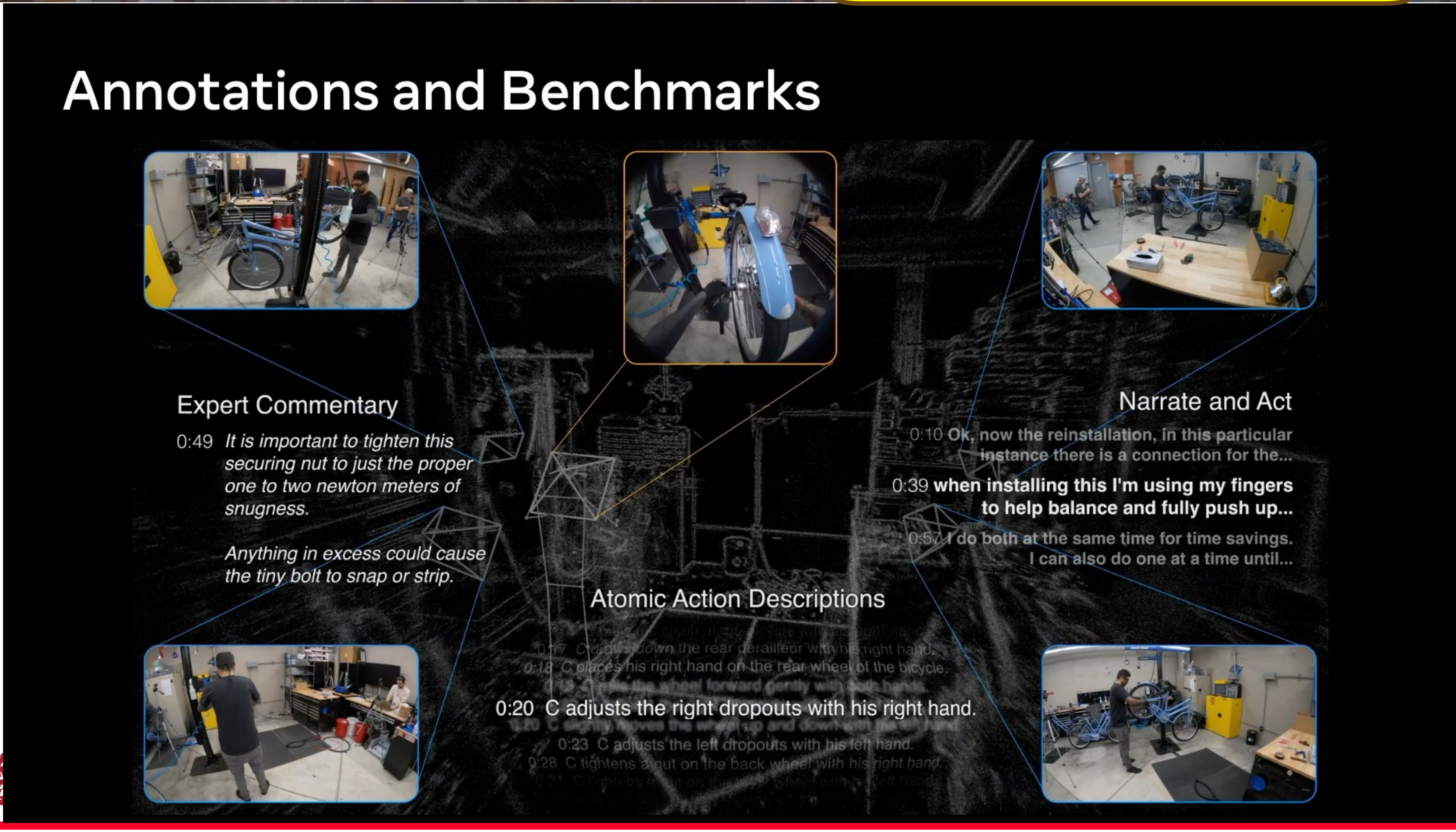
Dima Damen
MULA@CVPR2024

Ego4D Narrations



Ego-Exo4D Annotations

Fri Oral + Poster (Session 5 #292)



Multi-Modality in Egocentric Data

V

High frame-rate RGB footage from the camera wearer's perspective

A

One or many microphones, on the wearable device, best positioned to capture the sounds of actions and interactions

L

~~Speech in the video... or~~
Narrations/Captions added to index the videos

Multi-Modality in Egocentric Data

V

High frame-rate RGB footage from the camera wearer's perspective

A

One or many microphones, on the wearable device, best positioned to capture the sounds of actions and interactions

University of BRISTOL

Audio-Visual Egocentric Vision

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

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

Audio-Visual Egocentric Vision

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

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

Audio-Visual Egocentric Vision

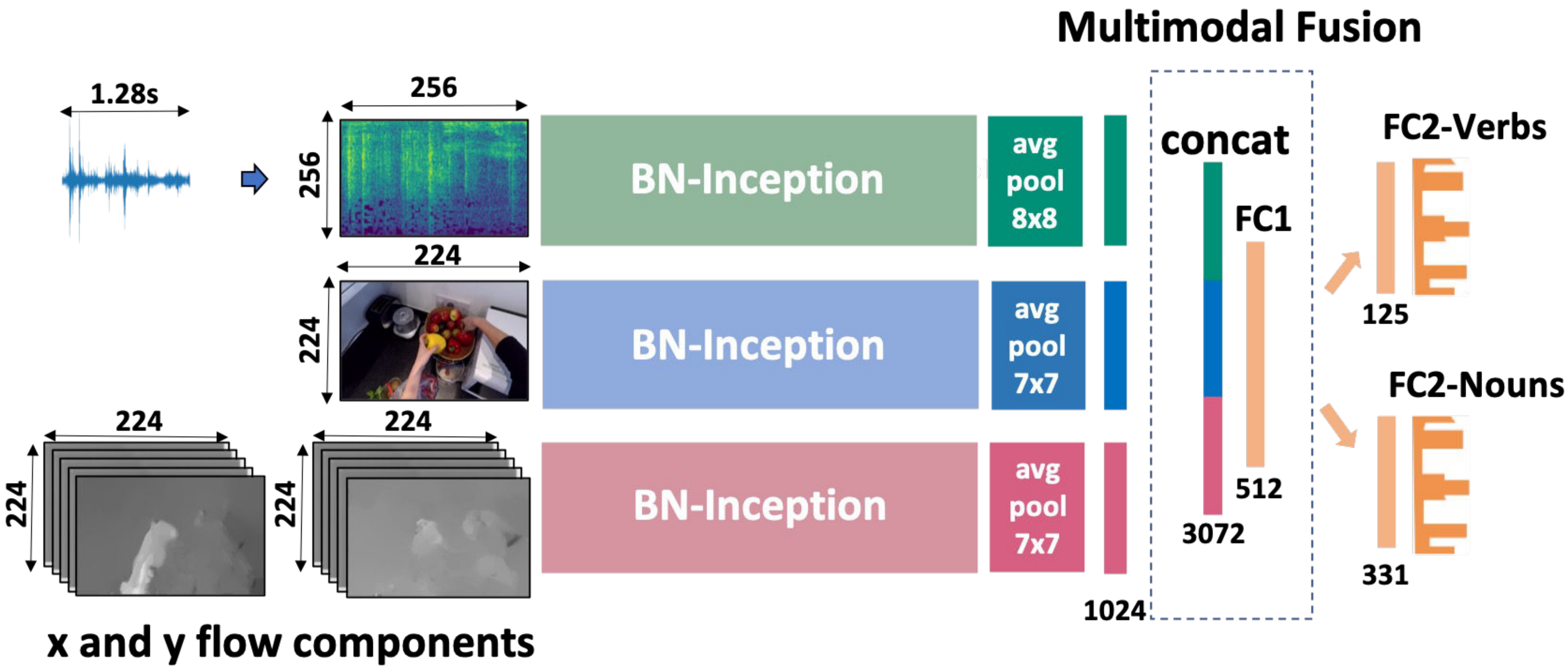
with: Vangelis Kazakos
Arsha Nagrani.
Andrew Zisserman
Jaesung Huh
Jacob Chalk

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

University of BRISTOL

The first attempt

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman



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

Objects that Sound

- musical Instruments
- animals and insects
- waterfall
- humans talking
- food processor

Actions that Sound

- put glass down
- close drawer
- turn-on tap
- chop garlic

Audio-Visual Egocentric Vision

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

Objects that Sound

- music
- anima
- water
- huma
- food p

Actions that Sound

- put gl
- close drawer
- turn-on tap
- chop garlic

Harmonic vs Percussive

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman

Harmonic Sounds

Percussive Sounds

EPIC-KITCHENS

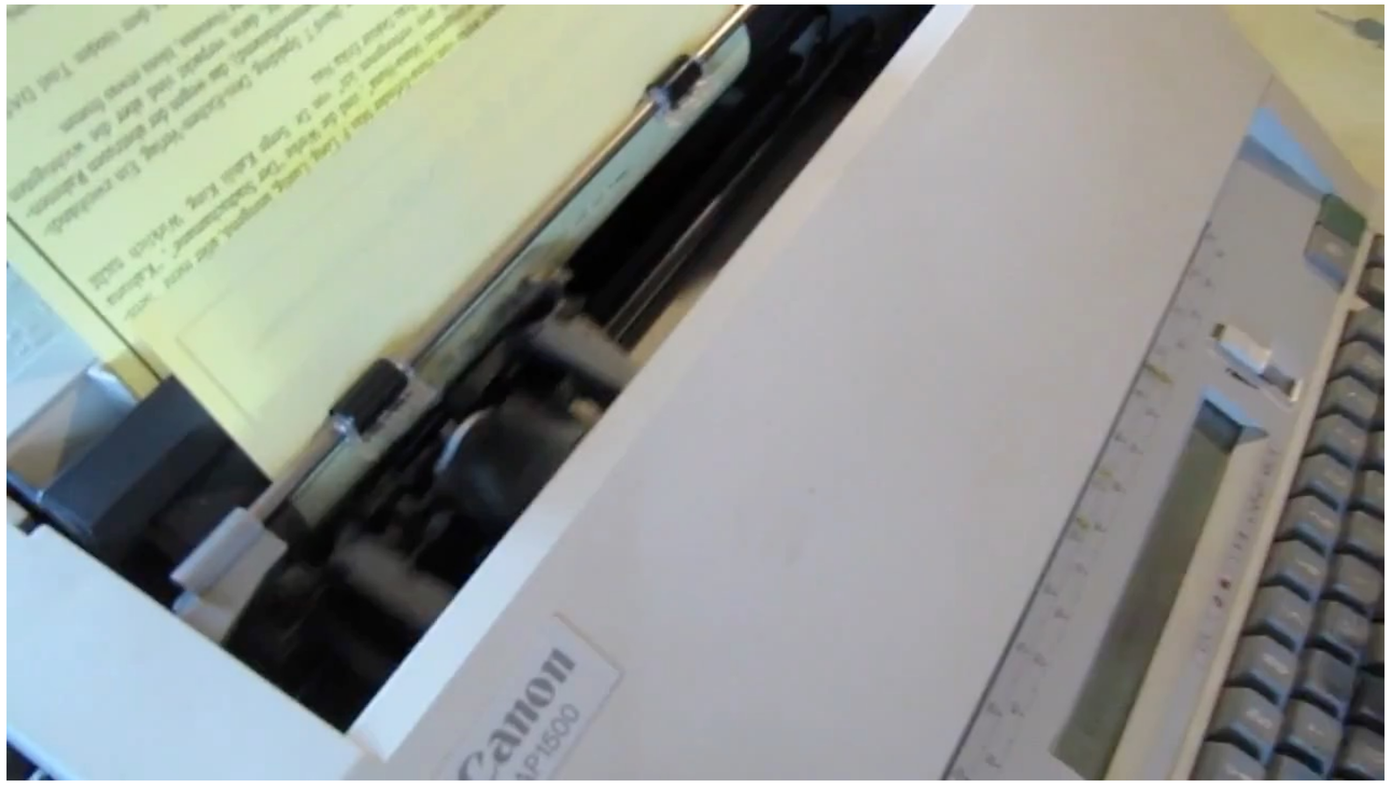
Harmonic vs Percussive

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman

VGG-Sound

Harmonic Sounds

Percussive Sounds

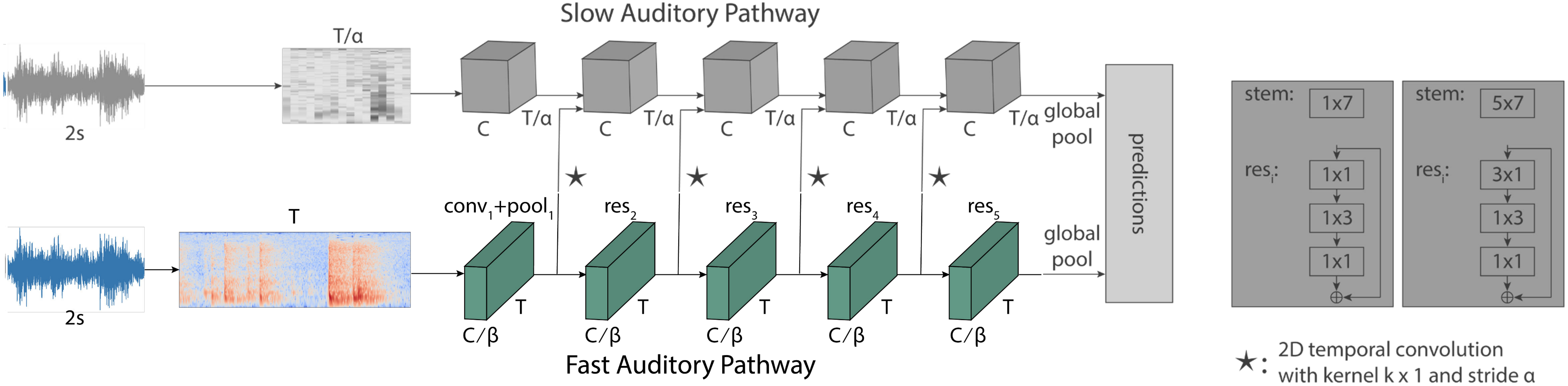


Auditory Slow-Fast

Outstanding Paper Award – ICASSP 2021

Audio Slow-Fast

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman



Audio Slow-Fast

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman

VGG-Sound

Model	Top-1	Top-5
Chen et al. [2]	51.00	76.40
McDonnell & Gao [3]	39.74	71.65
Slow	45.20	72.53
Fast	41.44	70.68
Slow-Fast (Proposed)	**52.46**	**78.12**

EPIC-KITCHENS

		Top-1 Accuracy (%)			
Split	Model	Verb	Noun	Action	# Param.
Test	Damen et al. [1]	42.12	21.51	14.76	10.67M
Test	Slow-Fast (Proposed)	**46.47**	**22.77**	**15.44**	26.88M

Audio Slow-Fast

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman

Slow stream

Animals

- baltimore oriole calling
- cheetah chirrup
- zebra braying
- dinosaurs bellowing
- horse neighing
- black capped chickadee calling
- cat hissing
- cuckoo bird calling
- mosquito buzzing
- bull bellowing
- whale calling

Scenes

- volcano explosion
- playing lacrosse
- hair dryer drying
- sea waves
- playing tympani
- blowtorch igniting
- opening/closing electric car windows
- thunder
- electric blender running
- playing shofar
- airplane flyby
- playing trumpet
- wind chime
- striking bowling

Fast stream

Percussive sounds

- footsteps on snow
- snake rattling
- tap dancing
- car engine knocking
- woodpecker pecking tree
- chopping wood
- people clapping
- lawn mowing
- typing on typewriter
- opening or closing car doors
- playing tennis
- railroad car
- playing tympani
- playing drum kit
- playing vibraphone
- popping pop corn

Voices

- singing choir
- people cheering
- people crowd
- child speech
- baby laughter

Others

- cat purring
- dog barking
- race car
- singing bowl
- vacuum cleaner cleaning floors
- toilet flushing
- dog growling
- splashing water

E Kazakos, A Nagrani, A Zisserman, D Damen (2021). Slow-Fast Auditory Streams For Audio Recognition. ICASSP

Audio Slow-Fast

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman

Slow stream		Fast stream	
Animals	baltimore oriole calling cheetah chirrup zebra braying dinosaurs bellowing horse neighing black capped chickadee calling cat hissing cuckoo bird calling mosquito buzzing bull bellowing whale calling	Percussive sounds	footsteps on snow snake rattling tap dancing car engine knocking woodpecker pecking tree chopping wood people clapping lawn mowing typing on typewriter opening or closing car doors playing tennis railroad car playing tympani playing drum kit playing vibraphone popping pop corn
Scenes	volcano explosion playing lacrosse hair dryer drying sea waves playing tympani blowtorch igniting opening/closing electric car windows thunder electric blender running playing shofar airplane flyby playing trumpet wind chime striking bowling	Voices	singing choir people cheering people crowd child speech baby laughter
		Others	cat purring dog barking race car singing bowl vacuum cleaner cleaning floors toilet flushing dog growling splashing water

Audio Slow-Fast

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman

TOWARDS LEARNING UNIVERSAL AUDIO REPRESENTATIONS

Luyu Wang, Pauline Luc, Yan Wu, Adrià Recasens, Lucas Smaira, Andrew Brock, Andrew Jaegle,

Table 2: **Evaluating frameworks and architectures on HARES.** We compare the impact of architecture choice under the classification and SimCLR objective. We also show the performance of several other recent strongly performing frameworks. Average scores are reported for tasks in each domain separately, and all three combined. All models are trained on AudioSet except for bidirectional CPC and Wav2Vec2.0, for which we also show results when they are trained on LibriSpeech (LS).

Architecture	#Params	Input format	Used in	Env.	Speech	Music	HARES	AudioSet (mAP)
				Classification/SimCLR				
BYOL-A CNN	5.3m	Spectrogram	[9]	69.4/69.9	61.4/69.8	57.6/63.1	63.1/68.2	32.2/32.2
EfficientNet-B0	4.0m	Spectrogram	[8]	71.1/63.8	43.5/40.7	48.0/44.0	53.8/49.2	34.5/26.2
CNN14	71m	Spectrogram	[11, 13]	74.6/66.4	56.0/37.3	56.4/44.8	62.3/48.9	37.8/28.8
ViT-Base	86m	Spectrogram	[12]	73.3/74.6	50.4/56.5	60.3/64.2	60.5/64.5	36.8/36.8
ResNet50	23m	Spectrogram	[19]	74.8/74.4	51.7/65.0	59.6/63.7	61.4/67.8	<u>38.4</u>/36.2
SF ResNet50	26m	Spectrogram	[17]	74.0/74.3	56.9/73.4	59.6/65.2	63.3/<u>71.7</u>	37.2/36.6
NFNet-F0	68m	Spectrogram	Ours	**76.1**/<u>76.0</u>	59.0/65.9	61.8/<u>65.5</u>	65.4/69.2	**39.3**/37.6
SF NFNet-F0	63m	Spectrogram	Ours	75.2/75.8	65.6/**77.2**	64.5/**68.6**	68.5/**74.6**	38.2/37.8

111.12

achieve state-of-the-art performance across all domains.

Index Terms— audio representations, representation evaluation, speech, music, acoustic scenes

supervised contrastive learning [8, 15], and comparing them across a large set of model architectures. We find that models trained with contrastive learning tend to generalize better in the speech and music domain, while performing comparably to supervised pretraining for environment sounds. We

Audio Slow-Fast

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman

- Project webpage: https://ekazakos.github.io/auditoryslowfast/
- Code & models: https://github.com/ekazakos/auditory-slow-fast

ICASSP 2023
4 - 10 JUNE,
RHODES ISLAND, GREECE

Play It Back: Iterative Attention for Audio Recognition

Alexandros Stergiou

Dima Damen

Motivation

with: Alexandros Stergiou

- Current Audio Recognition datasets contain examples of target classes intermixed with other irrelevant sounds

VGG-Sound

- Audio sources from YouTube videos
- Sounds emitted from human, animals, musical instruments, machinery or weather events

A Stergiou, D Damen (2023) Play It Back: Iterative Attention for Audio Recognition. ICASSP

Motivation

with: Alexandros Stergiou

- Current Audio-Recognition datasets contain examples of target classes intermixed with other irrelevant sounds

VGG-Sound

- Audio sources from YouTube videos
- Sounds emitted from human, animals, musical instruments, machinery or weather events

target class: *“ukulele”*

Motivation

- Current Audio-Recognition datasets contain examples of target classes intermixed with other irrelevant sounds

VGG-Sound

- Audio sources from YouTube videos
- Sounds emitted from human, animals, musical instruments, machinery or weather events

target class: *"people hiccup"*

Motivation

with: Alexandros Stergiou

- Current Audio-Recognition datasets contain examples of target classes intermixed with other irrelevant sounds

VGG-Sound

- Audio sources from YouTube videos
- Sounds emitted from human, animals, musical instruments, machinery or weather events

EPIC-KITCHENS

- Hand-Object Interaction Sounds
- Labelled with verb and noun classes

Motivation

with: Alexandros Stergiou

- Current Audio-Recognition datasets contain examples of target classes intermixed with other irrelevant sounds

VGG-Sound

- Audio sources from YouTube videos
- Sounds emitted from human, animals, musical instruments, machinery or weather events

EPIC-KITCHENS

- Hand-Object Interaction Sounds
- Labelled with verb and noun classes

target class: *"scrub plate"*

Motivation

with: Alexandros Stergiou

- Significant challenges to distinguish between similar sounds
  - Is it a Ukulele or Banjo?
  - Is it scrubbing a plate or a table?

Can you play it back?

Say again?

Motivation

with: Alexandros Stergiou

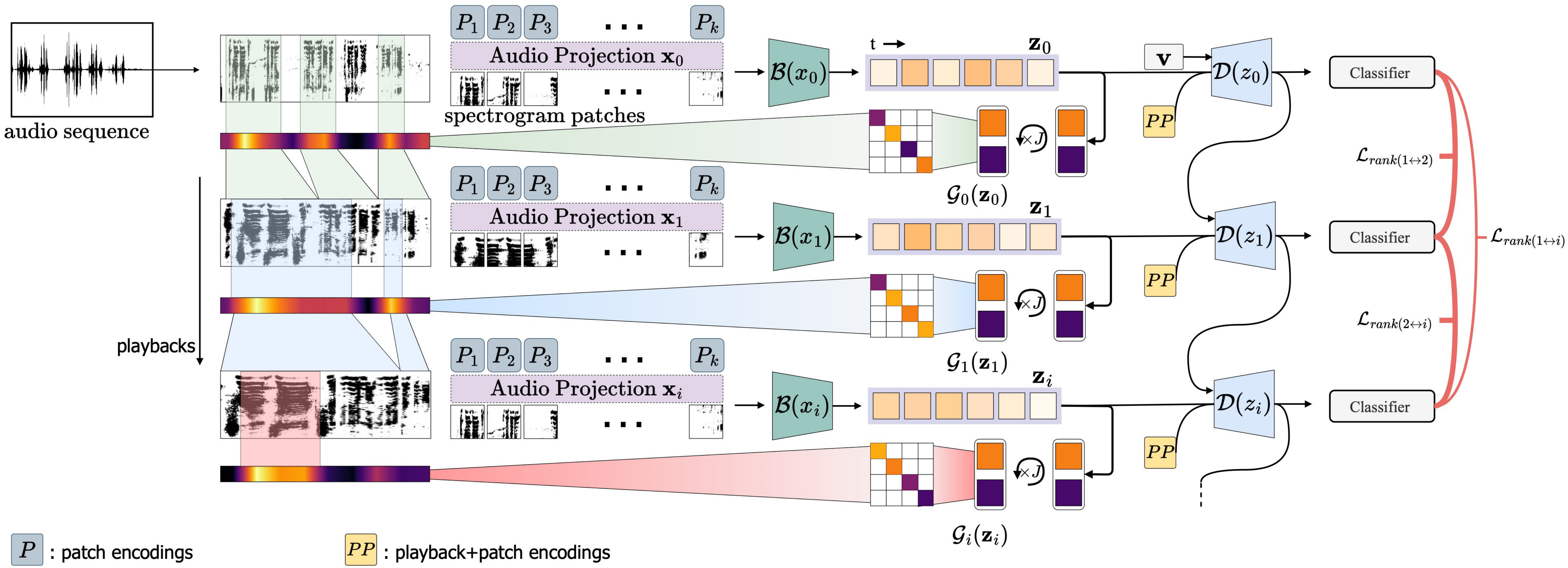
- Repeating sounds is essential for the development of **echoic memory**
- This memory is responsible for the memorisation of sounds [A
- Repeated listening to replays of sound stimulant ial for associating sound patterns [C]

How do we build an architecture that learns to repeat?

[A] Terry Clark, “Ec xplored and applied,” Journal of services marketing, 1987.
[B] Rael D St owan, Walter Ritter, and Daniel C Javitt, “Auditory sensory (” echoic”) me on in schizophrenia.,” The American journal of psychiatry, 1995.
[C] Ga Radvansky, Human Memory, Psychology Press, 2005.

Play-It-Back Architecture

with: Alexandros Stergiou



Results

with: Alexandros Stergiou

Model	GFLOPs	verb		noun		action	
		top-1	top-5	top-1	top-5	top-1	top5
Damen et al. [8]	N/A	42.6	75.8	22.3	44.6	14.5	28.2
MBT (A) [18]	34.2	44.3	-	22.4	-	13.0	-
Slow-Fast [12]	35.1	46.5	78.3	22.8	44.9	15.4	28.6
PlayItBackX3	122.8	**47.0**	**78.7**	**23.1**	**45.1**	**15.9**	**29.2**

Table 3: Comparisons to state-of-the-art for EPIC-KITCHENS-100. We report the top-1 and top-5 accuracies for the verb, noun, and action labels.

Performance over playbacks

with: Alexandros Stergiou

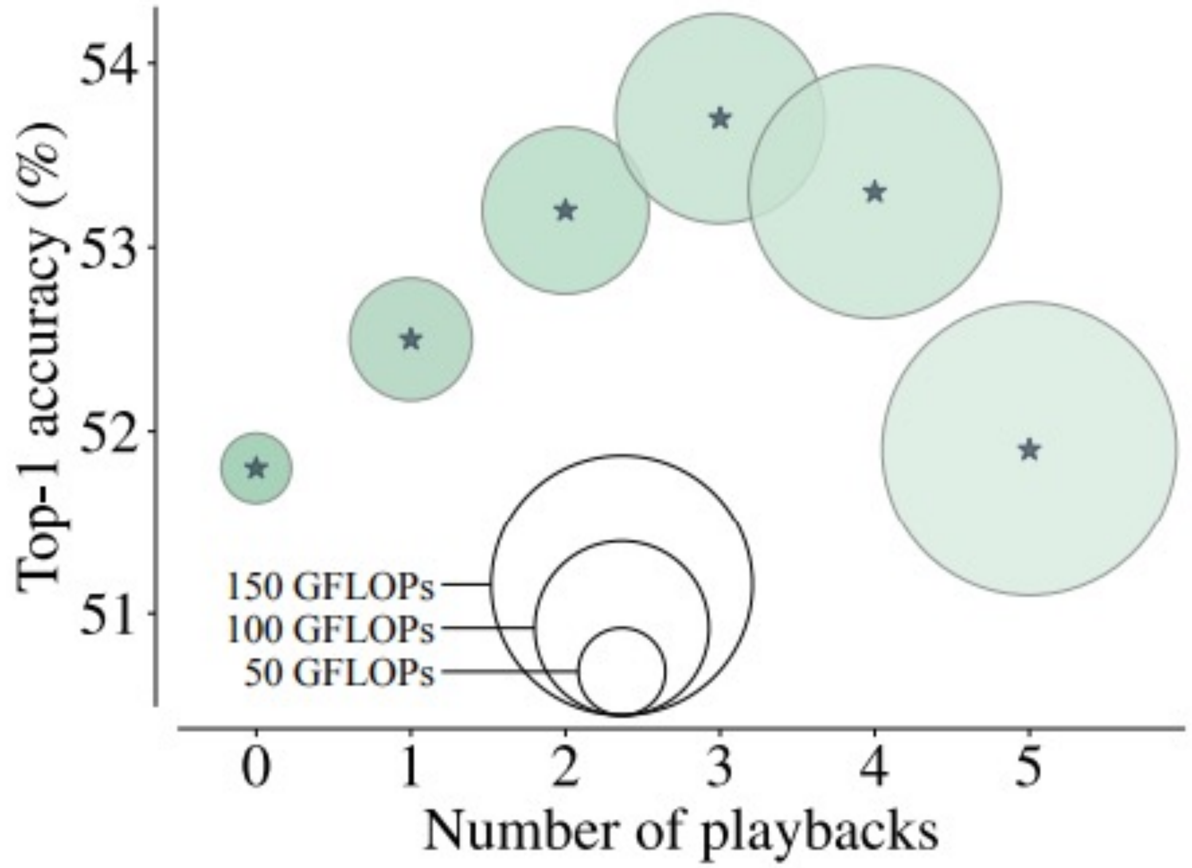
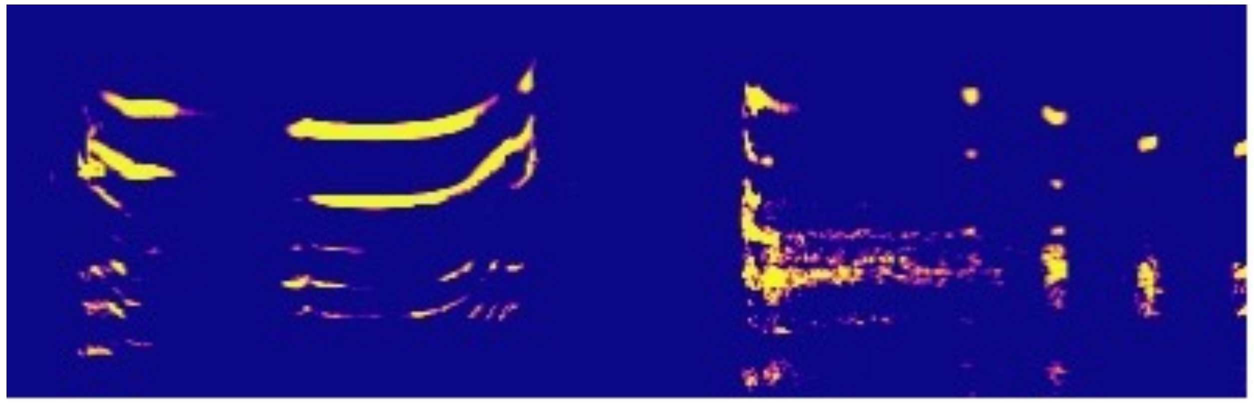


Fig. 3: VGG-Sound top-1 accuracy over different playback-numbers (N) with respect to the compute (in GFLOPs).

A Stergiou, D Damen (2023) Play It Back: Iterative Attention for Audio Recognition. ICASSP

Qualitative Results

with: Alexandros Stergiou



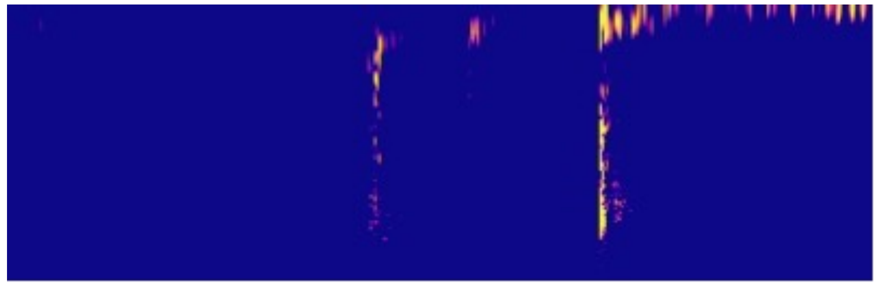
X0 → X3

GT: **people giggling**
PlayItBackX0: people screaming

GT: **people giggling**
PlayItBackX3: **people giggling**

Qualitative Results

with: Alexandros Stergiou



GT: **close fridge**

PlayItBackX0: open drawer

X0 → X3

GT: **close fridge**

PlayItBackX3: **close fridge**

Project website

Github code

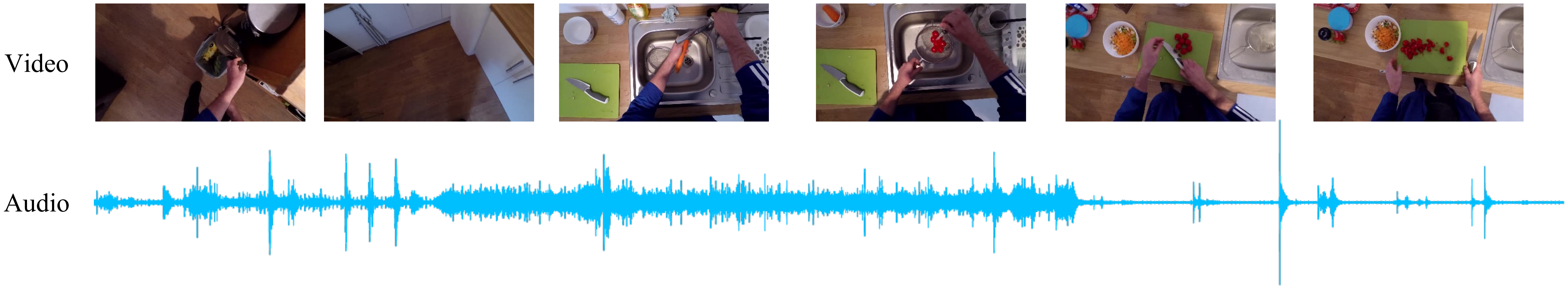
ICASSP 2023
4 - 10 JUNE, RHODES ISLAND, GREECE

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

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

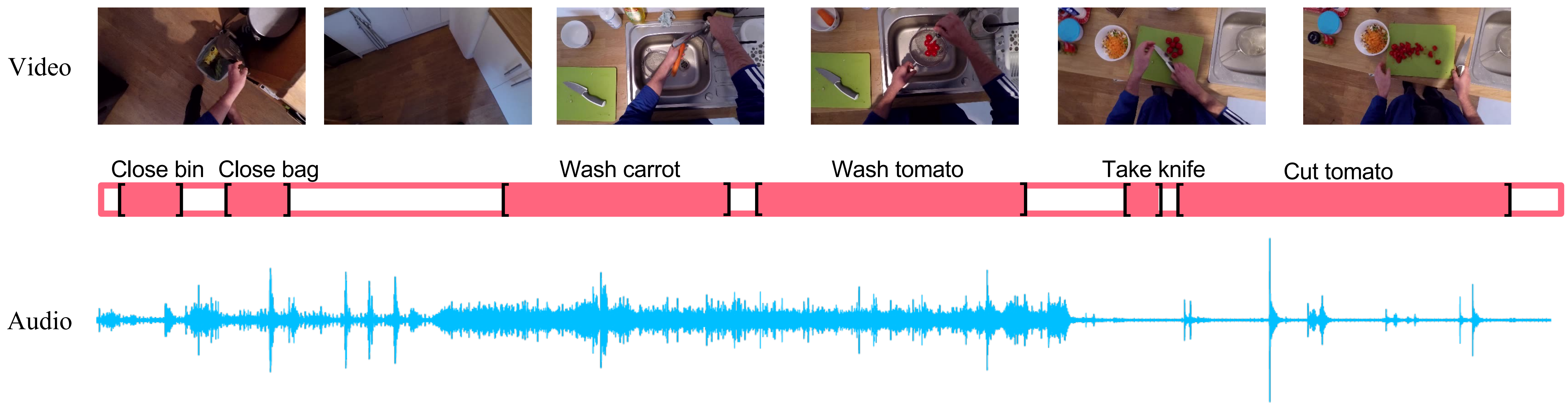
Motivation

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



Motivation

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

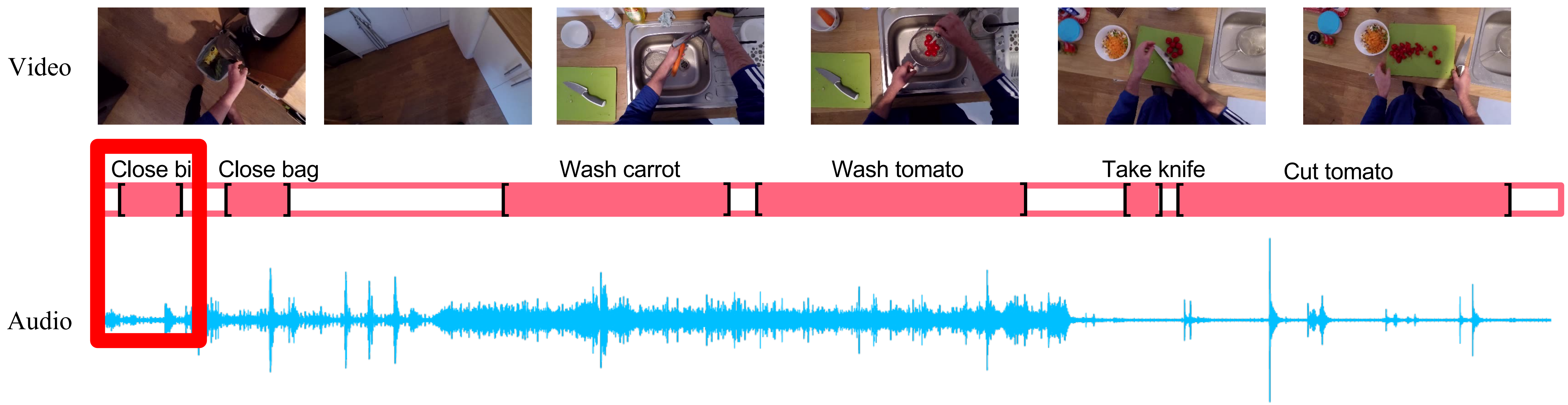


Motivation

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

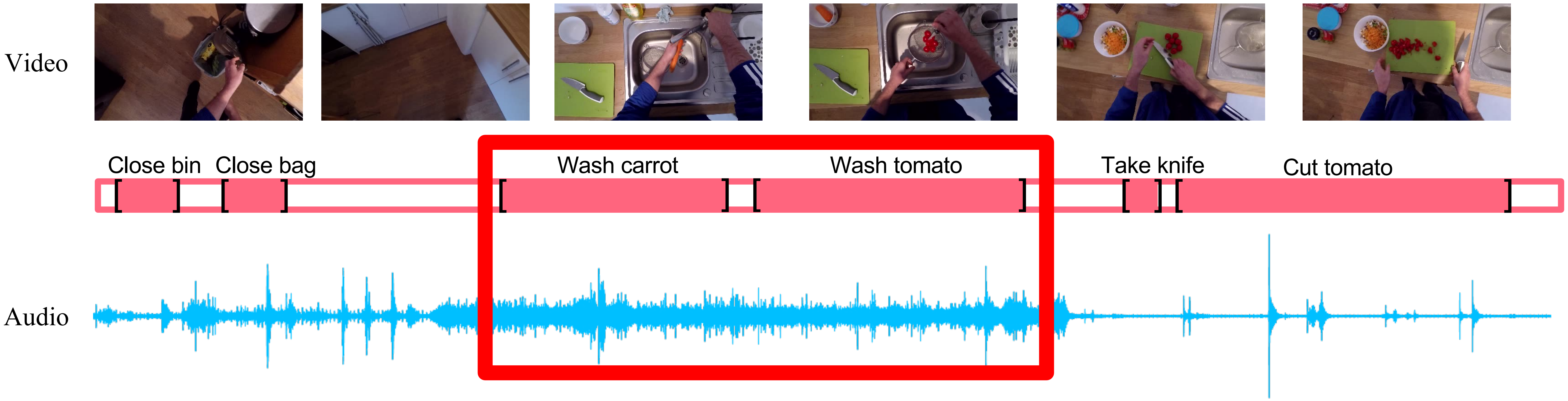
Motivation

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



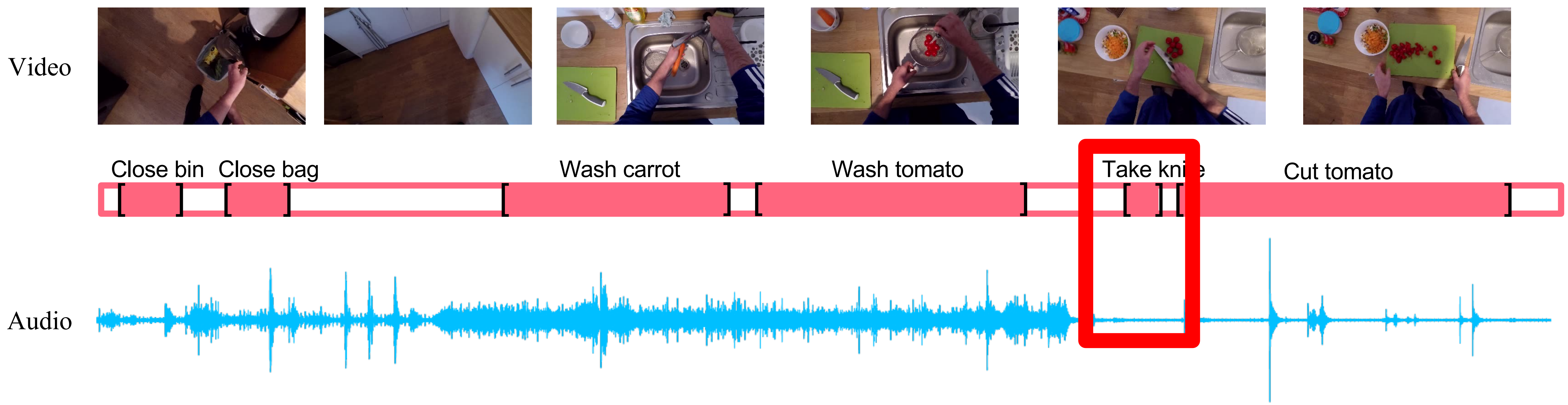
Motivation

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



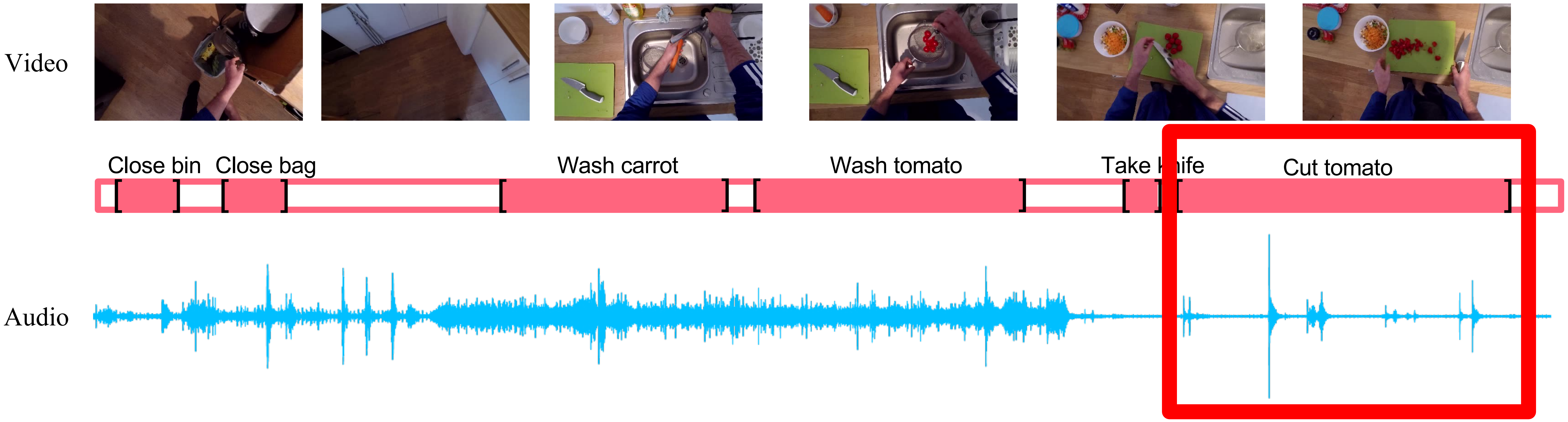
Motivation

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



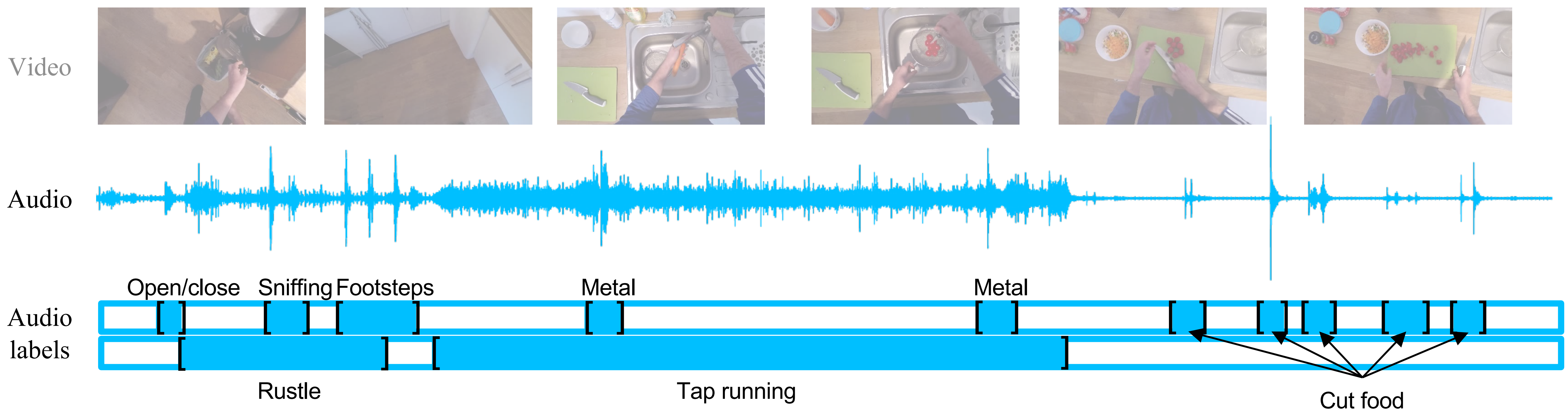
Motivation

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



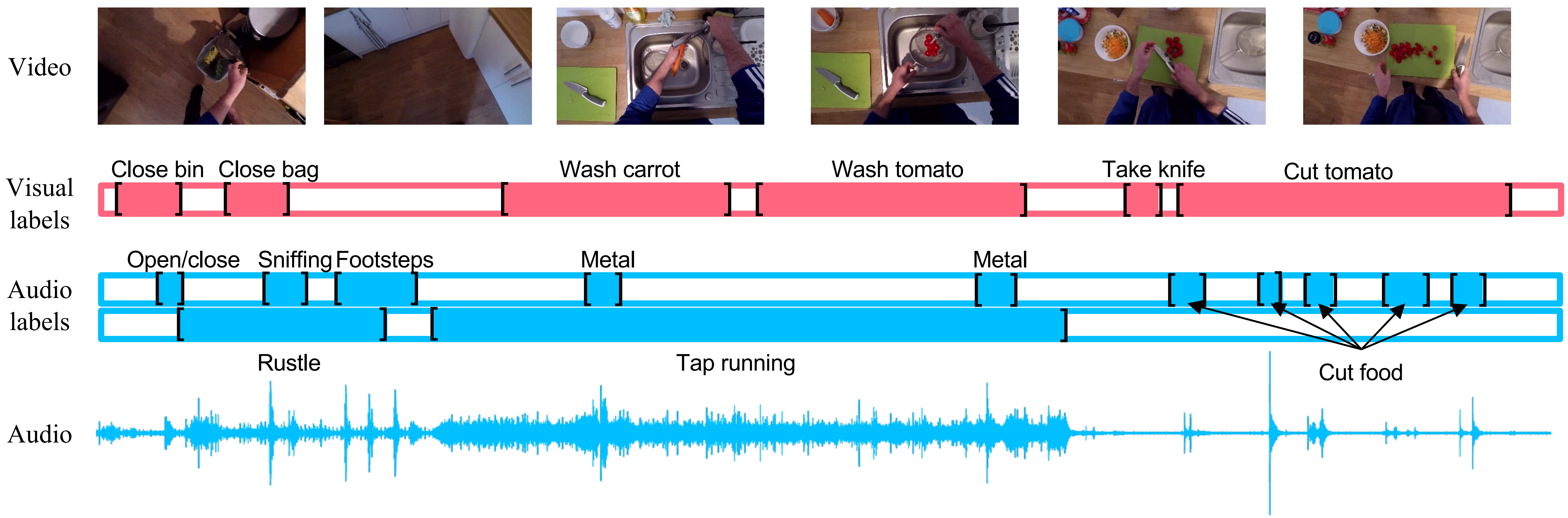
Motivation

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

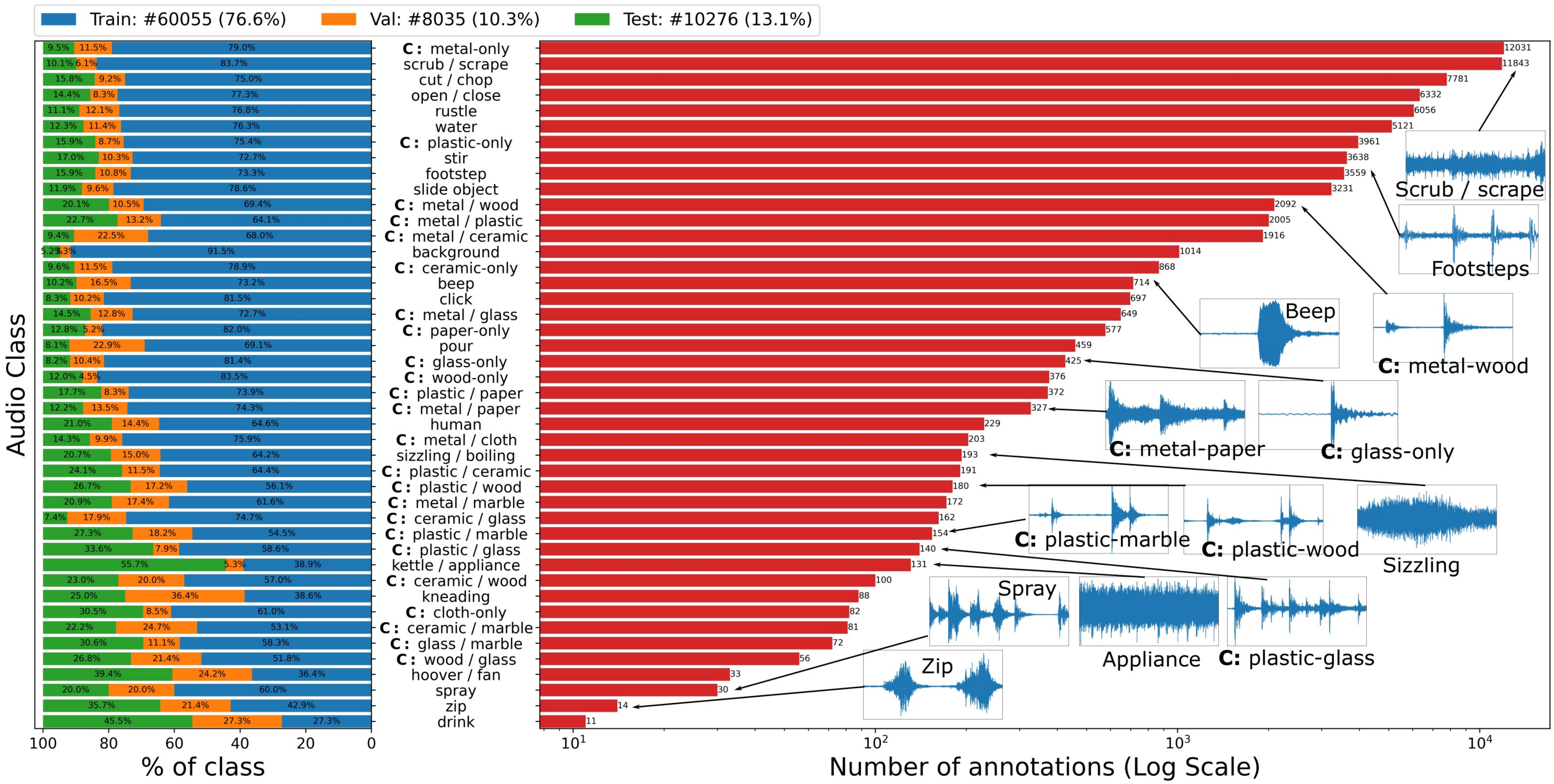
Visual Action Annotations
90K visual actions
97 verb classes
300 noun classes

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



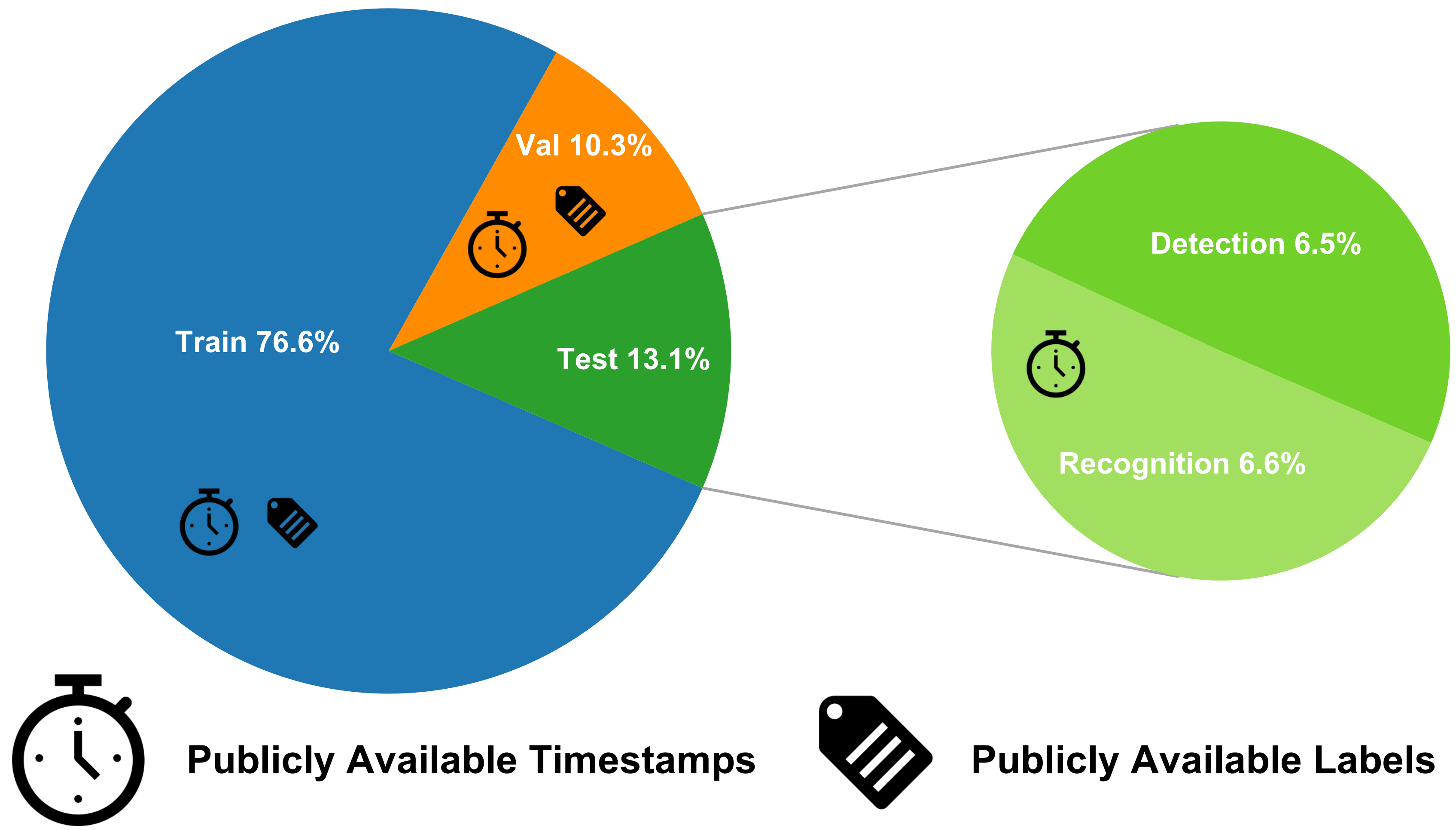
BRISTOL

MULA@CVPR2024

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

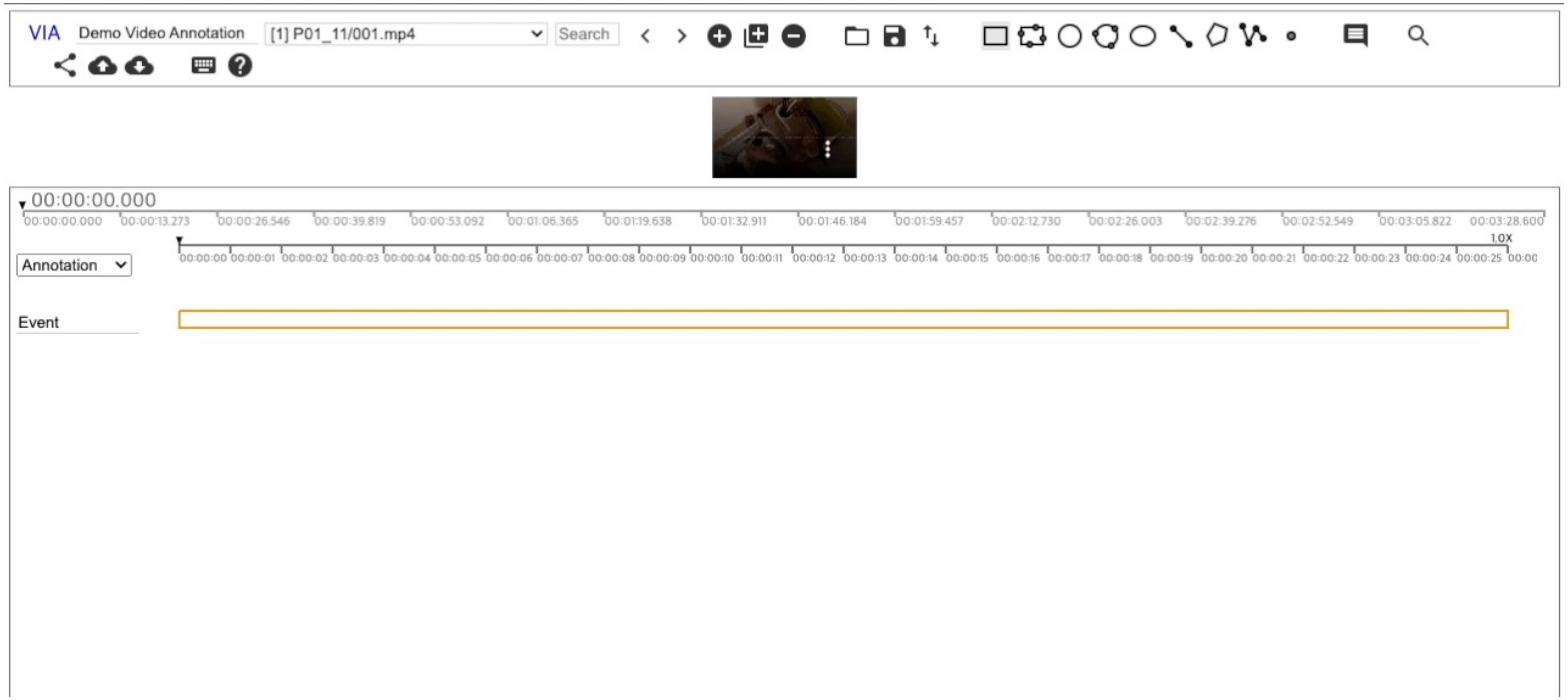


J Huh*, J Chalk* E Kazakos, D Damen, A Zisserman (2023). EPIC-Sounds: A Large-scale Dataset of Actions That Sound. ICASSP

Annotations Pipeline

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

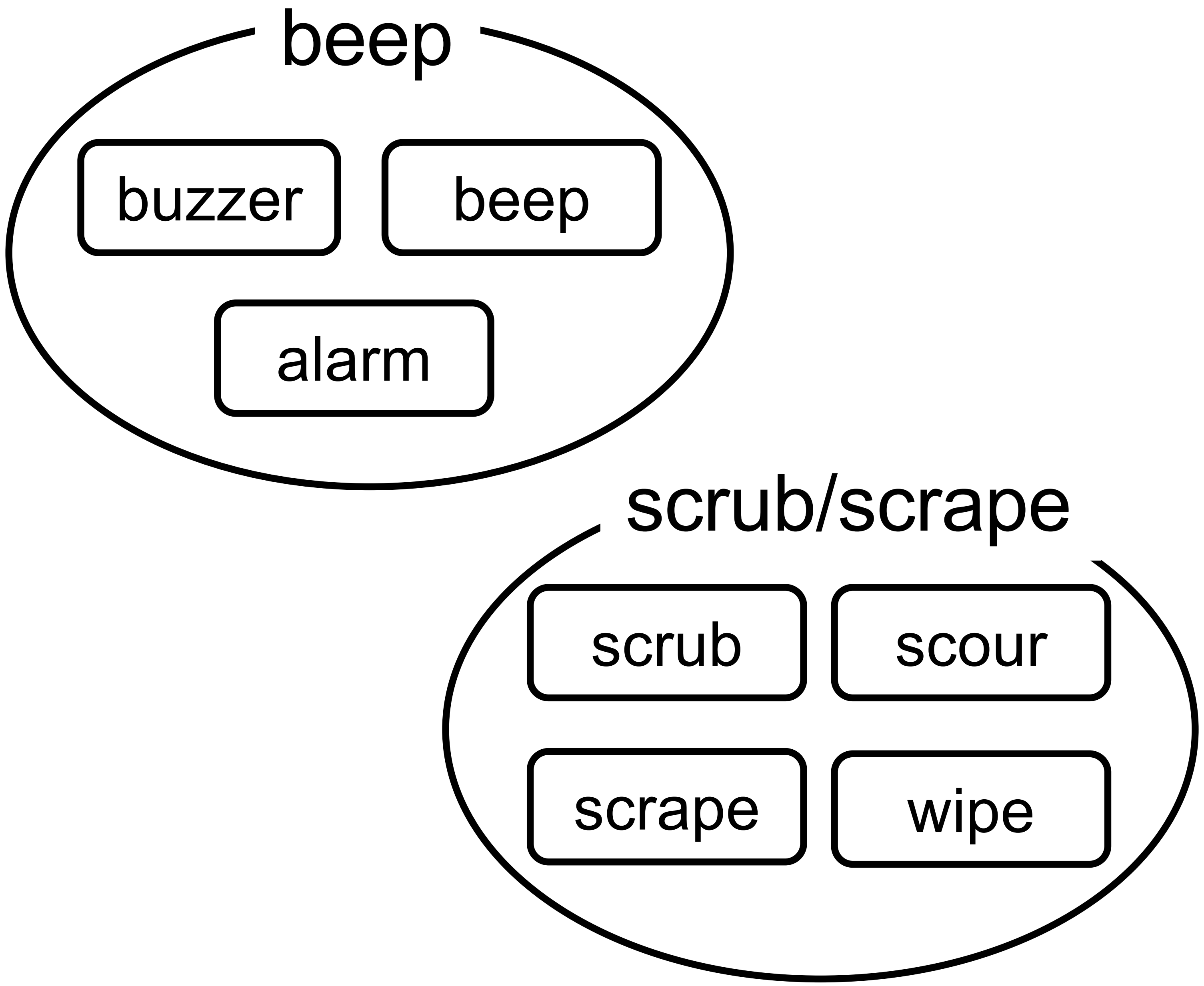
- We annotate all the distinctive sound events which consist of temporal intervals using free-form sound descriptions.
- Using VGG Image annotator tool



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

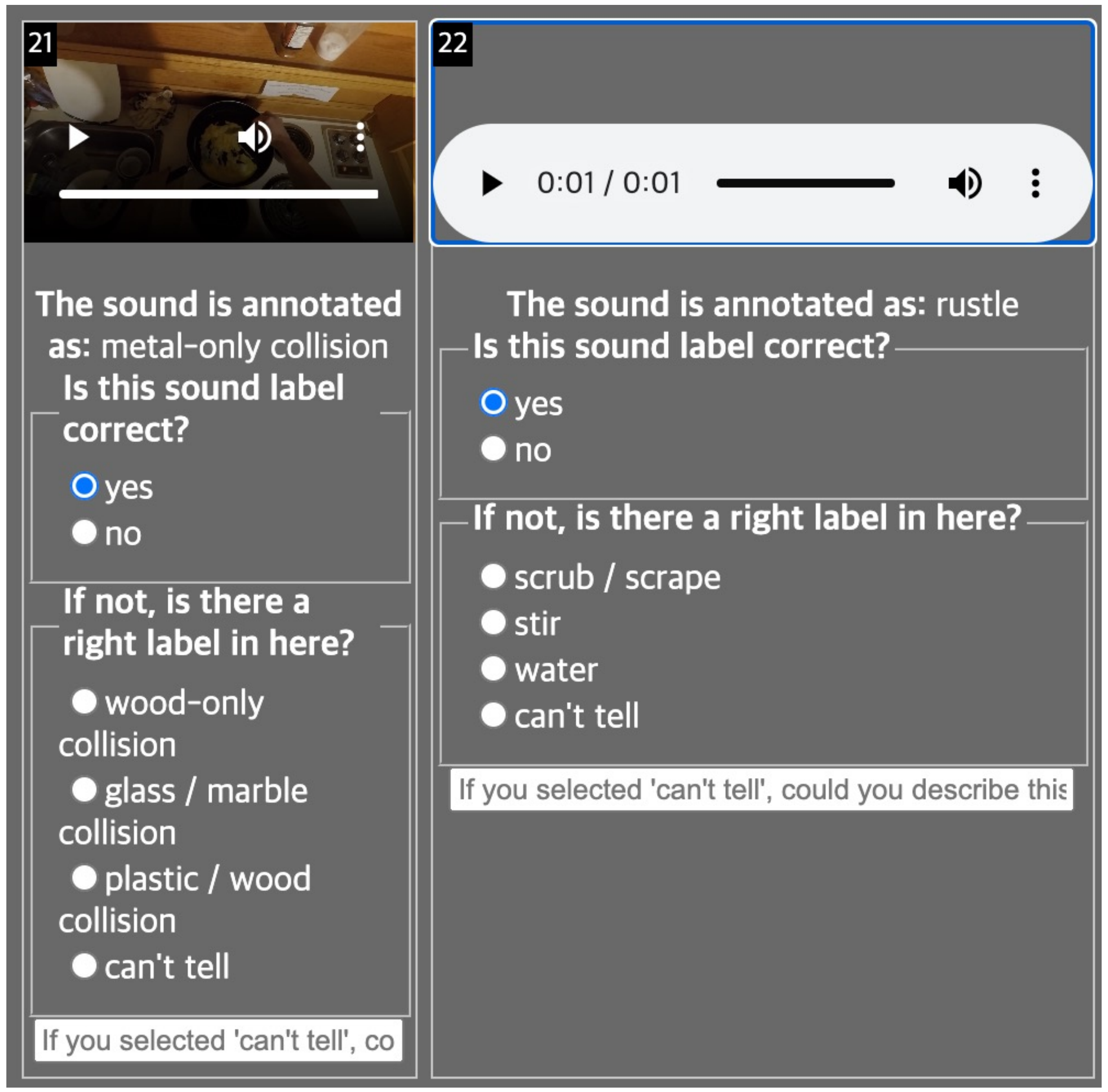
Glass

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

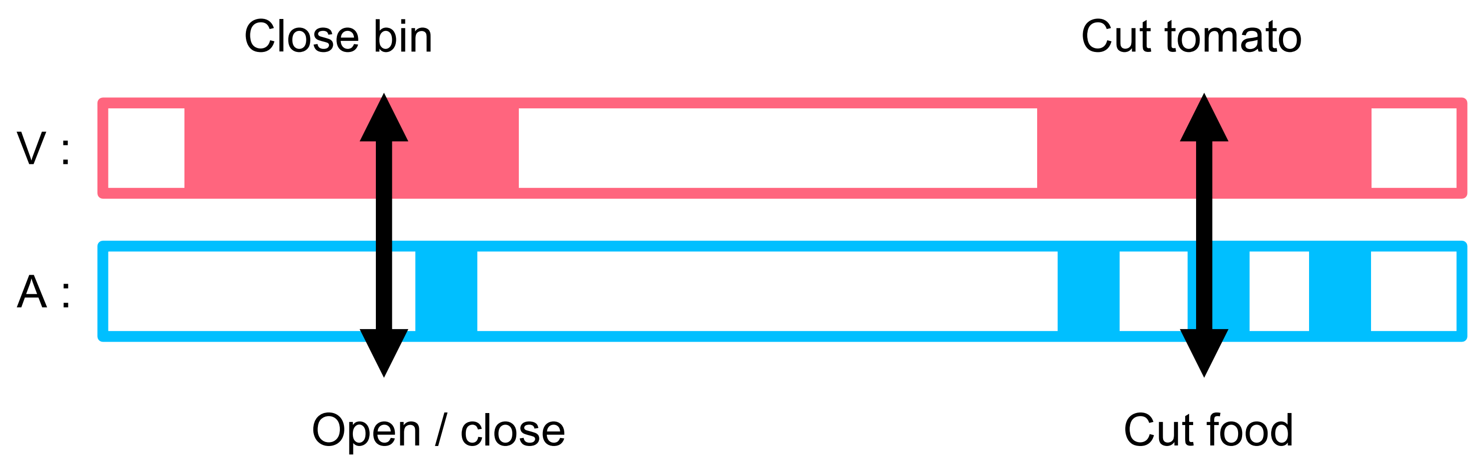
Post Processing

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

- Manual check on validation / test set



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



Non-categorized audio events

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

- There are around 39,000 audio events that we recognise the sound exists but no semantic label matching the 44 classes could be given.
- Because
  - Unable to assign the label
  - Collision sounds for which they could not be visually verified.
- We also released them in our website

Baselines and Model Checkpoints

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

Table 3: Results of the Baseline Models on the EPIC-SOUNDS validation and test splits. L: Linear-Probe; F: Fine-Tuning.

Split	Model		Top-1	Top-5	mCA	mAP	mAUC
Val	Chance	-	7.71	30.95	2.29	0.023	0.500
	SSAST [28]	L	28.74	64.87	7.14	0.079	0.755
	ASF [29]	L	45.53	79.33	13.48	0.172	0.789
	SSAST [28]	F	53.47	**84.56**	**20.22**	0.235	**0.879**
	ASF [29]	F	**53.75**	84.54	20.11	**0.254**	0.873
Test	Chance	-	7.22	30.11	2.27	0.023	0.500
	SSAST [28]	L	27.50	65.55	6.68	0.080	0.741
	ASF [29]	L	44.55	78.44	14.49	0.145	0.772
	SSAST [28]	F	53.75	83.76	**20.76**	**0.237**	**0.860**
	ASF [29]	F	**54.86**	**84.26**	20.30	0.232	0.823

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:

Dataset and Challenge

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

- Yesterday, the second EPIC-SOUDS Recognition Challenge and first EPIC-SOUNDS Detection Challenge

Test set												
#	User	Entries	Date of Last Entry	SLS			Training Modality	Top-1 Accuracy (%)	Top-5 Accuracy (%)	Per-Class Accuracy (%)	Mean Average Precision (%)	Mean Area Under Curve
				PT ▲	TL ▲	TD ▲	T_MOD ▲	Interaction ▲	Interaction ▲	Avg. ▲	Avg. ▲	Avg. ▲
1	**JMCarrot**	11	05/31/24	2.0 (1)	3.0 (1)	4.0 (1)	0.0 (2)	56.57 (1)	86.30 (1)	22.24 (3)	27.98 (2)	0.883 (2)
2	**TIM_method**	1	04/06/24	2.0 (1)	2.0 (2)	3.0 (2)	2.0 (1)	55.86 (2)	86.26 (2)	22.97 (1)	32.23 (1)	0.894 (1)
3	**CVCV**	6	05/31/24	2.0 (1)	3.0 (1)	4.0 (1)	0.0 (2)	55.66 (3)	85.93 (3)	21.69 (5)	27.89 (3)	0.878 (3)
4	stevenlau	6	05/31/23	2.0 (1)	3.0 (1)	4.0 (1)	0.0 (2)	55.43 (4)	85.52 (4)	21.84 (4)	26.98 (5)	0.877 (4)
5	Yuqi_Li	10	05/31/24	2.0 (1)	3.0 (1)	4.0 (1)	0.0 (2)	55.17 (5)	85.34 (6)	20.98 (9)	26.15 (6)	0.861 (6)
6	audi666	3	06/01/23	2.0 (1)	3.0 (1)	4.0 (1)	0.0 (2)	55.11 (6)	85.40 (5)	21.14 (8)	25.96 (9)	0.856 (7)
7	EPIC_AUDITORY_SLOWFAST	1	01/25/23	2.0 (1)	3.0 (1)	3.0 (2)	0.0 (2)	54.80 (7)	85.18 (8)	20.77 (10)	26.01 (8)	0.850 (10)
8	DXLong	6	05/31/24	2.0 (1)	3.0 (1)	3.0 (2)	0.0 (2)	54.78 (8)	85.40 (5)	21.43 (6)	26.06 (7)	0.854 (8)
9	WJB	8	05/25/24	2.0 (1)	3.0 (1)	4.0 (1)	0.0 (2)	54.50 (9)	85.32 (7)	21.40 (7)	27.41 (4)	0.876 (5)
9	WJB	8	05/25/24	(1)	(1)	(1)	0.0 (2)	54.50 (9)	85.32 (7)	21.40 (7)	27.41 (4)	0.876 (5)

Test set													
#	User	Entries	Date of Last Entry	SLS			Training Modality	mAP@0.1 (%)	mAP@0.2 (%)	mAP@0.3 (%)	mAP@0.4 (%)	mAP@0.5 (%)	Avg. mAP (%)
				PT ▲	TL ▲	TD ▲	T_MOD ▲	Interaction ▲	Interaction ▲	Interaction ▲	Interaction ▲	Interaction ▲	Interaction ▲
1	**shuming**	3	05/31/24	2.0 (1)	3.0 (1)	4.0 (1)	0.0 (2)	19.81 (1)	17.24 (1)	14.82 (1)	12.48 (1)	9.74 (1)	14.82 (1)
2	**TIM_method**	1	04/06/24	2.0 (1)	3.0 (1)	3.0 (2)	2.0 (1)	15.71 (2)	13.27 (2)	11.36 (2)	9.34 (2)	7.30 (2)	11.40 (2)
3	**AABC**	11	05/30/24	0.0 (2)	3.0 (1)	3.0 (2)	0.0 (2)	11.22 (3)	9.75 (3)	8.55 (3)	7.13 (3)	5.66 (3)	8.46 (3)
4	Yuqi_Li	2	05/28/24	2.0 (1)	3.0 (1)	4.0 (1)	0.0 (2)	9.69 (4)	8.75 (4)	7.58 (4)	6.57 (4)	5.32 (4)	7.58 (4)
5	EPIC_ACTIONFORMER	1	03/01/24	2.0 (1)	3.0 (1)	3.0 (2)	0.0 (2)	9.57 (6)	8.51 (5)	7.38 (5)	6.22 (5)	5.05 (5)	7.35 (5)
6	CVCV	8	05/31/24	2.0 (1)	3.0 (1)	3.0 (2)	0.0 (2)	9.61 (5)	8.40 (6)	7.25 (6)	5.98 (6)	4.45 (6)	7.14 (6)
7	fly_to	3	05/31/24	2.0 (1)	3.0 (1)	3.0 (2)	0.0 (2)	9.36 (7)	8.25 (7)	6.97 (7)	5.77 (7)	4.37 (7)	6.94 (7)

CVPR
JUNE 17-21, 2024
SEATTLE, WA

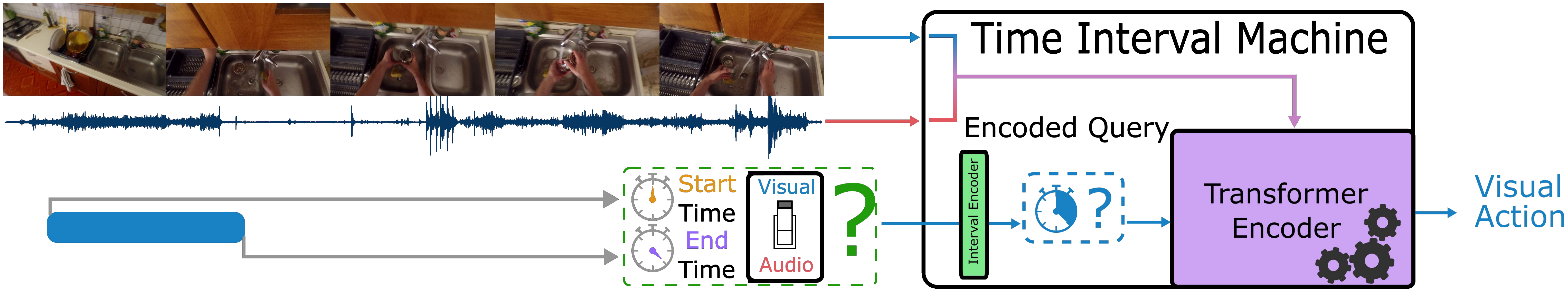
TIM: A Time Interval Machine for Audio-Visual Action Recognition

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

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

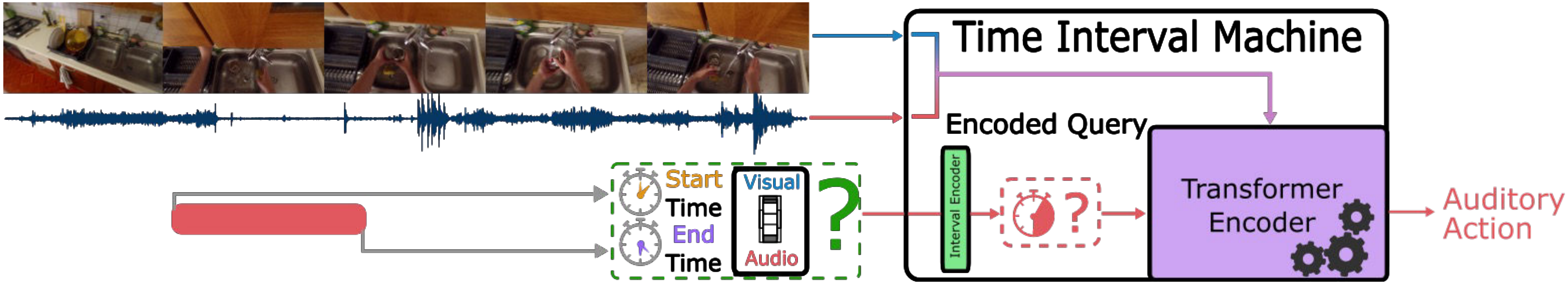
TIM: A Time-Interval Audio-Visual Machine

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



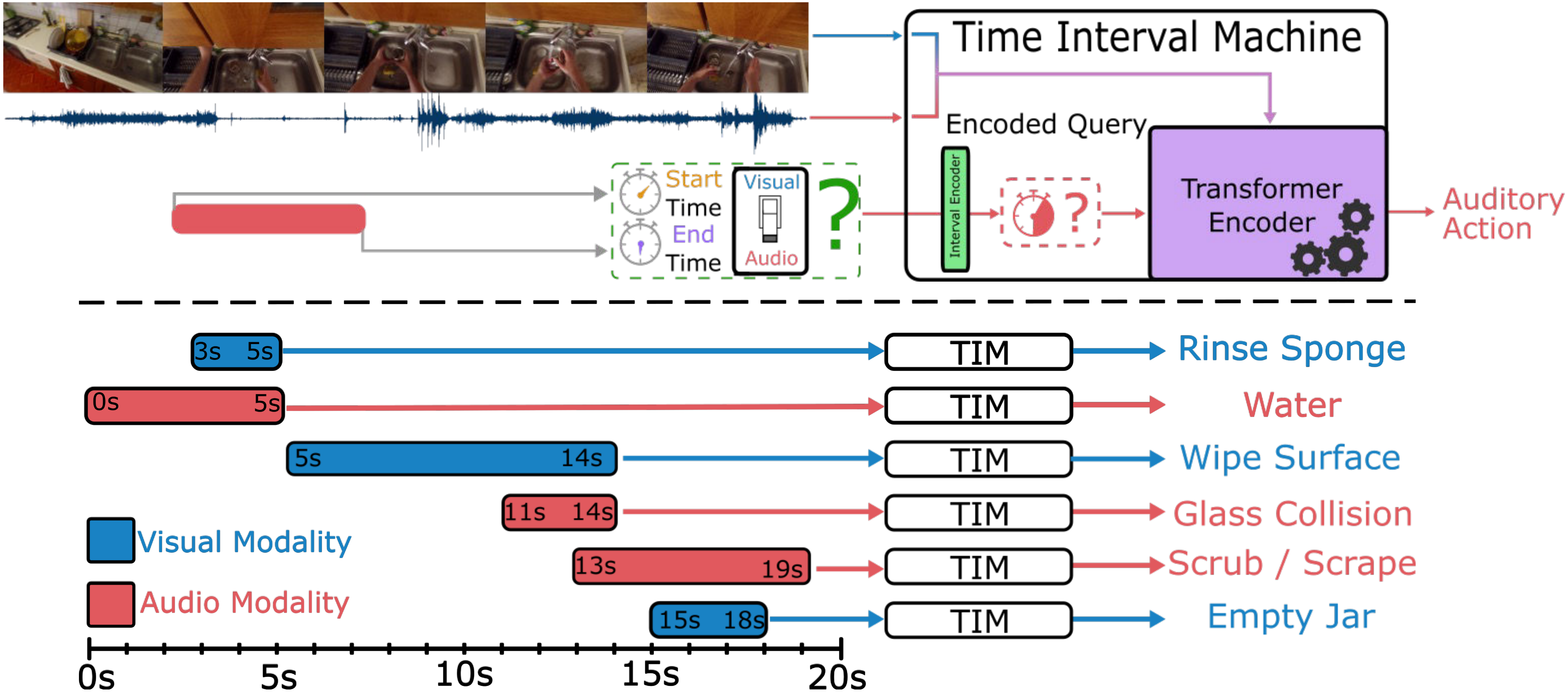
TIM: A Time-Interval Audio-Visual Machine

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

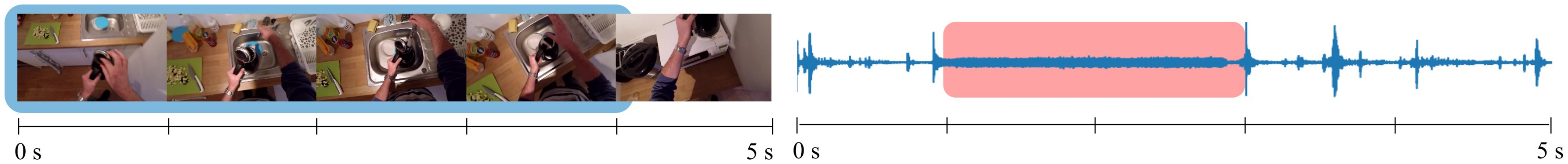


University of BRISTOL

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

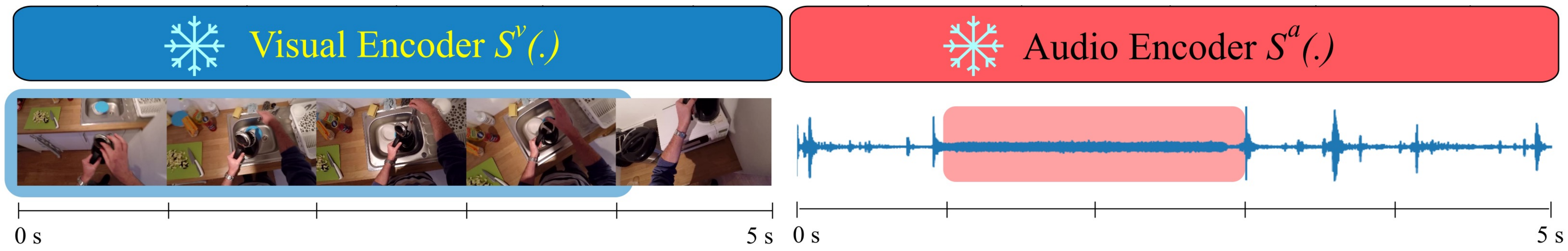
TIM: A Time-Interval Audio-Visual Machine

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

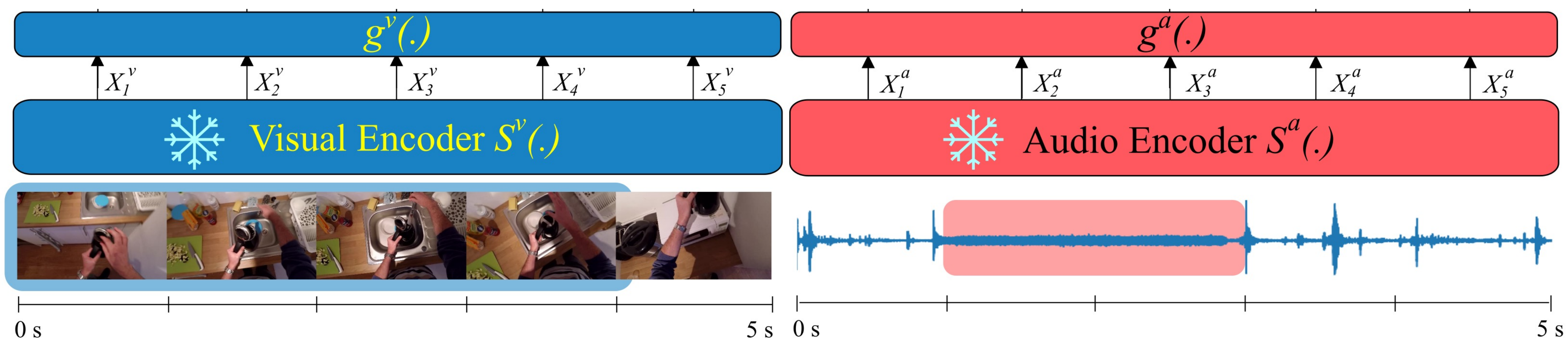


University of BRISTOL

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

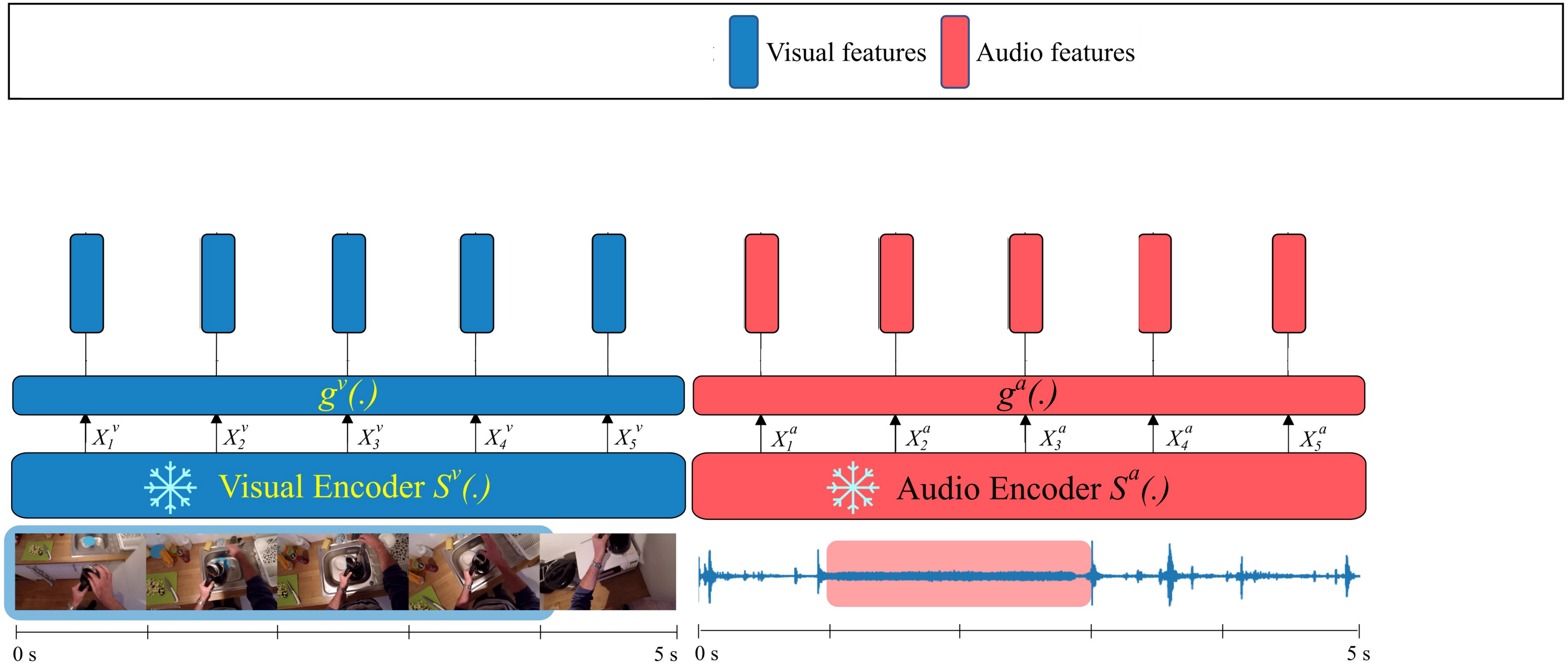
TIM: A Time-Interval Audio-Visual Machine

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



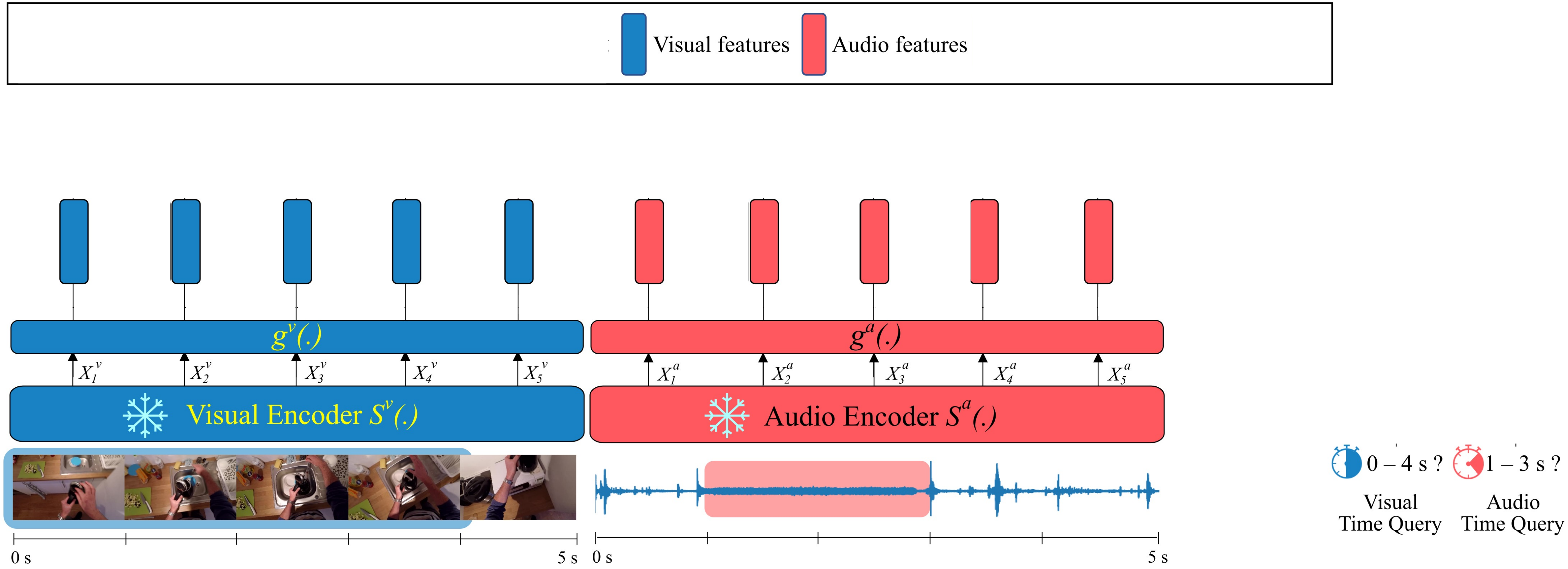
TIM: A Time-Interval Audio-Visual Machine

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



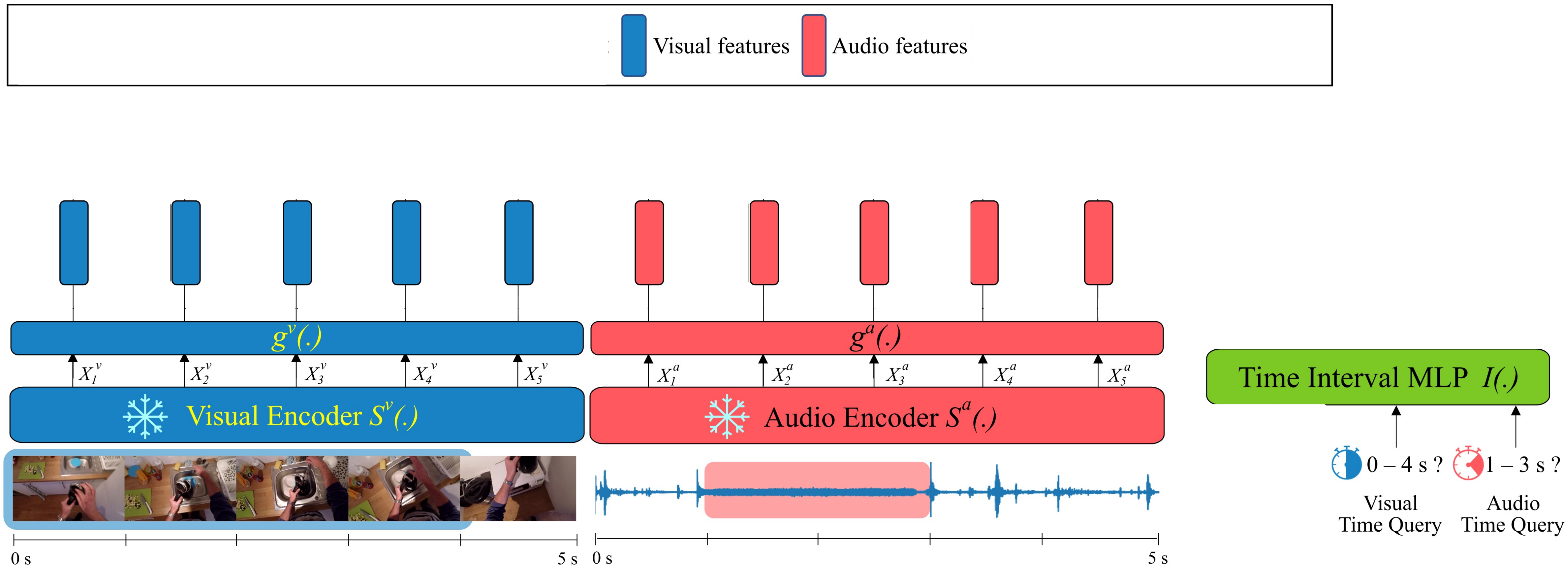
TIM: A Time-Interval Audio-Visual Machine

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



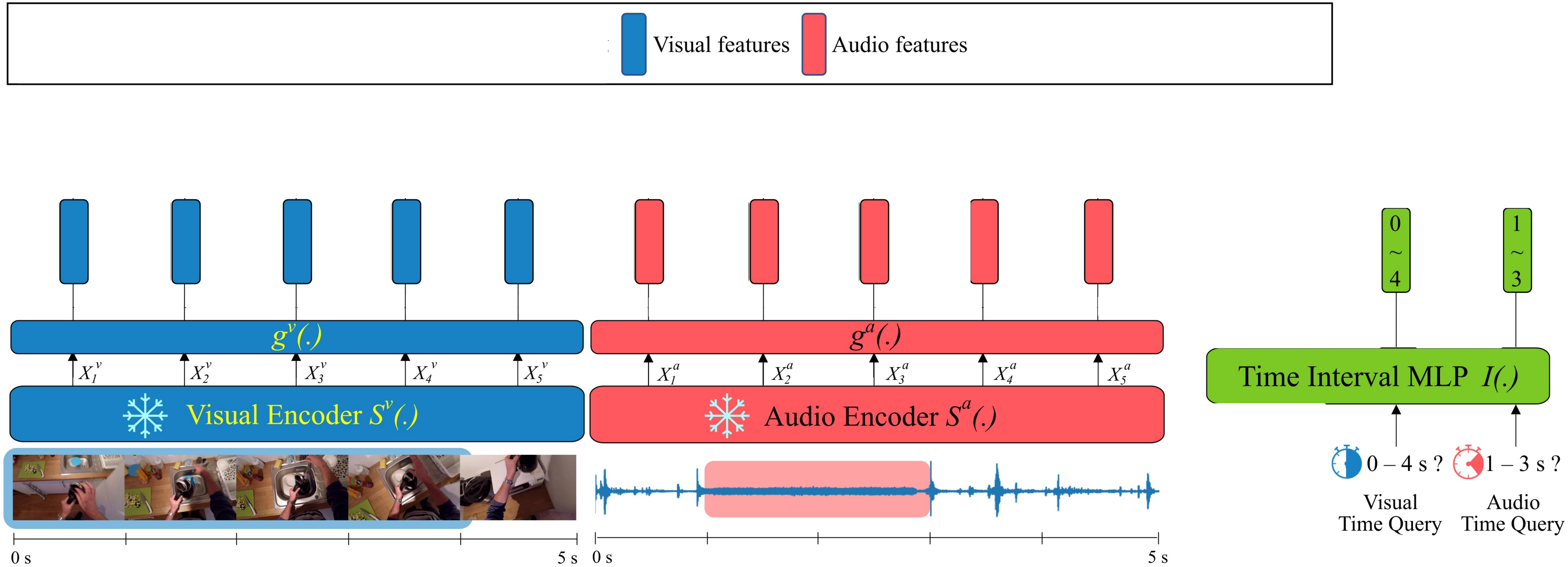
TIM: A Time-Interval Audio-Visual Machine

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



TIM: A Time-Interval Audio-Visual Machine

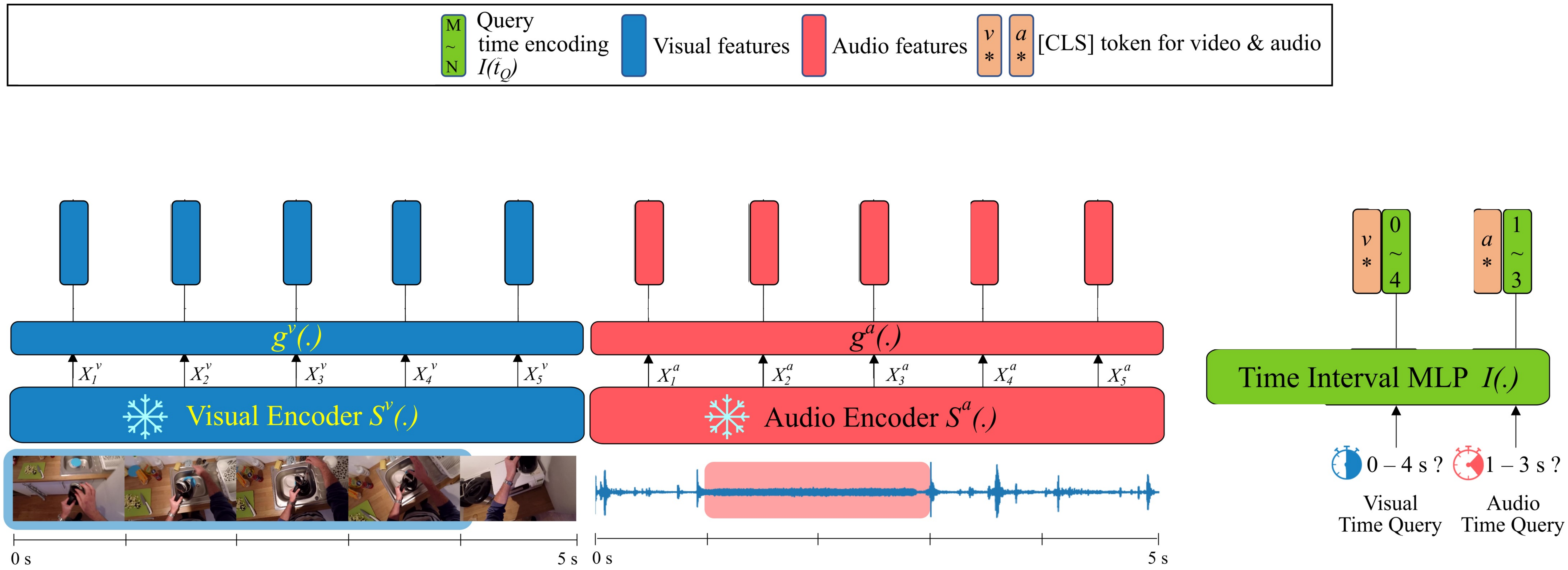
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



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

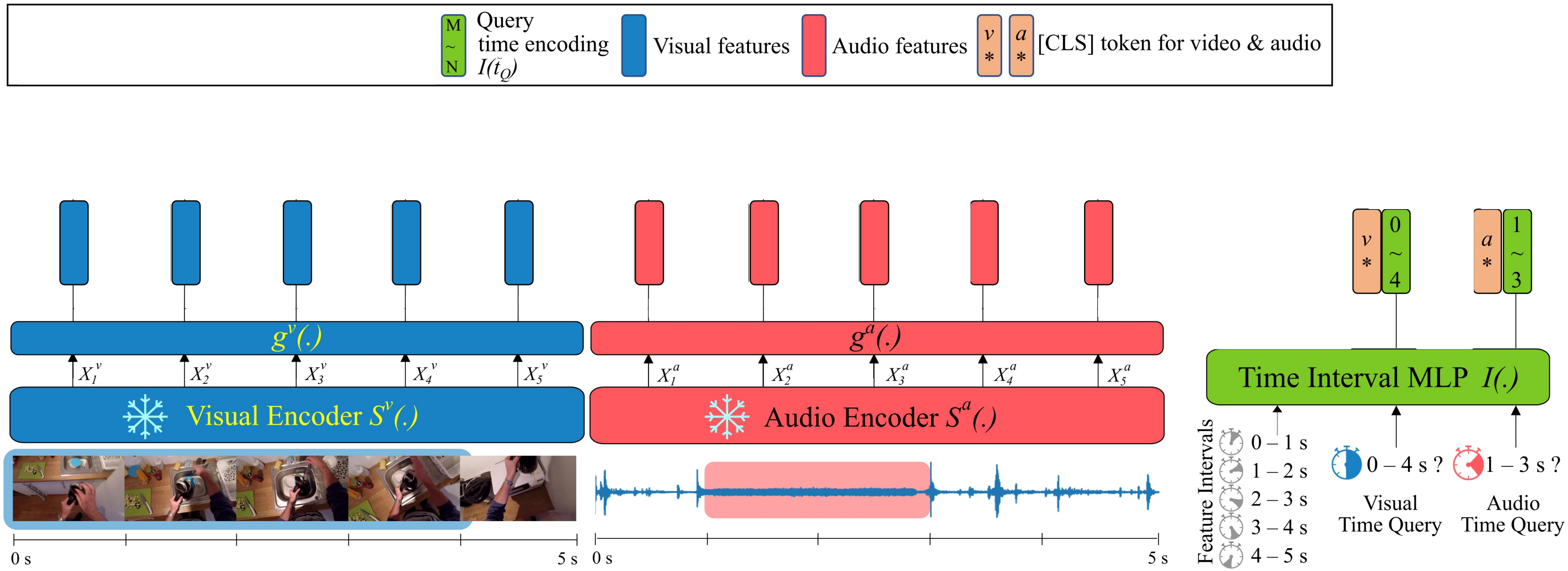
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



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

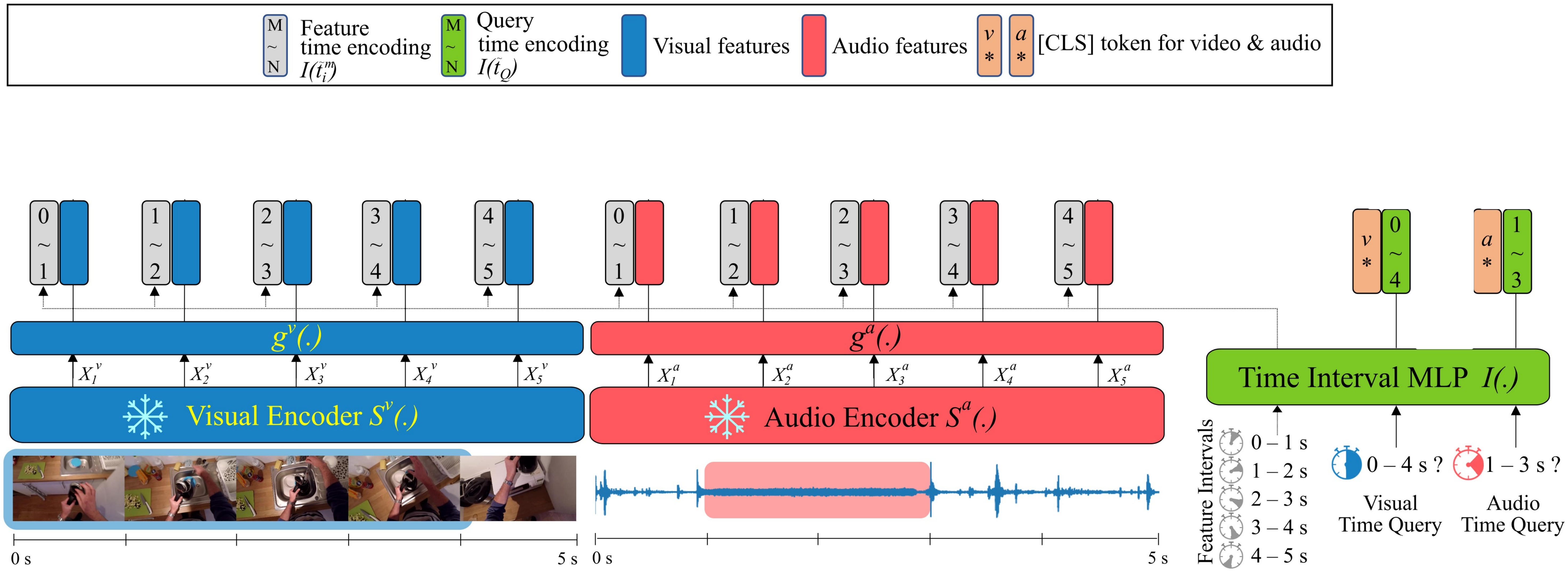
TIM: A Time-Interval Audio-Visual Machine

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



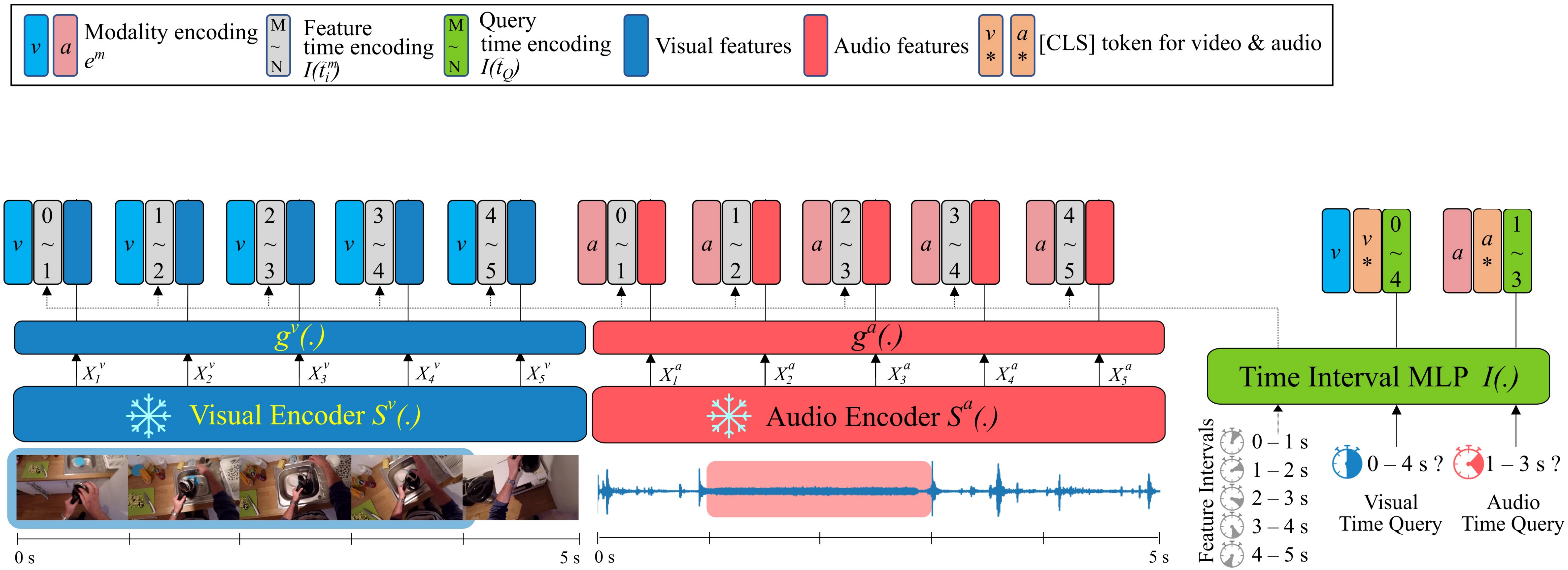
TIM: A Time-Interval Audio-Visual Machine

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



TIM: A Time-Interval Audio-Visual Machine

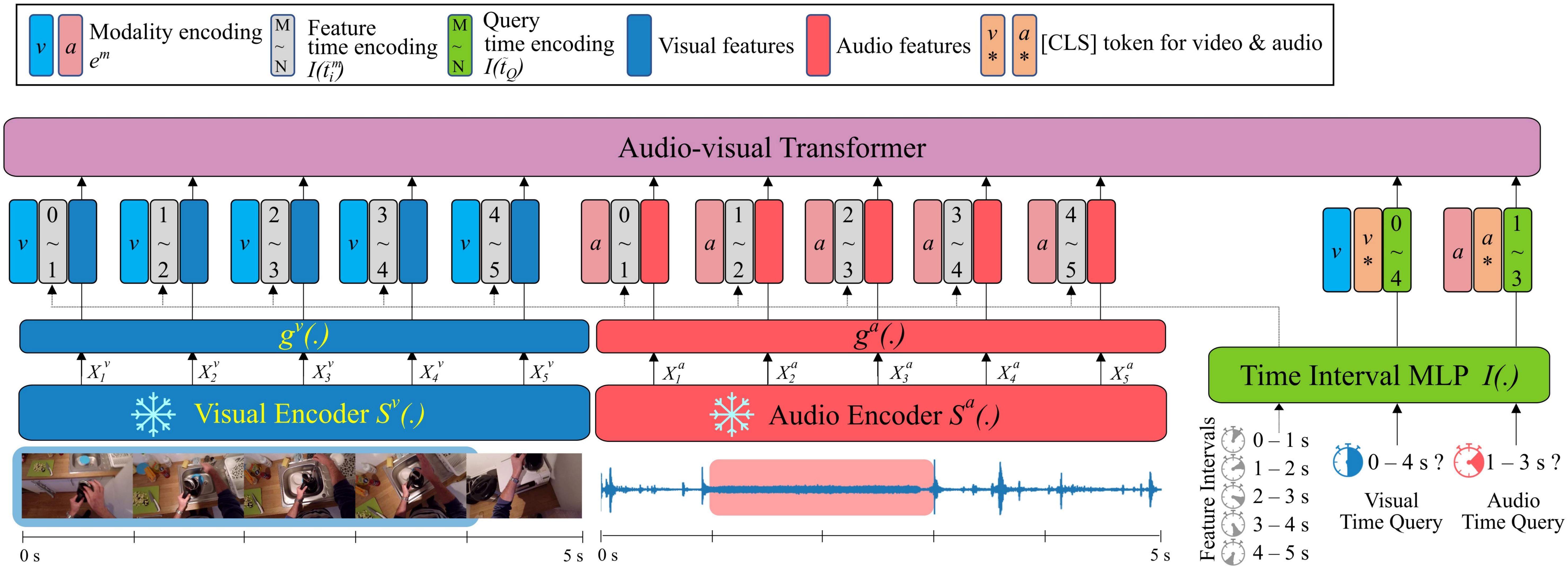
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman



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

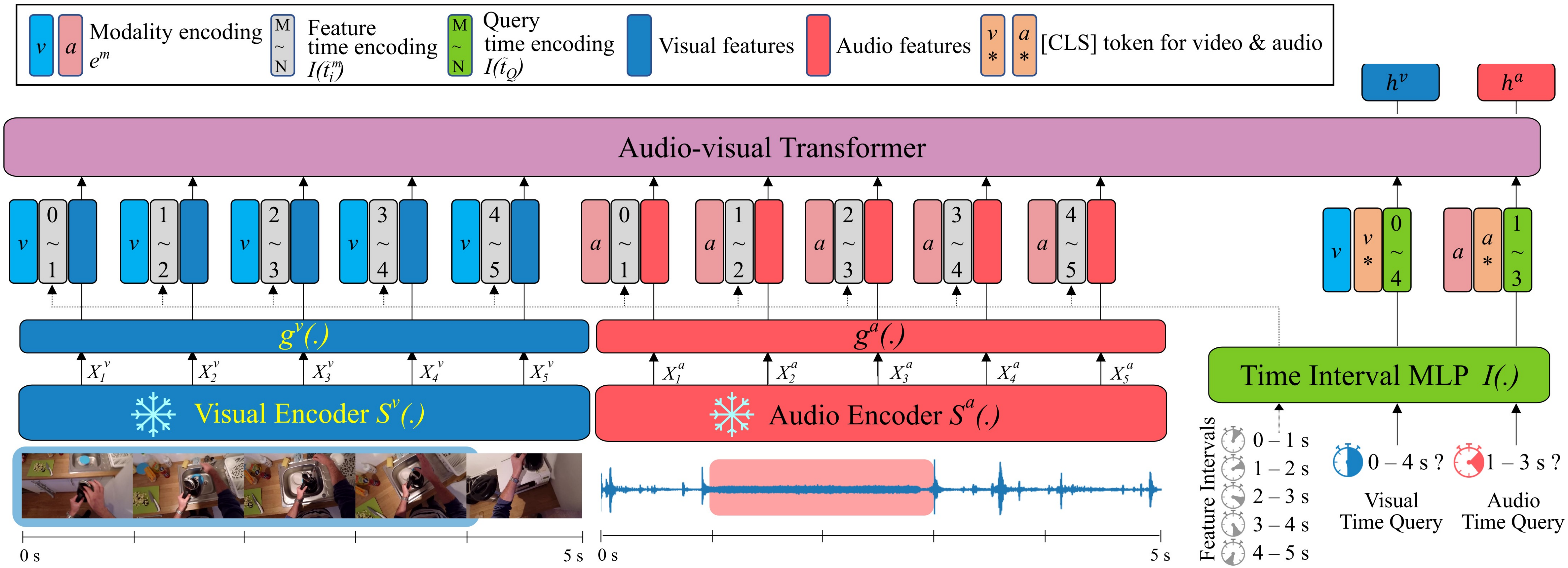


University of BRISTOL

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

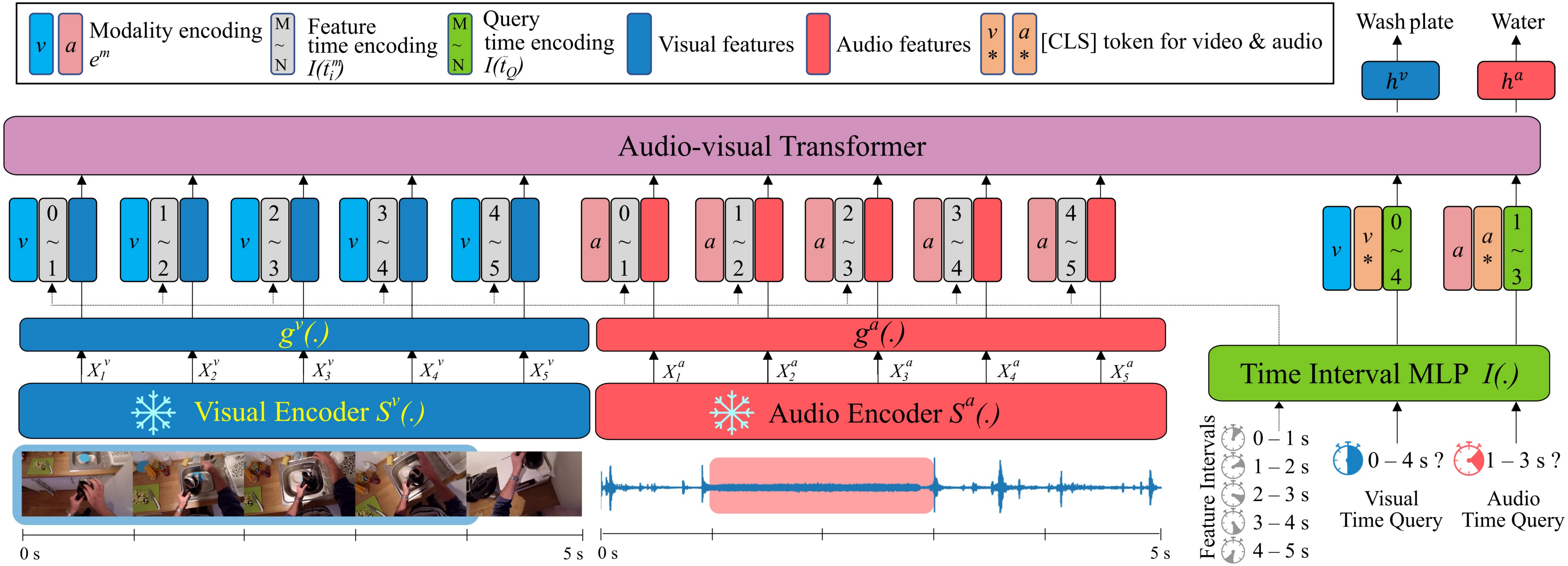


University of BRISTOL

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

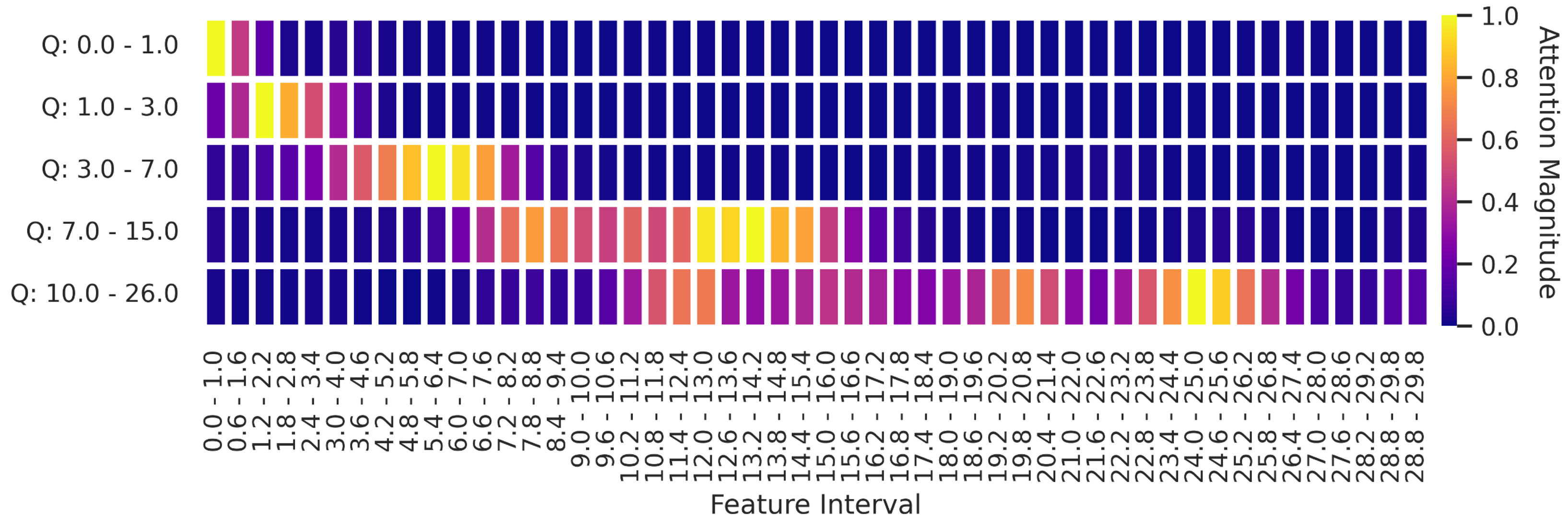
Model	xp	LLM	Verb	Noun	Action
Visual-only models					
MFormer-HR [37]	336p	✗	67.0	58.5	44.5
MoViNet-A6 [27]	320p	✗	72.2	57.3	47.7
MeMViT [55]	224p	✗	71.4	60.3	48.4
Omnivore [14]	224p	✗	69.5	61.7	49.9
MTV [59]	280p	✗	69.9	63.9	50.5
LaViLa (TSF-L) [63]	224p	✓	72.0	62.9	51.0
AVION (ViT-L) [62]	224p	✓	73.0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

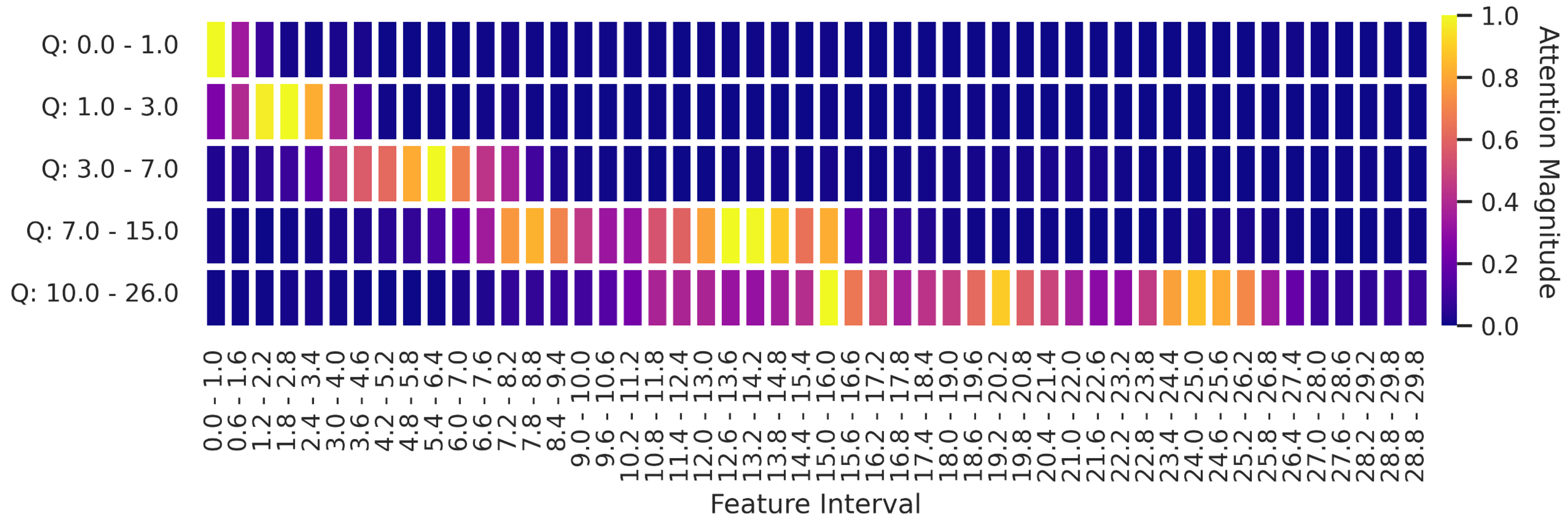
Perception Test Action				
Model	MLP (V)	MTCN [25](A+V)	**TIM (V)**	**TIM (A+V)**
Top-1 acc	43.7	51.2	56.1	**61.1**
Perception Test Sound				
Model	MLP (A)	MTCN [25](A+V)	**TIM (A)**	**TIM (A+V)**
Top-1 acc	50.6	52.9	54.8	**56.1**

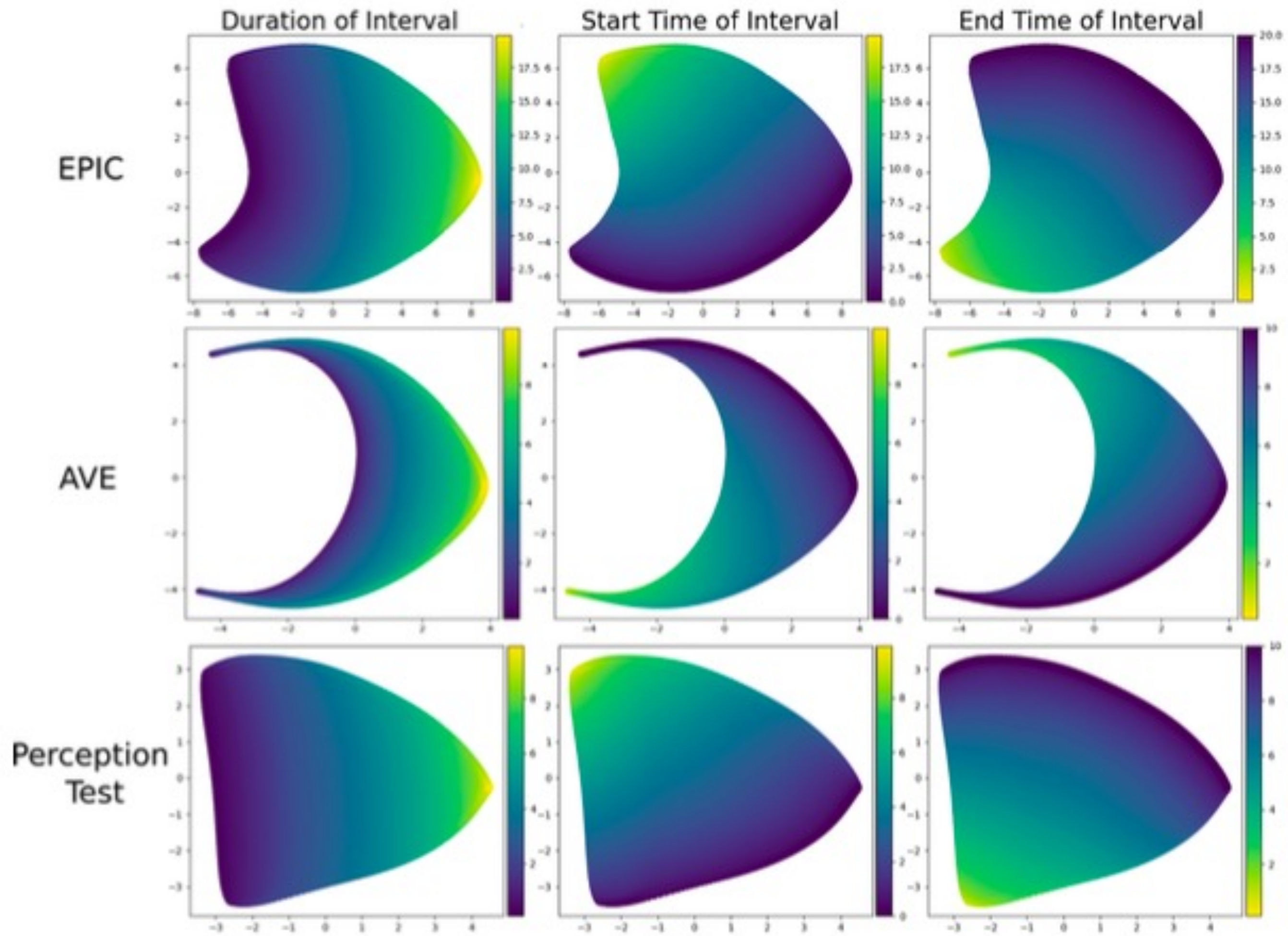
Table 5. Comparisons to trained recognition baselines on the Perception Test validation split. We show both action and sound recognition and the benefit of including audio-visual in TIM for both challenges. **V** : visual and **A** : audio input features. MLP is the result by training an MLP classifier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

TIM: A Time-Interval Audio-Visual Machine

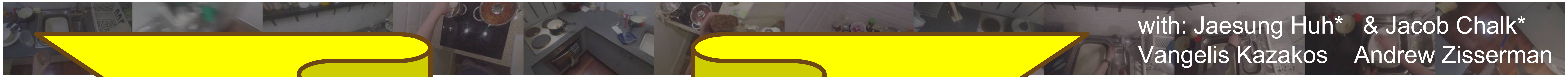
with: Jacob Chalk* Jaesung Huh*
Vangelis Kazakos Andrew Zisserman







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



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-Modality in Egocentric Data

V

High frame-rate RGB footage from the camera wearer's perspective

L

Speech in the video... or
Narrations/Captions added to index the videos

ICCV23
PARIS

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

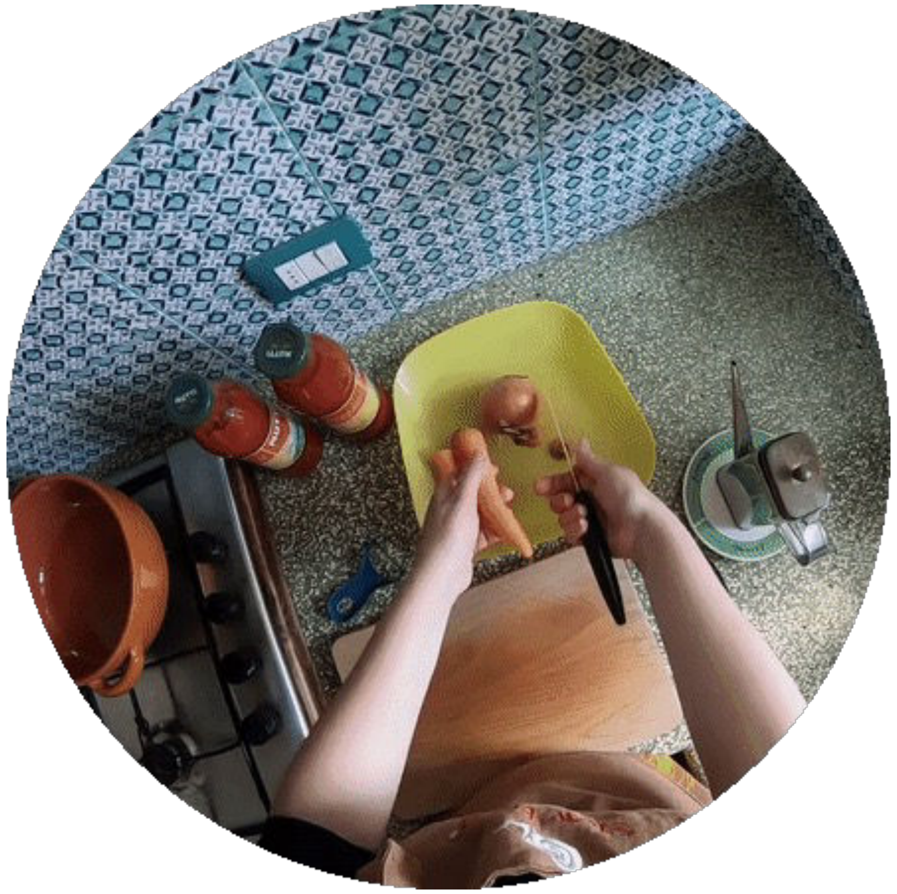
Chiara Plizzari, Toby Perrett, Barbara Caputo, Dima Damen

University of BRISTOL

Politecnico di Torino

Generalisation across Scenarios and Locations

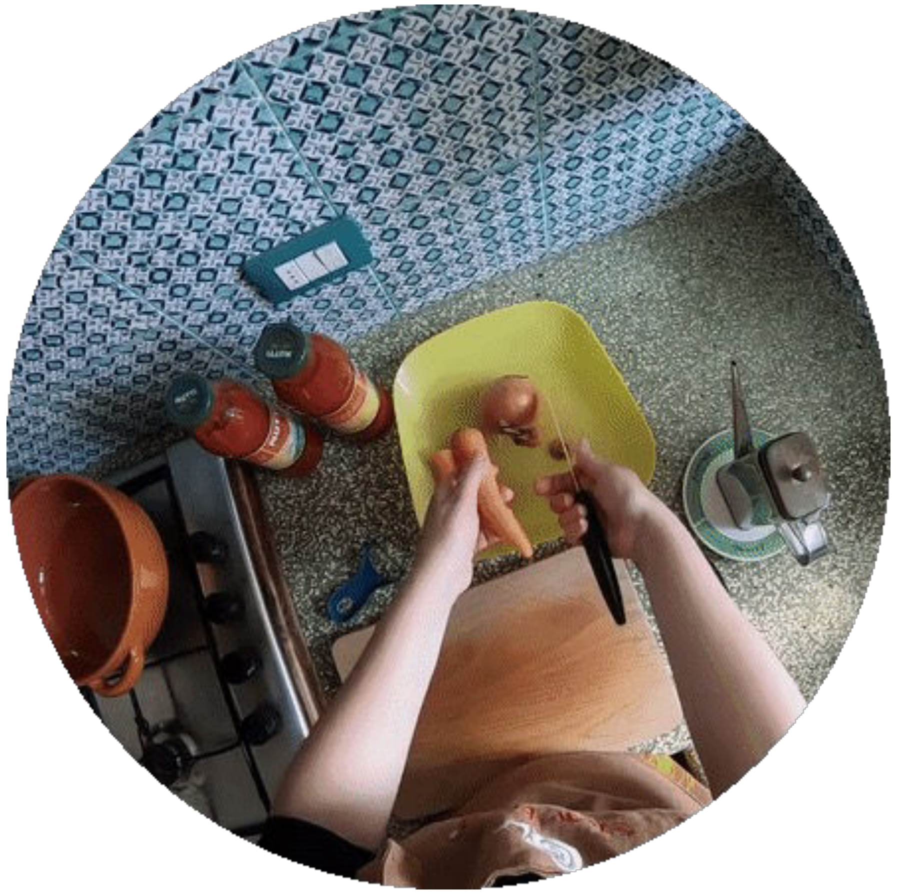
with: Chiara Plizzari
Toby Perrett

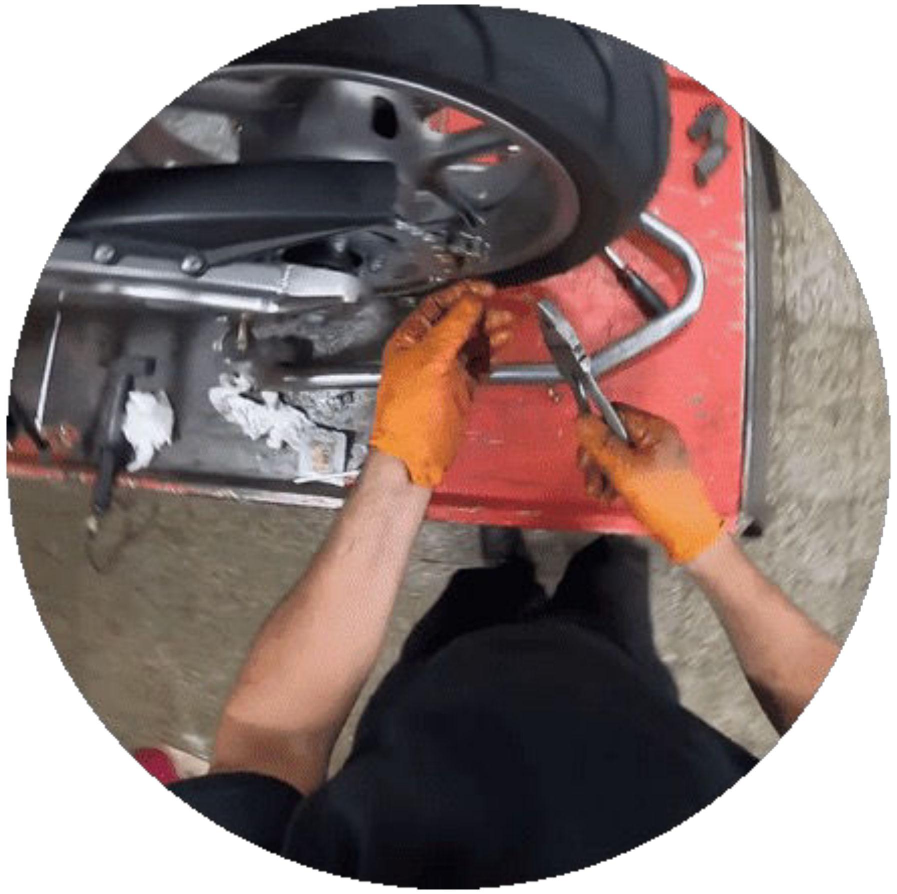


Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett

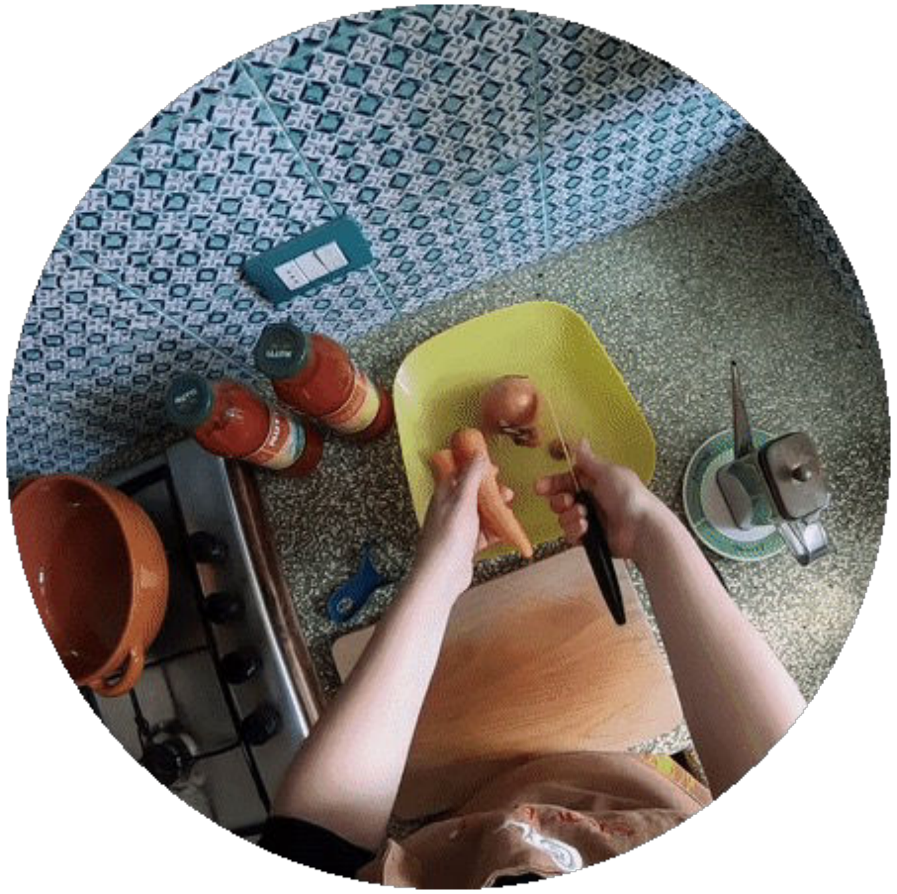


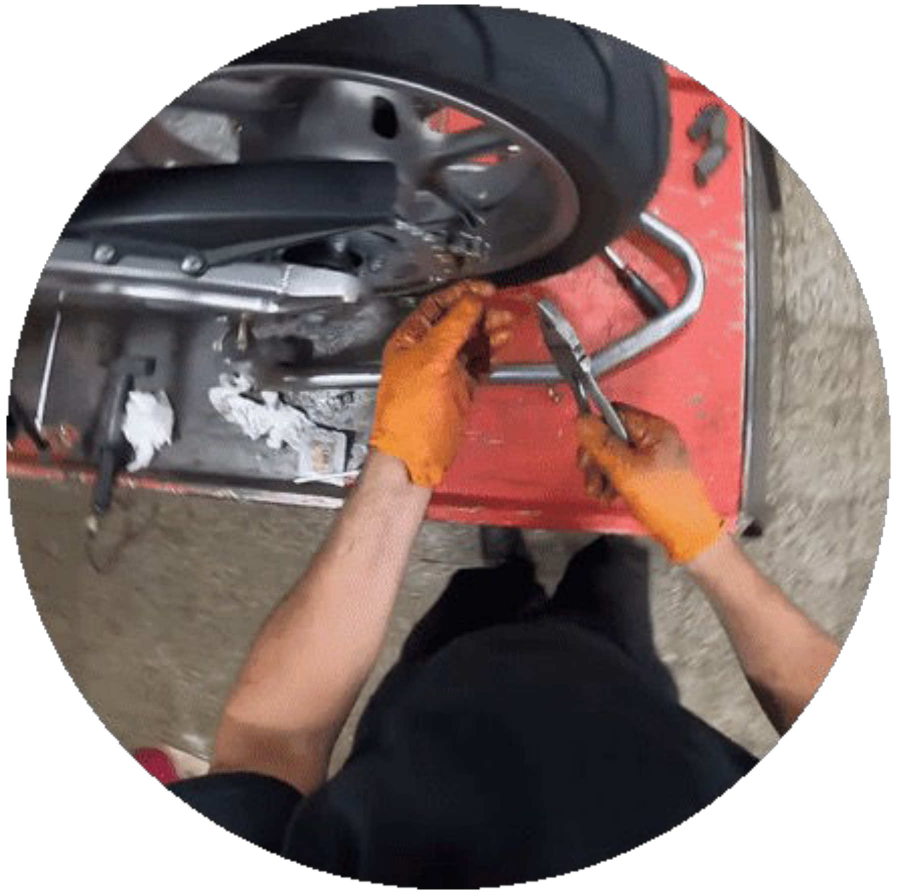


Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett

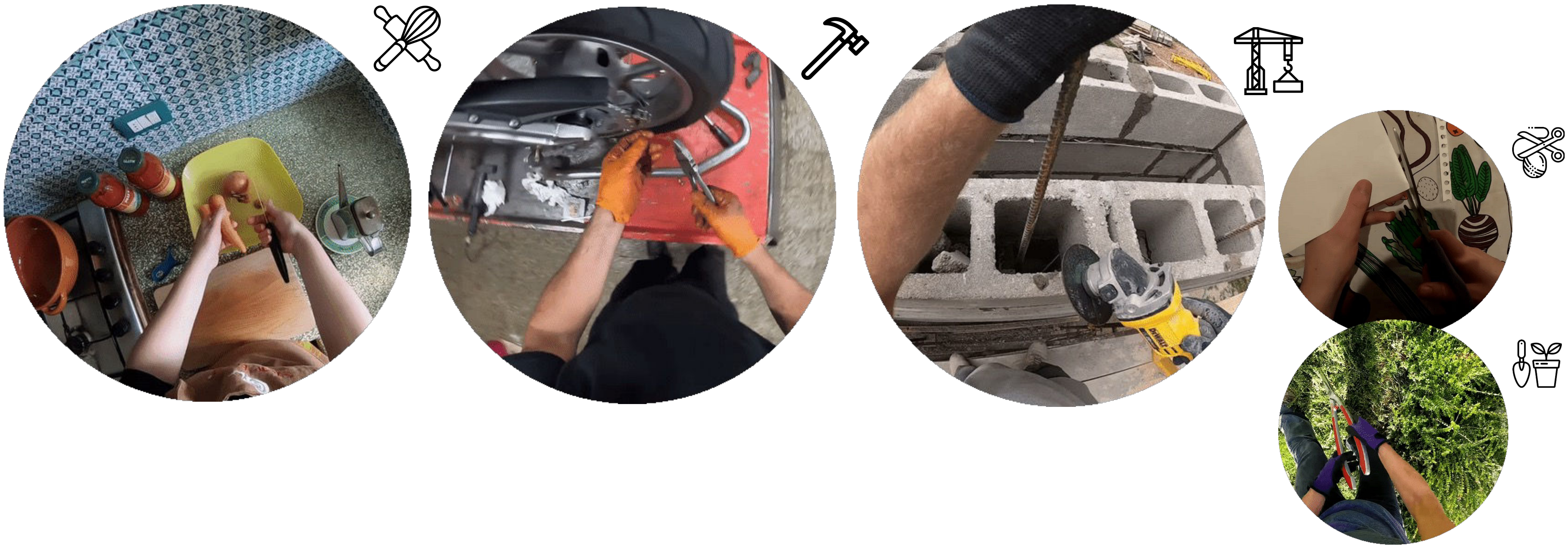




Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett



University of BRISTOL

Plizzari et al (2023). What 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[...]national 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

EGO 4D

13 locations

10 scenarios

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).

Dataset: ARGO1M

with: Chiara Plizzari
Toby Perrett

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

EGO4D

13 locations

10 scenarios

60 action classes

University of Bristol

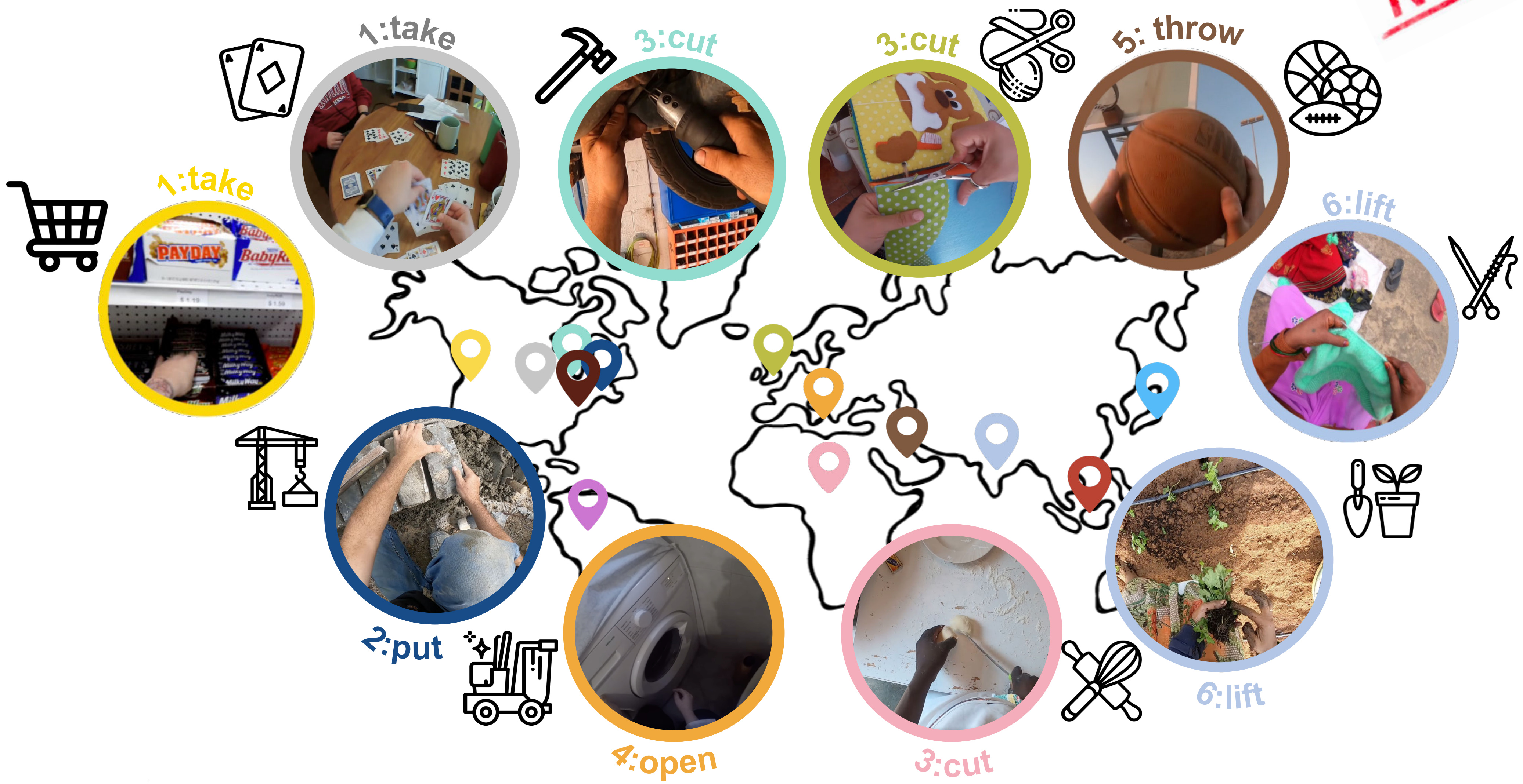
Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

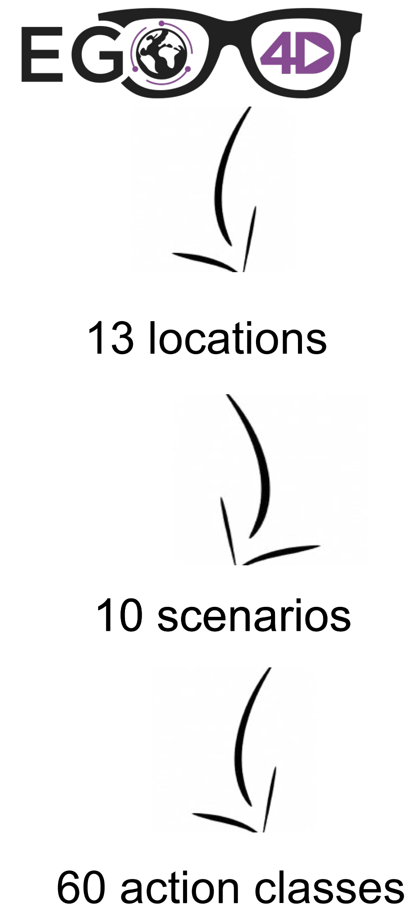
Dataset: ARGO1M

with: Chiara Plizzari
Toby Perrett

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

NEW **1.1M samples**





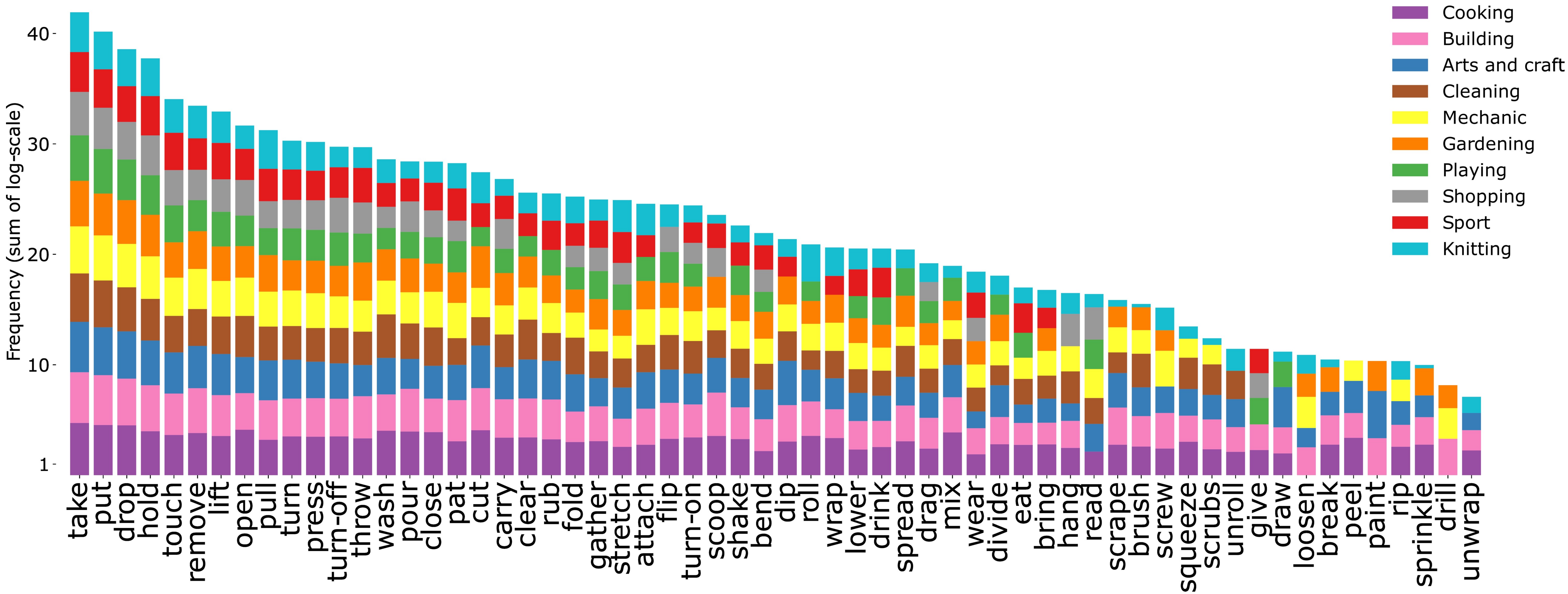
University of BRISTOL

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).

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



ARGO1M Splits

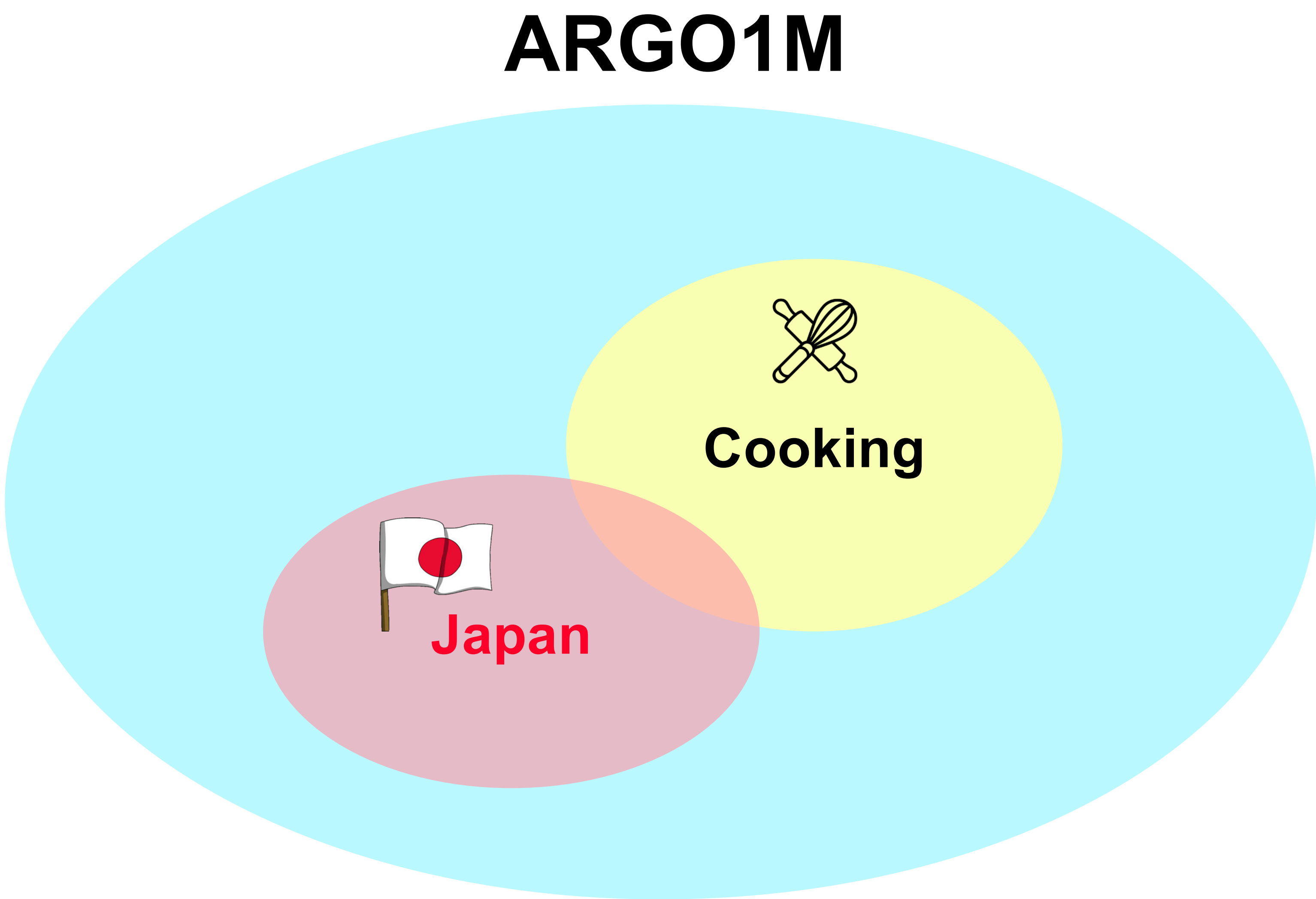
with: Chiara Plizzari
Toby Perrett

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

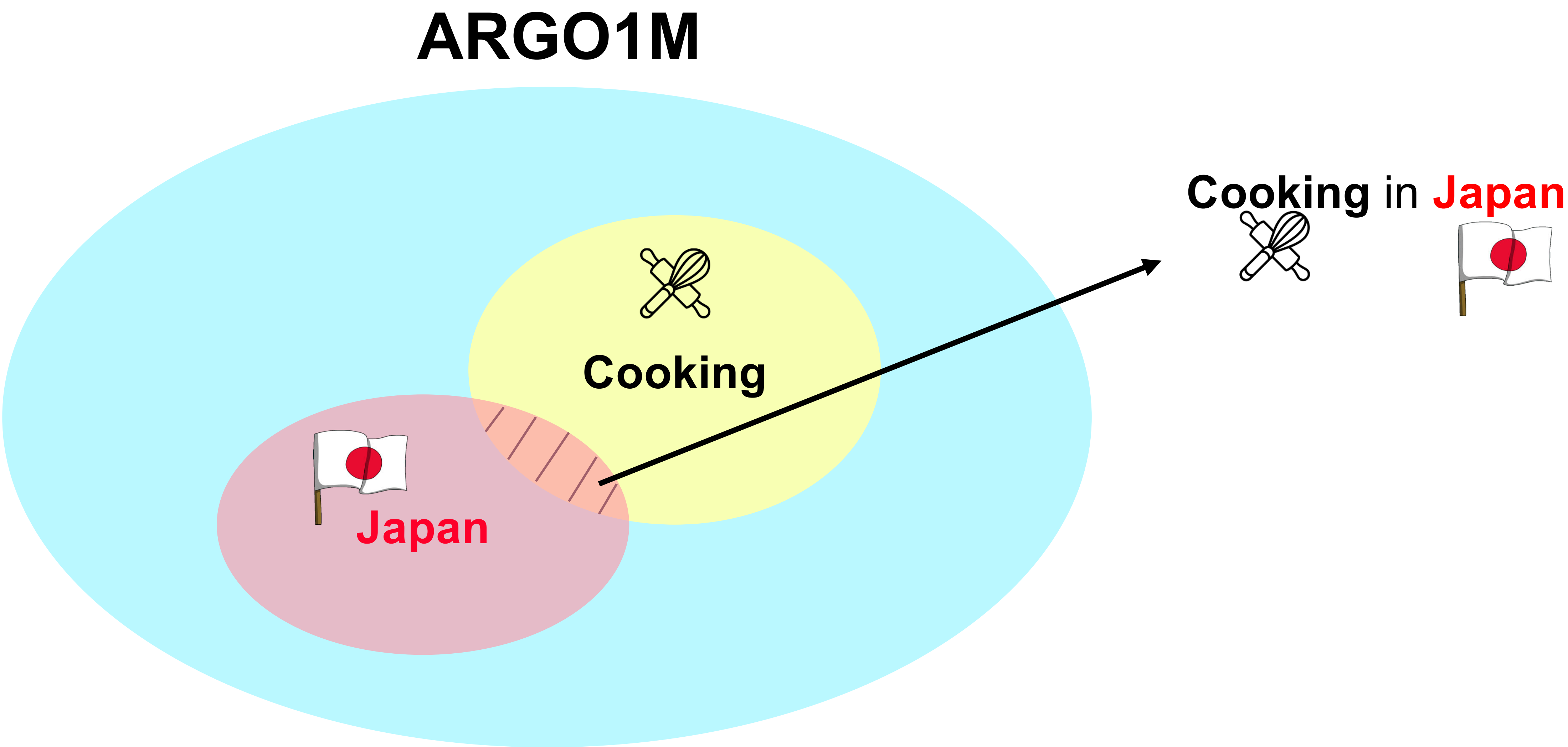
with: Chiara Plizzari
Toby Perrett



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).

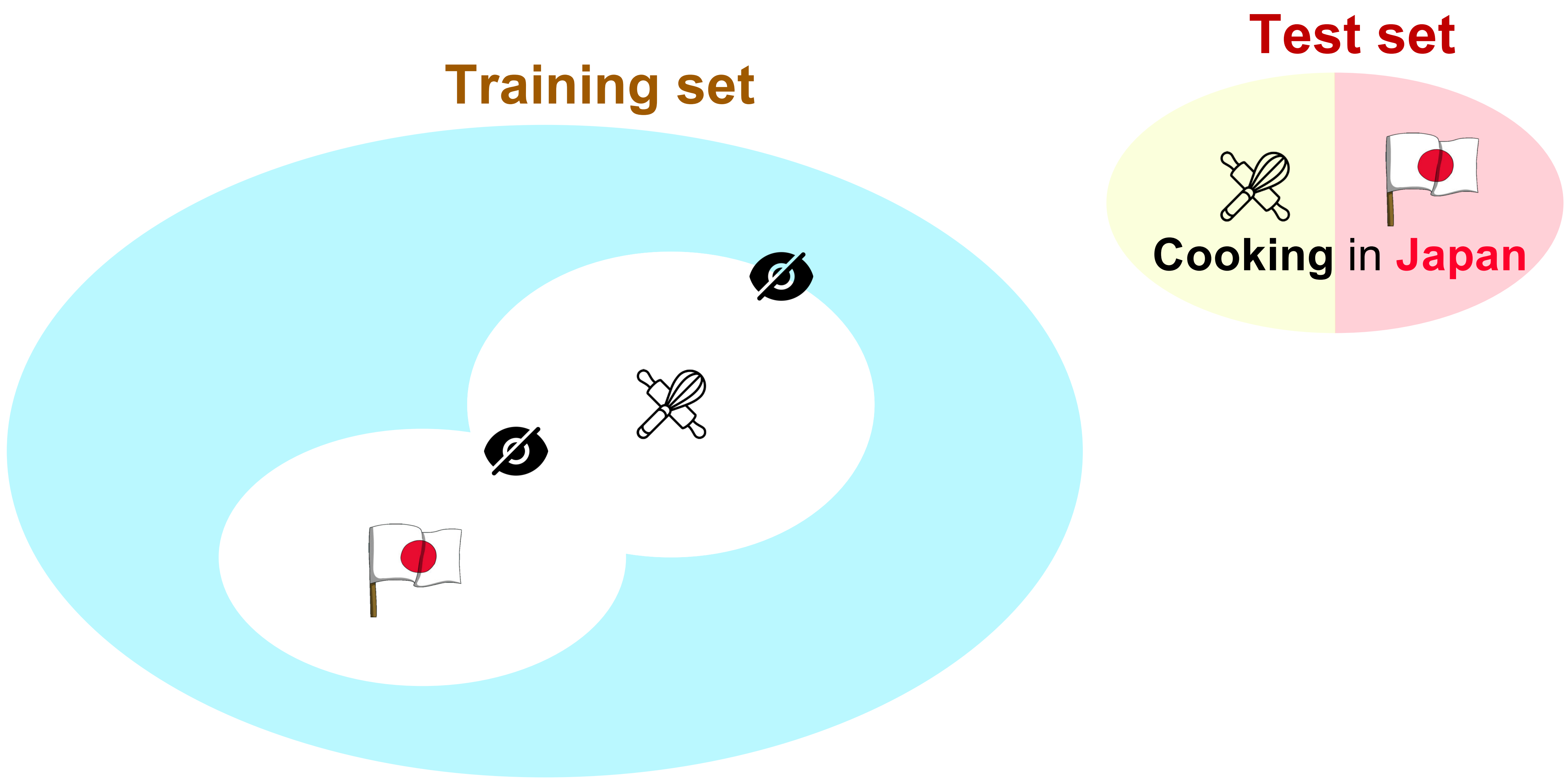
ARGO1M Splits

with: Chiara Plizzari
Toby Perrett



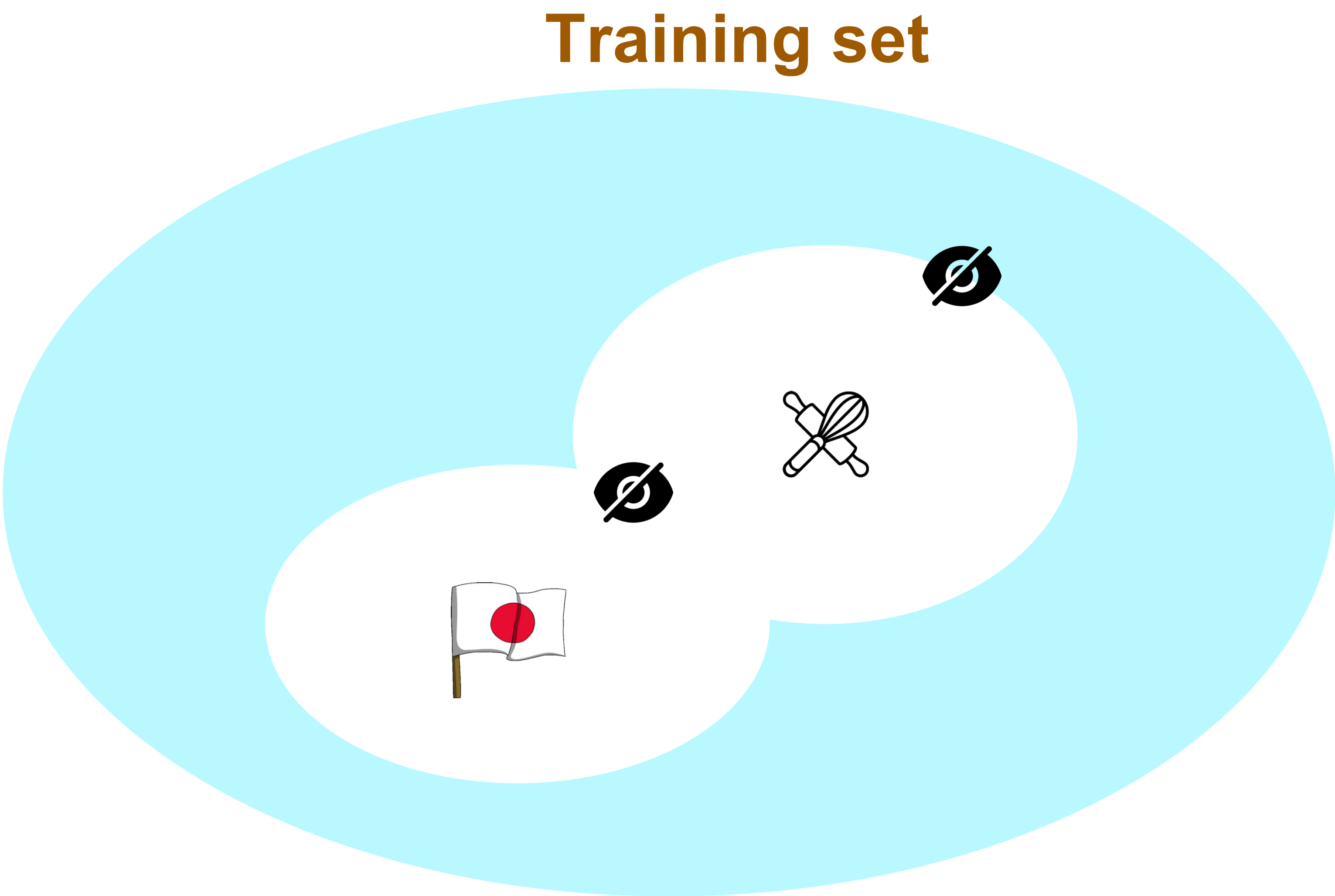
ARGO1M Splits

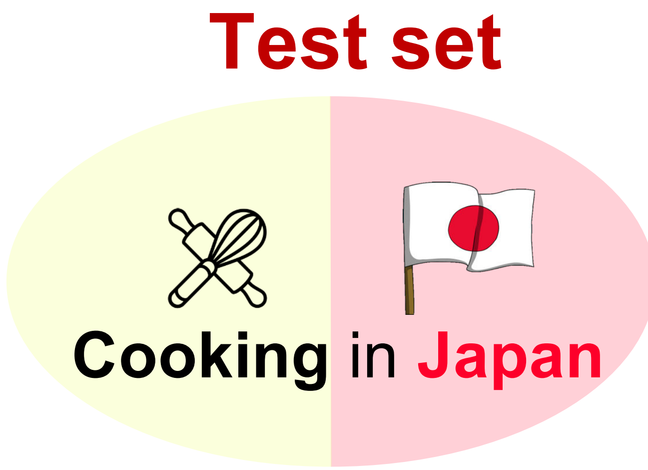
with: Chiara Plizzari
Toby Perrett

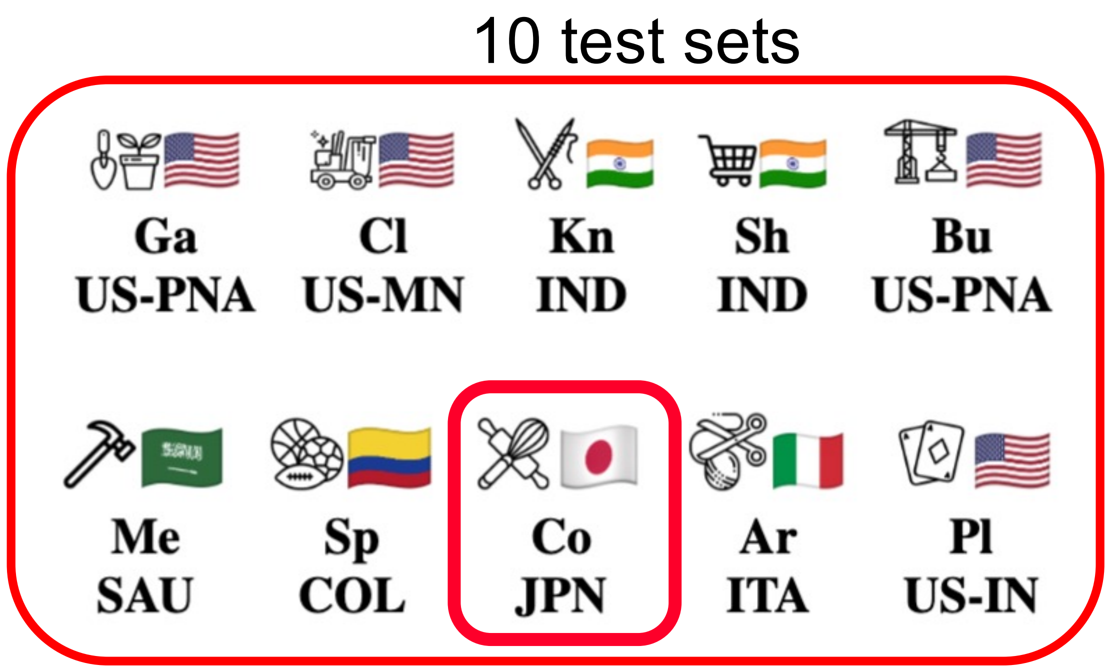


ARGO1M Splits

with: Chiara Plizzari
Toby Perrett

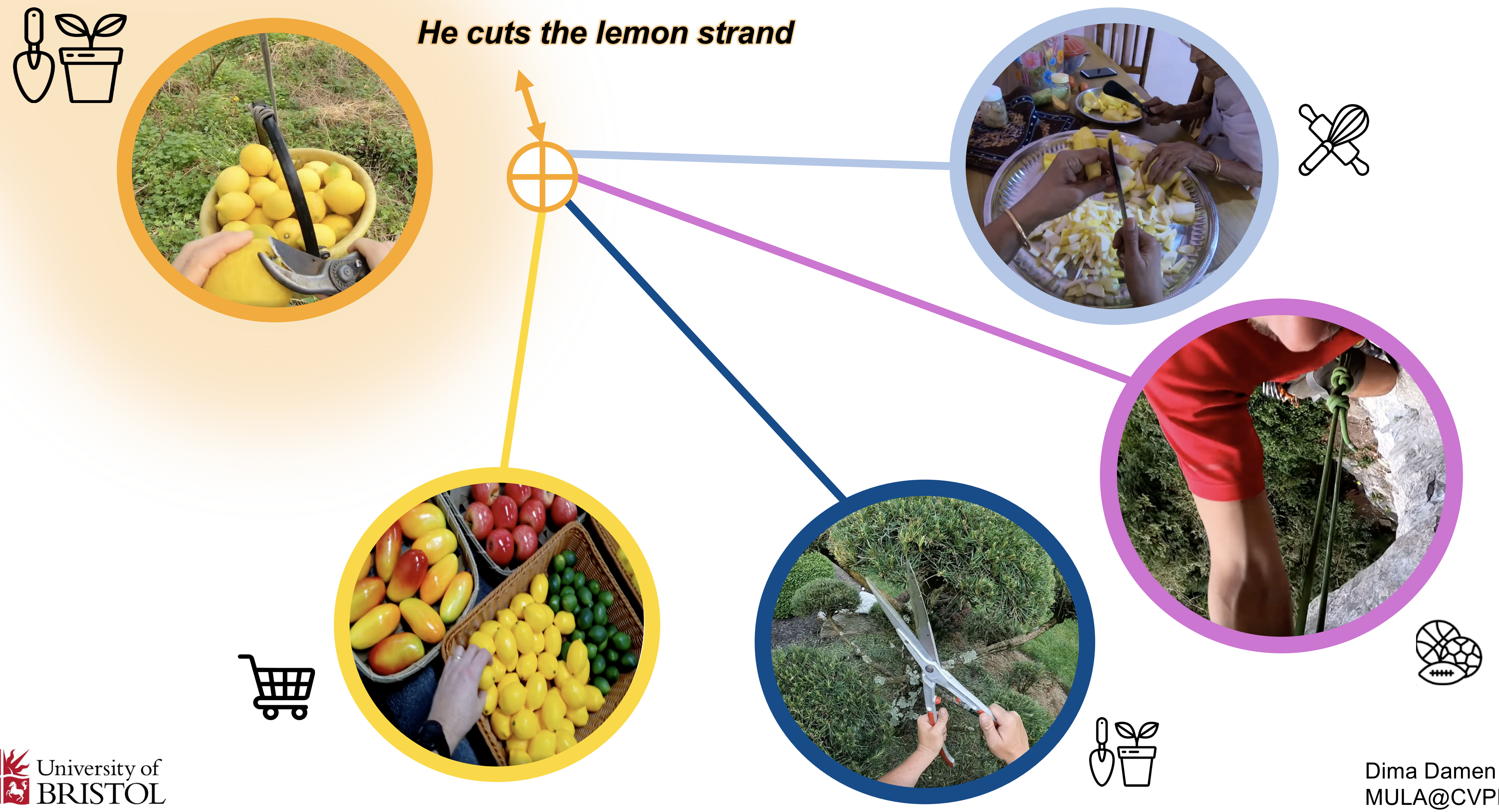






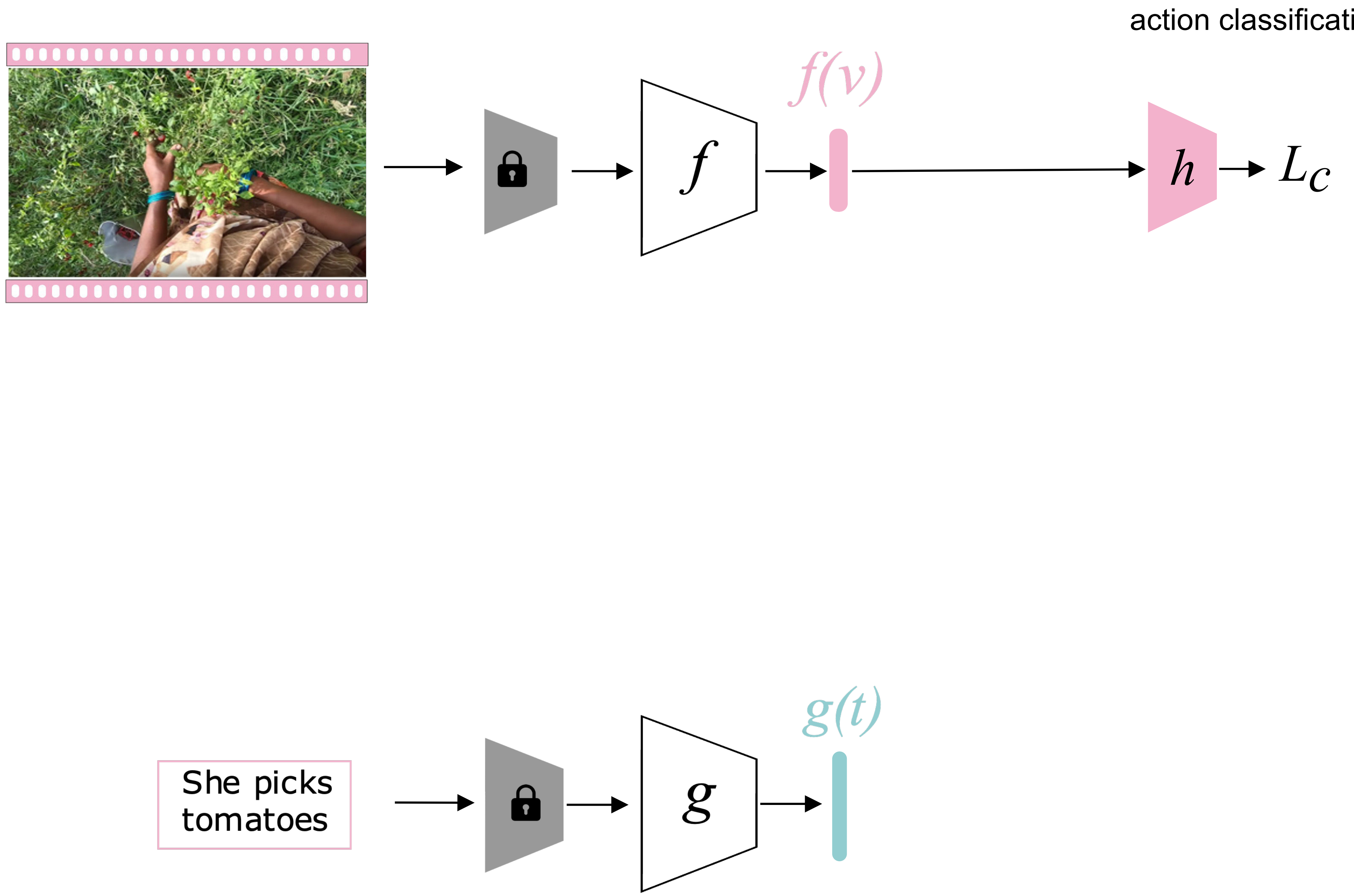
Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett



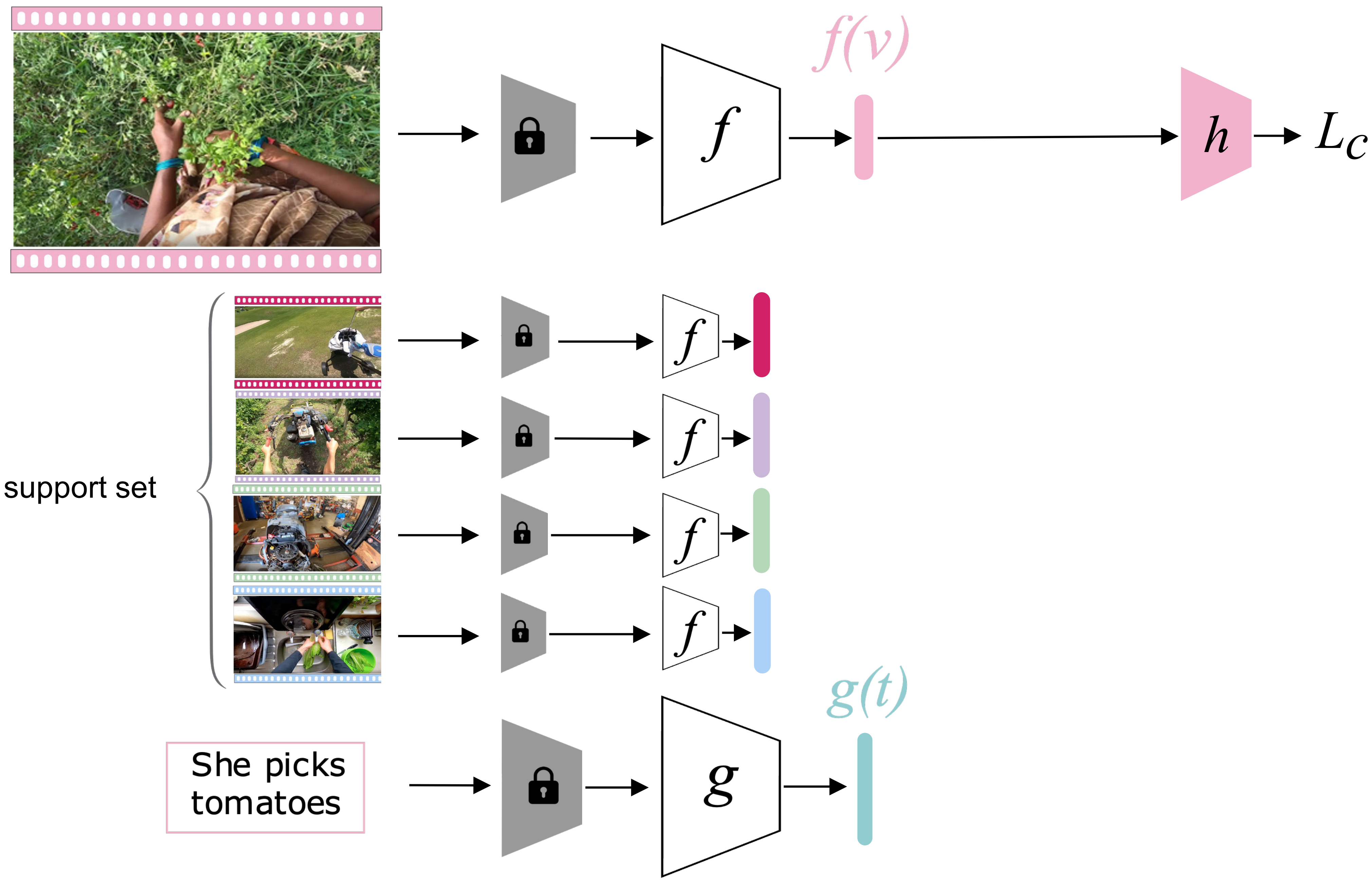
Proposed method: CIR

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).

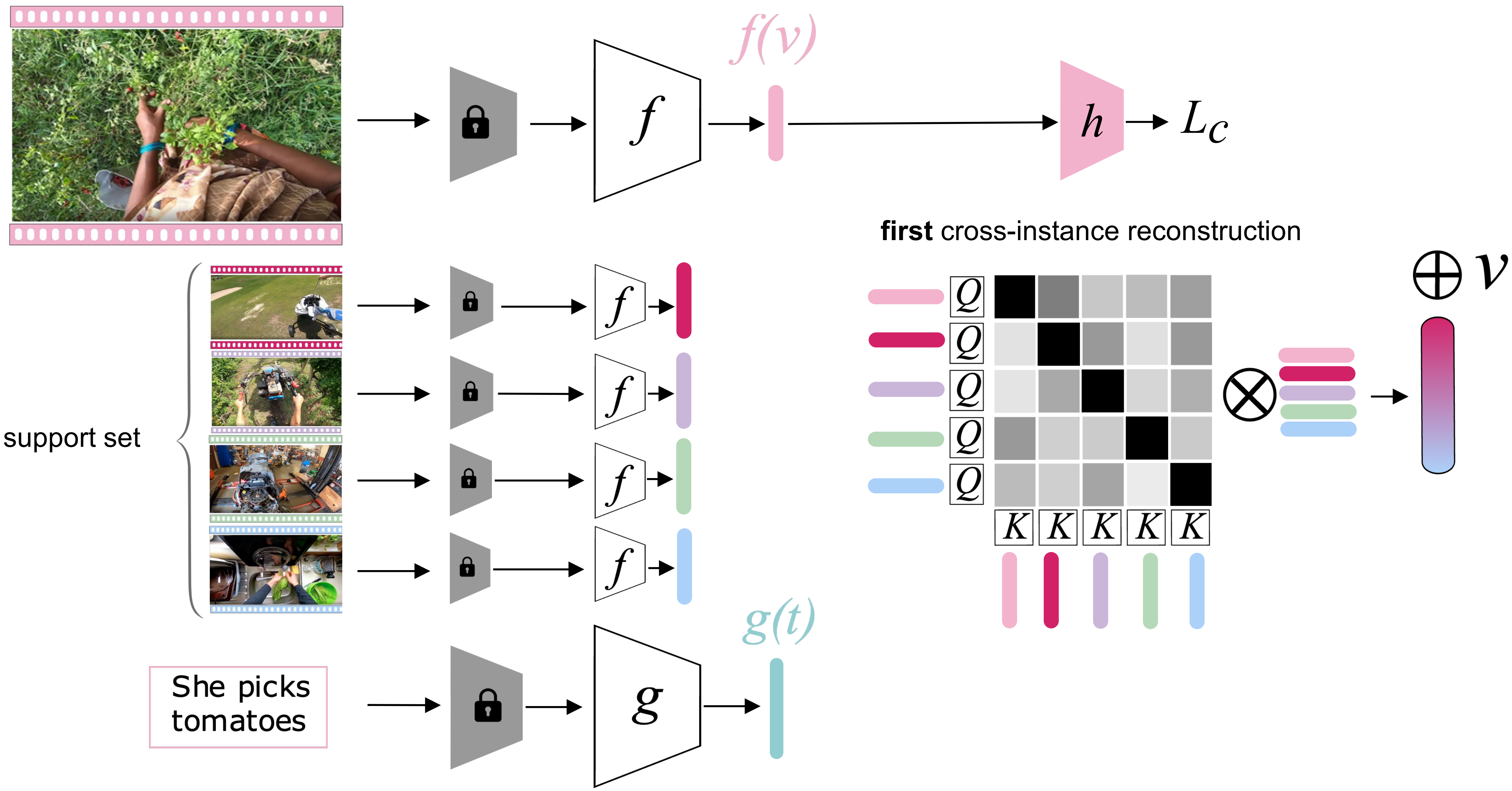
Proposed method: CIR



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).

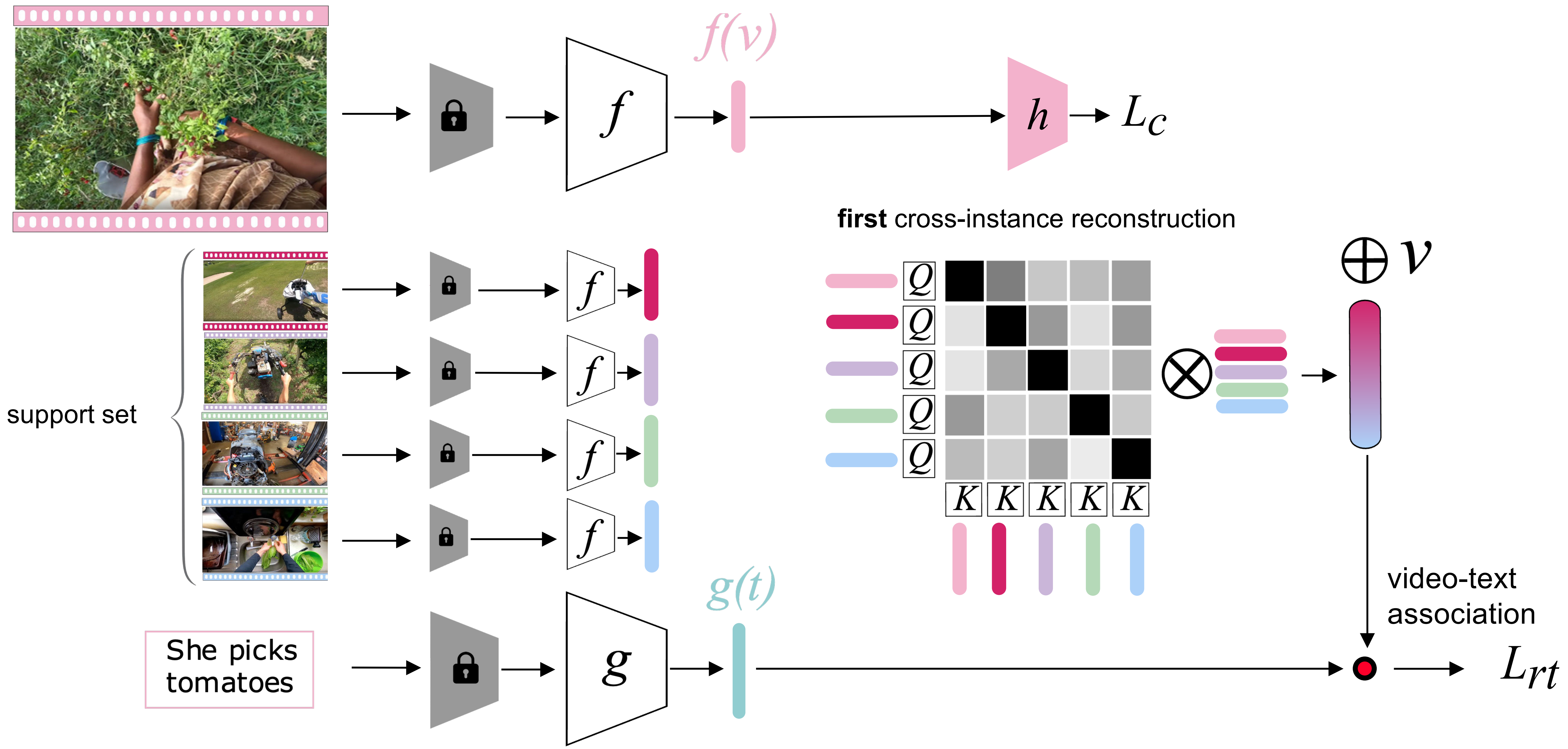
Proposed method: CIR

with: Chiara Plizzari
Toby Perrett



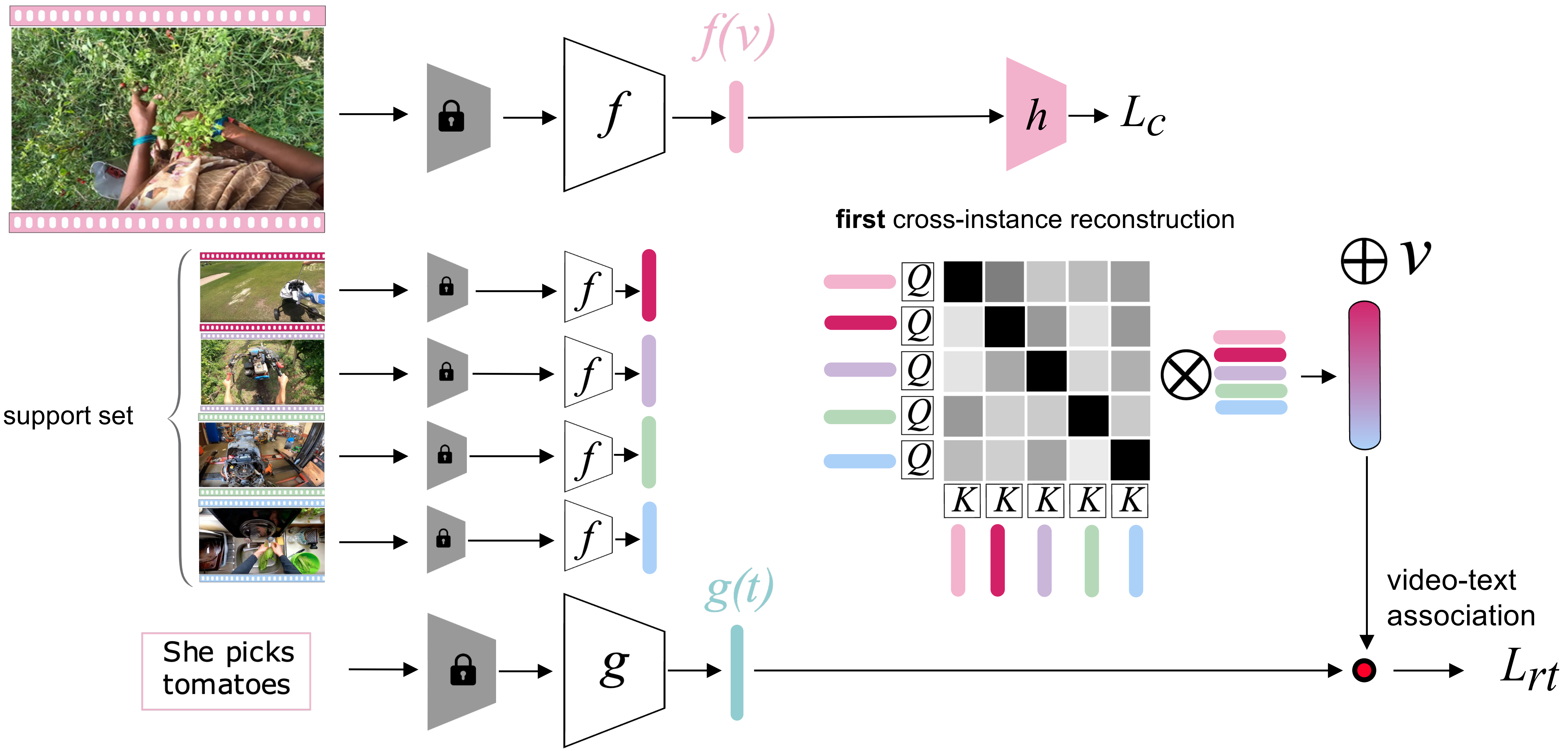
Proposed method: CIR

with: Chiara Plizzari
Toby Perrett



Proposed method: CIR

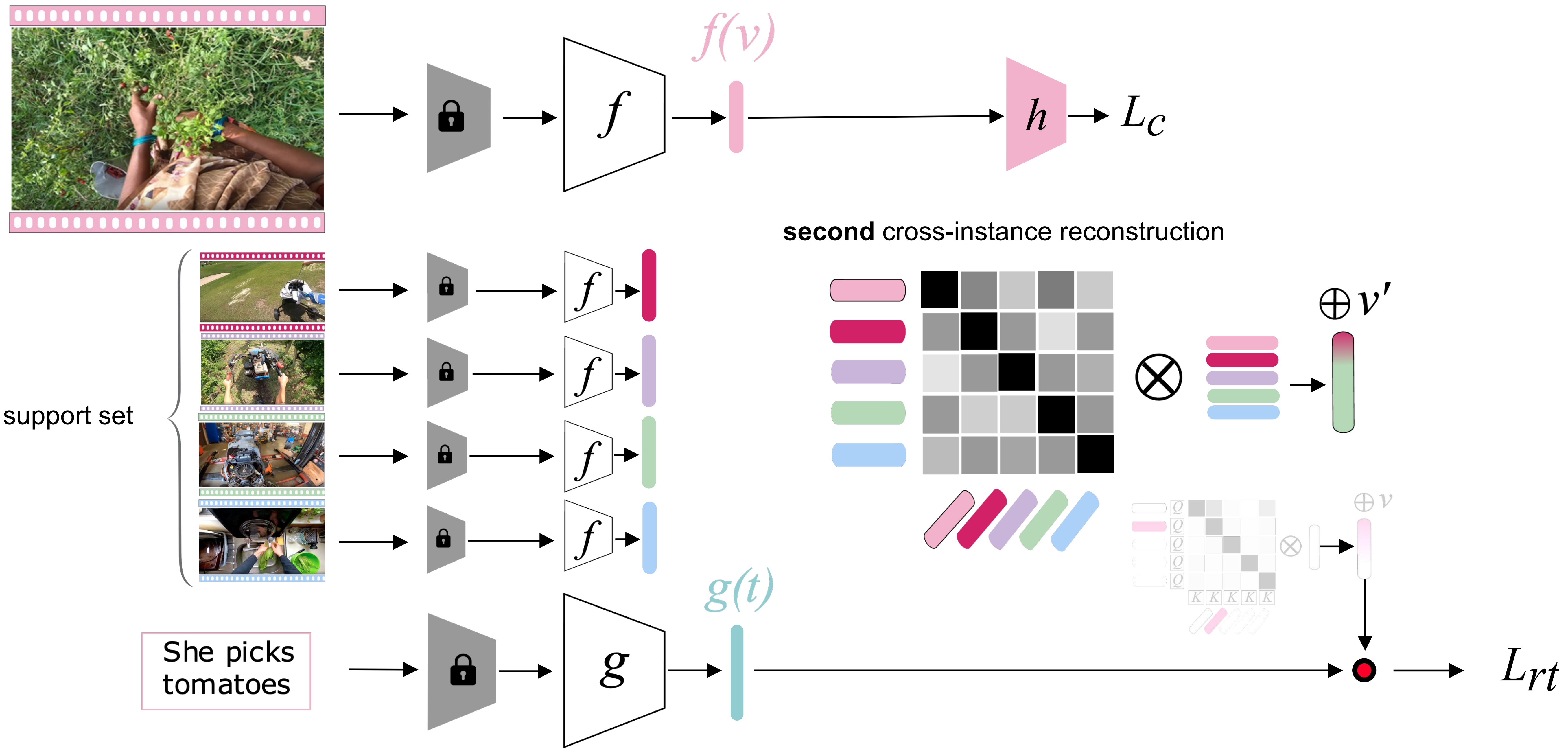
with: Chiara Plizzari
Toby Perrett



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



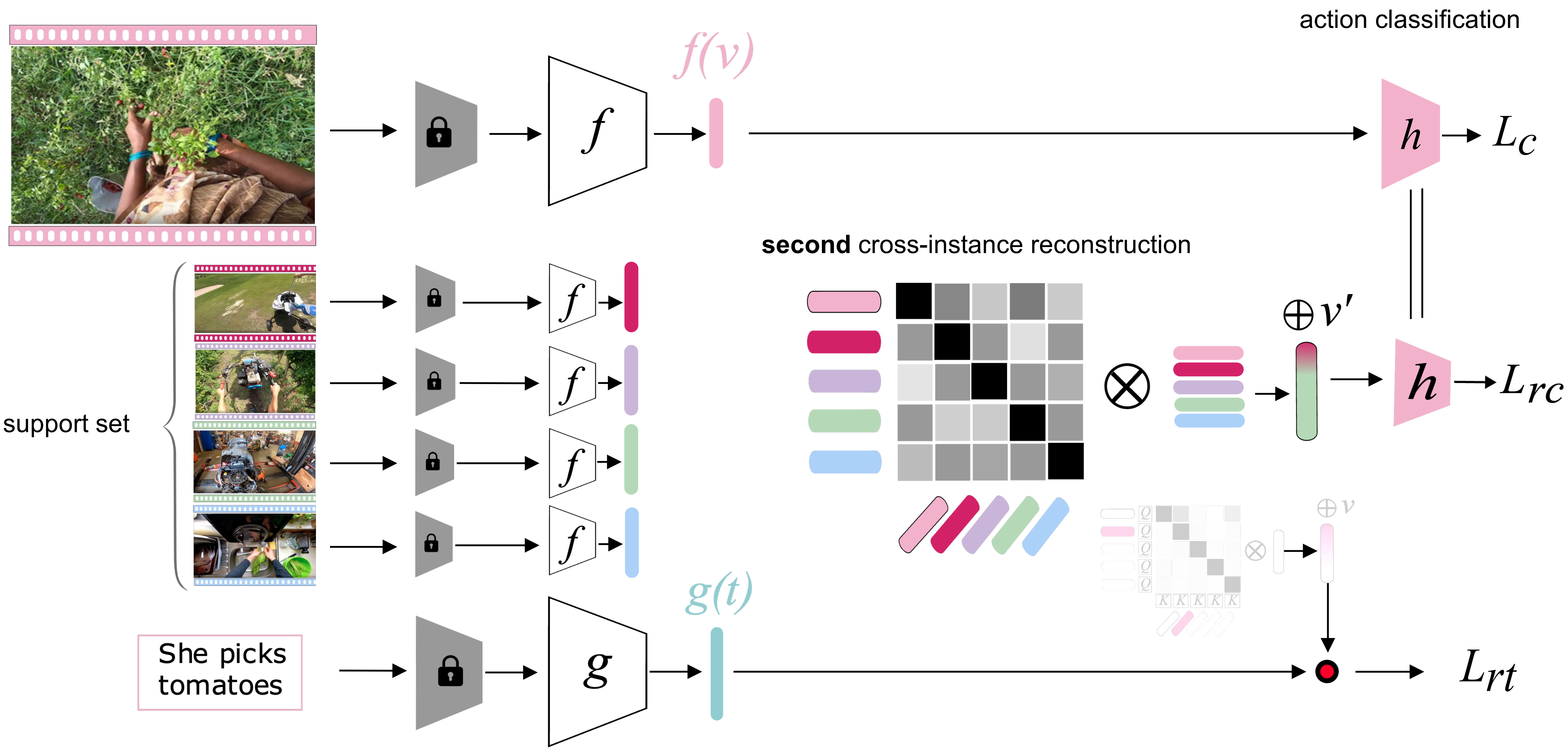
University of BRISTOL

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).

Dima Damen
MULA@CVPR2024

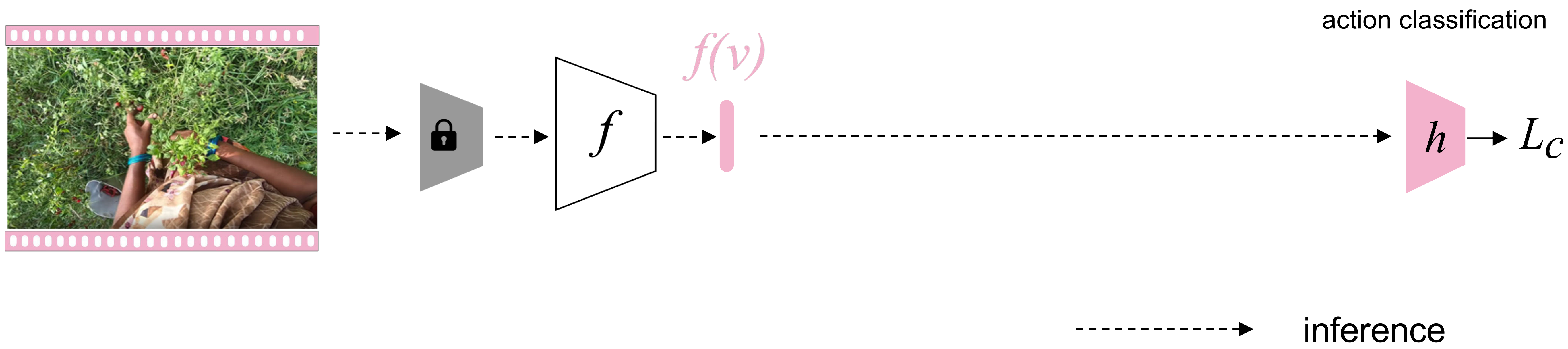
Proposed method: CIR

with: Chiara Plizzari
Toby Perrett



Proposed method: CIR

with: Chiara Plizzari
Toby Perrett



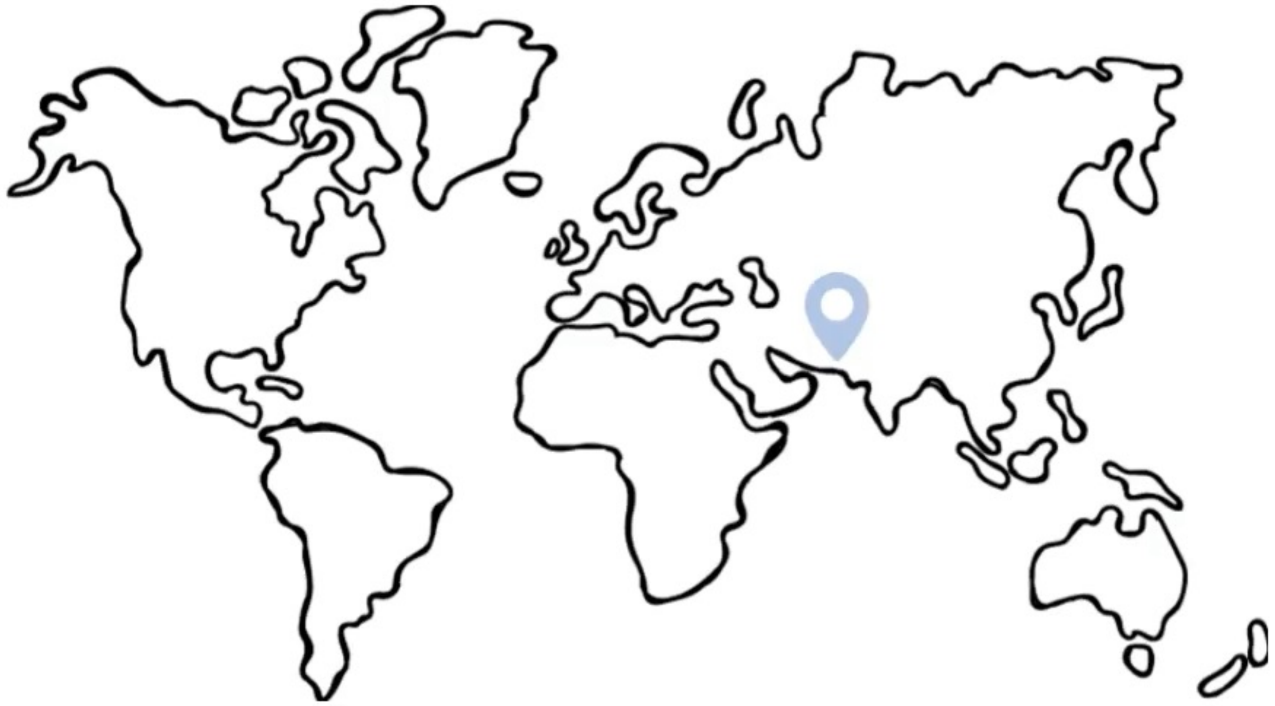
Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Examples

Chiara Plizzari
Toby Perrett
Dima Damen

#C C drops the cut vegetables

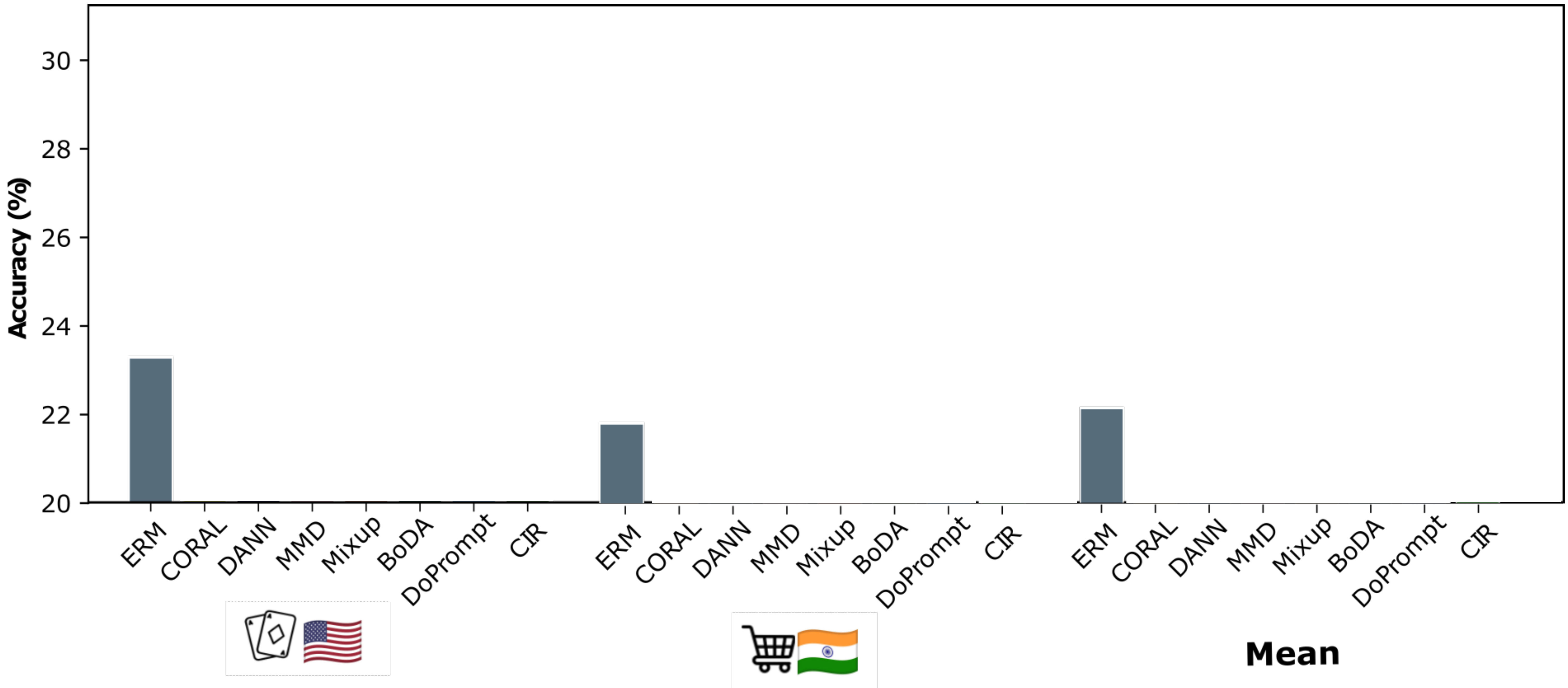
query



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

We propose and address a new generalisation problem: can a model trained for action recognition successfully classify actions when they are performed within a previously unseen scenario and in a previously unseen location? To answer this question, we introduce the Action Recognition Generalisation Over scenarios and locations dataset (ARGO1M), which contains 1.1M video clips from the large-scale Ego4D dataset, across 10 scenarios and 13 locations. We demonstrate recognition models struggle to generalise over 10 proposed test splits, each of an unseen scenario in an unseen location. We thus propose CIR, a method to represent each video as a Cross-Instance Reconstruction of videos from other domains. Reconstructions are paired with text narrations to guide the learning of a domain generalisable representation. We provide extensive analysis and ablations on ARGO1M that show CIR outperforms prior domain generalisation works on all test splits. Code and data: https://chiaraplizzari.github.io/what-can-a-cook/.

1. Introduction

A notable distinction between human and machine intelligence is the ability of humans to generalise. We can see an example of the action "cut" performed by a cook in Italy, and recognise the same action performed in a different geographic *location*, *e.g.* India, despite having never visited. We can also recognise actions within new *scenarios*, such as a mechanic cutting metal, even if we are unfamiliar with the tools they use.

This problem is known as domain generalisation [62], where a model trained on a set of labelled data fails to generalise to a different distribution in inference. The gap between distributions is known as *domain shift*. To date, works have focused on generalising over visual domain shifts [25, 46, 31, 10, 39]. In this paper, we introduce the *scenario shift*, where the same action is performed as part

Training Test

: Cooking : Building : Arts and crafts : Mechanic : Gardening

Figure 1: Problem statement and samples from the ARGO1M dataset. The same action, e.g. "cut", is performed differently based on the *scenario* and the *location* in which it is carried out. We aim to generalise so as to recognise the same action within a new scenario, *unseen* during training, and in an *unseen location*, *e.g.*, *Mechanic* ( ) in *India* ( ).

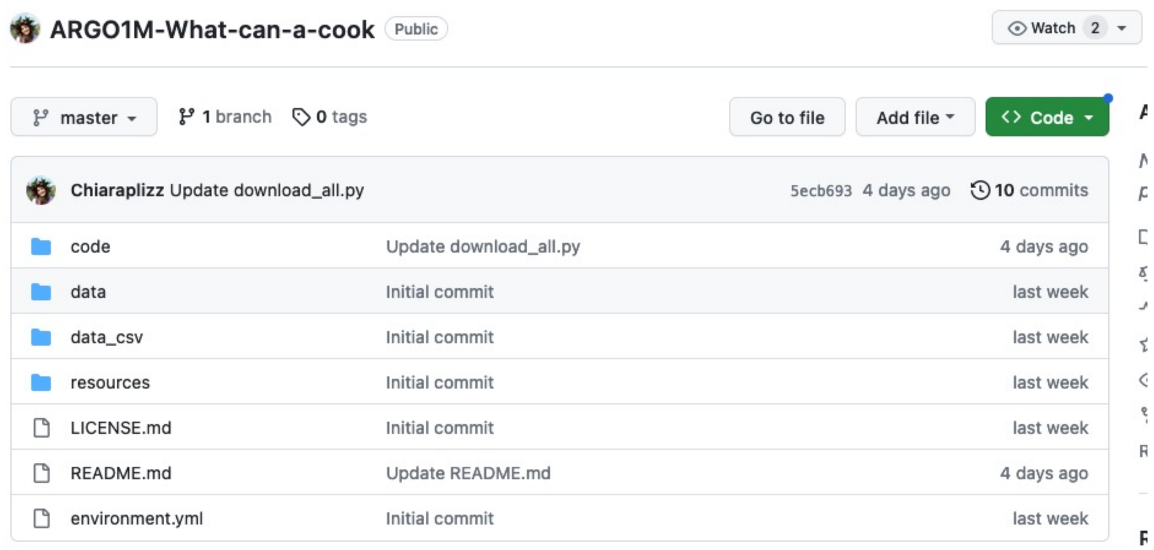
of a different activity, impacting the tools used, objects interacted with, goals and behaviour. We combine this with the location shift, generalising over both simultaneously.

In Fig. 1, the action 'cut' is performed using a knife whilst cooking ( ), pliers whilst building ( ) and scissors for arts and crafts ( ). Tools are not specific for a scenario and can vary over locations – e.g. in Fig. 1, seaweed sheets are cut with scissors while cooking in Japan. Generalising would be best achieved by learning the notion of "cutting" as separating an object into two or more pieces, regardless of the tool or background location. Successful generalisation can thus enable recognising metal being "cut" by a mechanic in India using an angle grinder (Fig. 1 Test).

Our investigation is enabled by the recent introduction of the Ego4D [17] dataset of egocentric footage from around the world. We curate a setup specifically for action generalisation, called ARGO1M. It contains 1.1M action clips of 60 classes from 73 unique scenario/location combinations.

To tackle the challenge of ARGO1M, we propose a new method for domain generalisation. We represent each video

*Work carried during Chiara's research visit to the University of Bristol



ARGO1M Dataset
CIR Method
Code and Models

RELEASED

Wed (Session 2)
Poster # 172

CVPR JUNE 17-21, 2024 SEATTLE, WA

GenHowTo: Learning to Generate Actions and State Transformations from Instructional Videos


Tomáš Souček


Dima Damen


Michael Wray


Ivan Laptev


Josef Šivic

University of BRISTOL

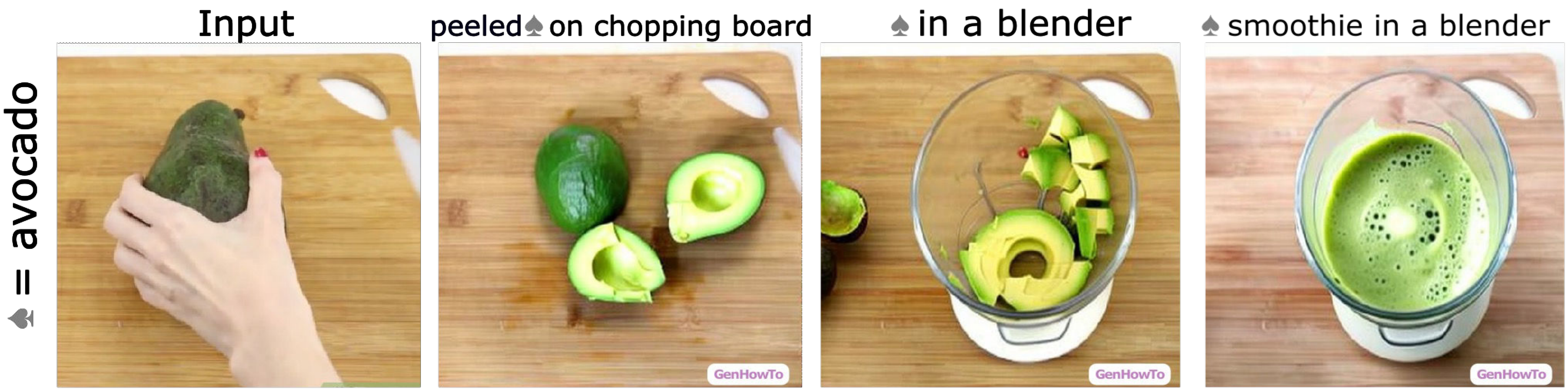
CTU CZECH TECHNICAL UNIVERSITY IN PRAGUE

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

GenHowTo…

with: Tomas Soucek, Michael Wray, Ivan Laptev, Josef Sivic

- Hands transform objects….



GenHowTo...

with: Tomas Soucek
Ivan Laptev
Michael Wray
Josef Sivic

Input	GenHowTo	EF-DDPM	InstructPix2Pix

Prompt: a frosted cake with strawberries around the top

Prompt: a person kneading dough on a cutting board

Prompt: a person cutting a fish on a cutting board

GenHowTo…

with: Tomas Soucek
Ivan Laptev
Michael Wray
Josef Sivic

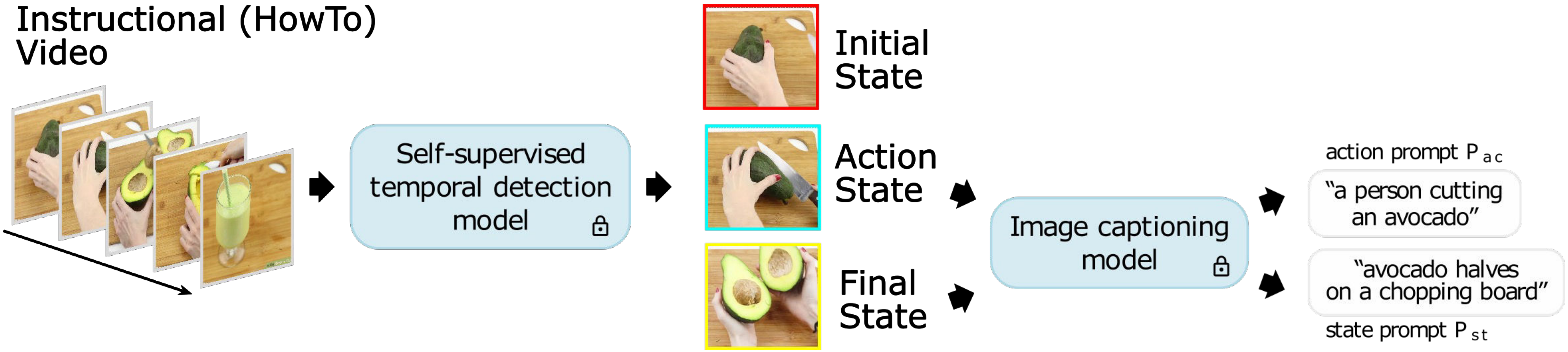
- Two contributions…. Dataset & Method

T Soucek et al (2023). GenHowTo: Learning to Generate Actions and State Transformations from Instructional Videos. ArXiv

GenHowTo...

with: Tomas Soucek
Ivan Laptev
Michael Wray
Josef Sivic

- Two contributions…. Dataset & Method

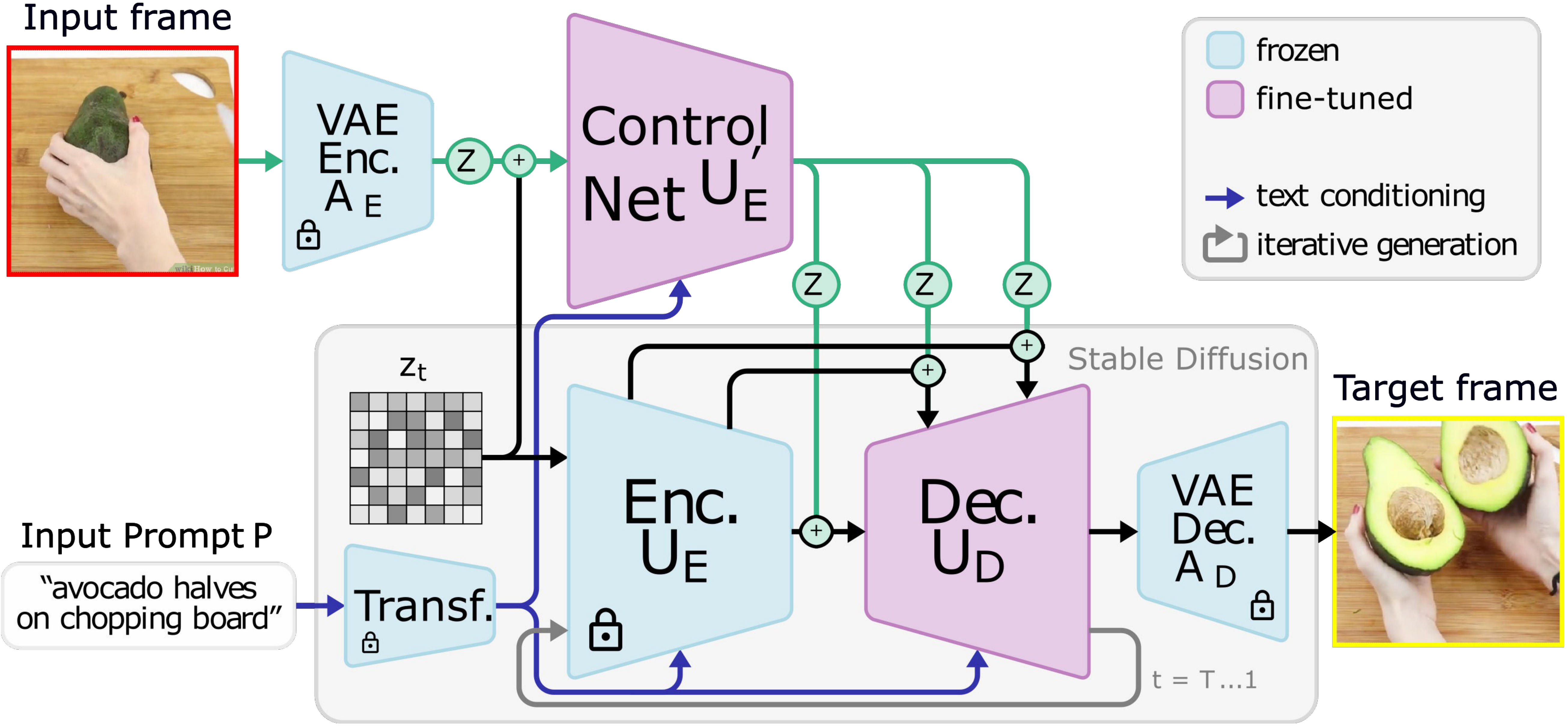


Tomas Soucek, Jean-Baptiste Alayrac, Antoine Miech, Ivan Laptev, and Josef Sivic (2022). Multi-task learning of object state changes from uncurated videos.

GenHowTo…

with: Tomas Soucek Michael Wray
Ivan Laptev Josef Sivic

- Two contributions…. Dataset & Method



GenHowTo...

with: Tomas Soucek
Ivan Laptev
Michael Wray
Josef Sivic

T Soucek et al (2023). GenHowTo: Learning to Generate Actions and State Transformations from Instructional Videos. ArXiv

GenHowTo...

with: Tomas Soucek
Ivan Laptev
Michael Wray
Josef Sivic

- Qualitative Evaluation…
  - Initial vs Final State
  - Binary Classifier

	Method	$\text{Acc}_{\text{ac}} \uparrow$	$\text{Acc}_{\text{st}} \uparrow$
	test set categories unseen during training		
(a)	Stable Diffusion	0.51	0.50
(b)	Edit Friendly DDPM	0.60	0.61
(c)	InstructPix2Pix	0.55	0.63
(d)	*CLIP (manual prompts)*	0.52	0.62
(e)	**GenHowTo**	**0.66**	**0.74**
	test set categories seen during training		
(f)	Edit Friendly DDPM†	0.69	0.80
(g)	**GenHowTo†**	**0.77**	**0.88**
(h)	*Real images*	0.96	0.97

† Models trained also on the test set *categories*.

GenHowTo...

with: Tomas Soucek
Ivan Laptev
Michael Wray
Josef Sivic

a person is wrapping a tortilla on a plate

REAL IMAGE —— **GENERATED**

a plate with two burritos on it

REAL IMAGE —— **GENERATED**

a man pouring beer into a glass

REAL IMAGE —— **GENERATED**

a man sitting at a table holding a glass of beer

REAL IMAGE —— **GENERATED**

T Soucek et al (2023). GenHowTo: Learning to Generate Actions and State Transformations from Instructional Videos. ArXiv

Wed (Session 2)
Poster # 172

CVPR
JUNE 17-21, 2024
SEATTLE, WA

GenHowTo: Learning to Generate Actions and State Transformations from Instructional Videos

Tomáš Souček

Dima Damen

Michael Wray

Ivan Laptev

Josef Šivic

University of BRISTOL

CTU CZECH TECHNICAL UNIVERSITY IN PRAGUE

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Multi-Modality in Egocentric Data

V High frame-rate RGB footage from the camera wearer's perspective

A One or many microphones, on the wearable device, best positioned to capture the sounds of actions and interactions

L Speech in the video... or
Narrations/Captions added to index the videos

Short Detour…

The Perception Test

with: Viorica Patraucean, Joao Carreira, Andrew Zisserman
+ DeepMind Team

Training

Training

What comes to mind when you hear the word “Dataset”?

Training

Training

V Patraucean et al (2022). Perception Test: A Diagnostic Benchmark for Multimodal Video Models.

The Perception Test

with: Viorica Patraucean, Joao Carreira, Andrew Zisserman
+ DeepMind Team

Training and Pretraining
Large-scale Diversity Weak/sparse supervision
Kinetics-400, -600, -700 HowTo100M Ego4D

V Patraucean et al (2022). Perception Test: A Diagnostic Benchmark for Multimodal Video Models.

The Perception Test

with: Viorica Patraucean, Joao Carreira, Andrew Zisserman
+ DeepMind Team

Training and Pretraining	**Fine-Grained Actions**
Large-scale Diversity Weak/sparse supervision	Fine-grained actions Subtle variations Crowd-sourced
Kinetics-400, -600, -700 HowTo100M Ego4D	Charades Something-Something EPIC-KITCHENS Ego4D

V Patraucean et al (2022). Perception Test: A Diagnostic Benchmark for Multimodal Video Models.

The Perception Test

with: Viorica Patraucean, Joao Carreira, Andrew Zisserman
+ DeepMind Team

Training and Pretraining	**Fine-Grained Actions**	**Audio-Visual**
Large-scale Diversity Weak/sparse supervision	Fine-grained actions Subtle variations Crowd-sourced	Audio-Visual Input Video classes only
Kinetics-400, -600, -700 HowTo100M Ego4D	Charades Something-Something EPIC-KITCHENS Ego4D	Audioset VGG-Sound EPIC-KITCHENS Ego4D

V Patraucean et al (2022). Perception Test: A Diagnostic Benchmark for Multimodal Video Models.

The Perception Test

with: Viorica Patraucean, Joao Carreira, Andrew Zisserman
+ DeepMind Team

Training and Pretraining	Fine-Grained Actions	Audio-Visual	Test Set
Large-scale Diversity Weak/sparse supervision	Fine-grained actions Subtle variations Crowd-sourced	Audio-Visual Input Video classes only	A split of the training set
Kinetics-400, -600, -700 HowTo100M Ego4D	Charades Something-Something EPIC-KITCHENS Ego4D	Audioset VGG-Sound EPIC-KITCHENS Ego4D	All

V Patraucean et al (2022). Perception Test: A Diagnostic Benchmark for Multimodal Video Models.

The Perception Test

with: Viorica Patraucean, Joao Carreira, Andrew Zisserman
+ DeepMind Team

Is a “test” not a “test set”

V Patraucean et al (2022). Perception Test: A Diagnostic Benchmark for Multimodal Video Models.

The Perception Test

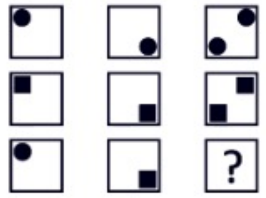
with: Viorica Patraucean, Joao Carreira, Andrew Zisserman
+ DeepMind Team

Memory

Visual discrimination
Change detection
Sequencing (order of objects, actions)
Event recall

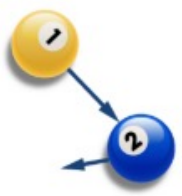
Abstraction

Object, action & event counting
Feature matching (shape, colour)
Patterns discovery
Pattern breaking



Physics (and Geometry)

Object permanence
Spatial relations and containment
Object attributes (material, size, colour)
Motion & occluded interactions
Solidity & collisions
Conservation
Stability



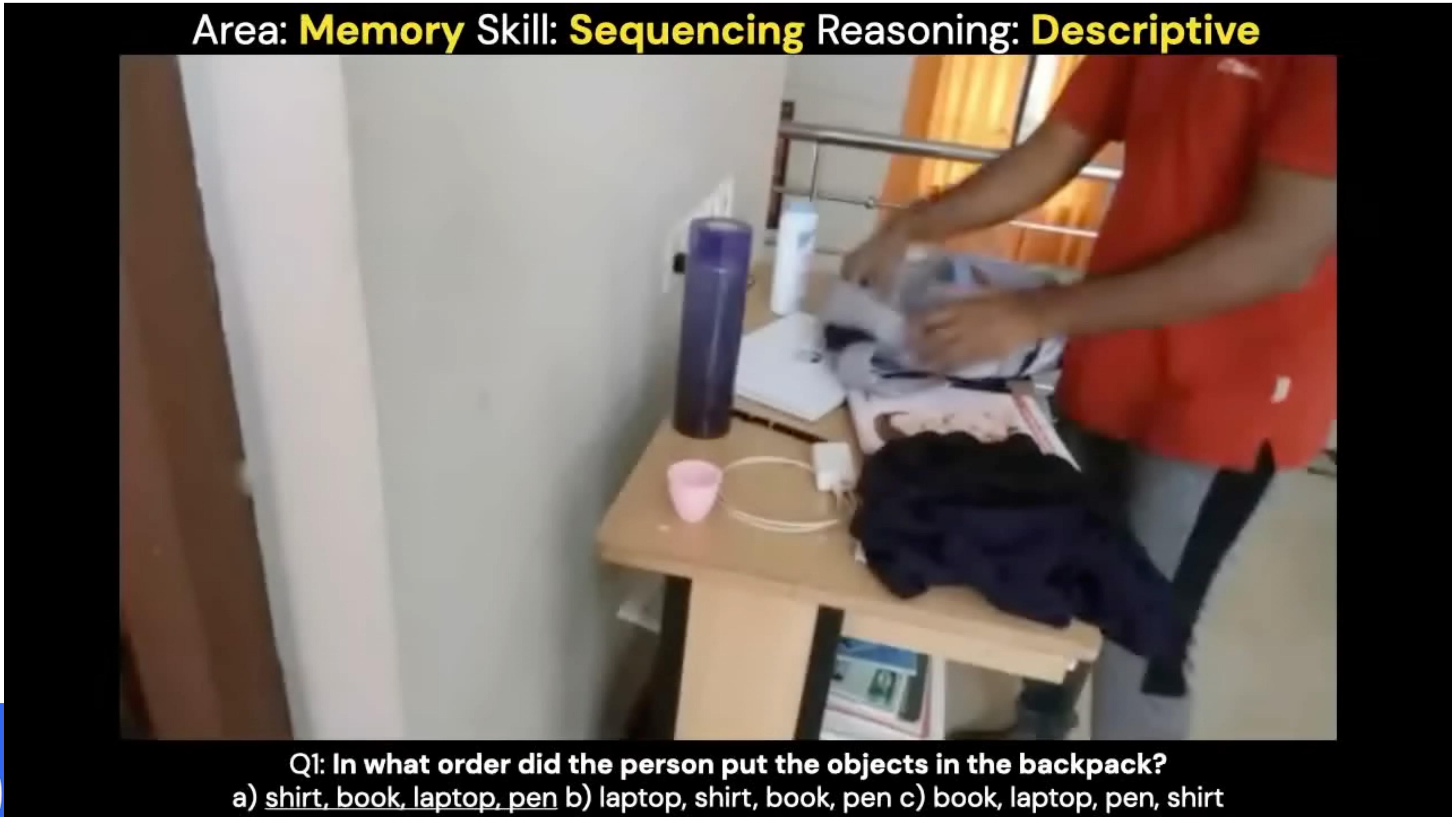
Semantics

Distractor actions & objects
Task completion & adversarial actions
Object & part recognition
Action & sound recognition
Place & state recognition
General knowledge
Language

V Patraucean et al (2022). Perception Test: A Diagnostic Benchmark for Multimodal Video Models.

The Perception Test

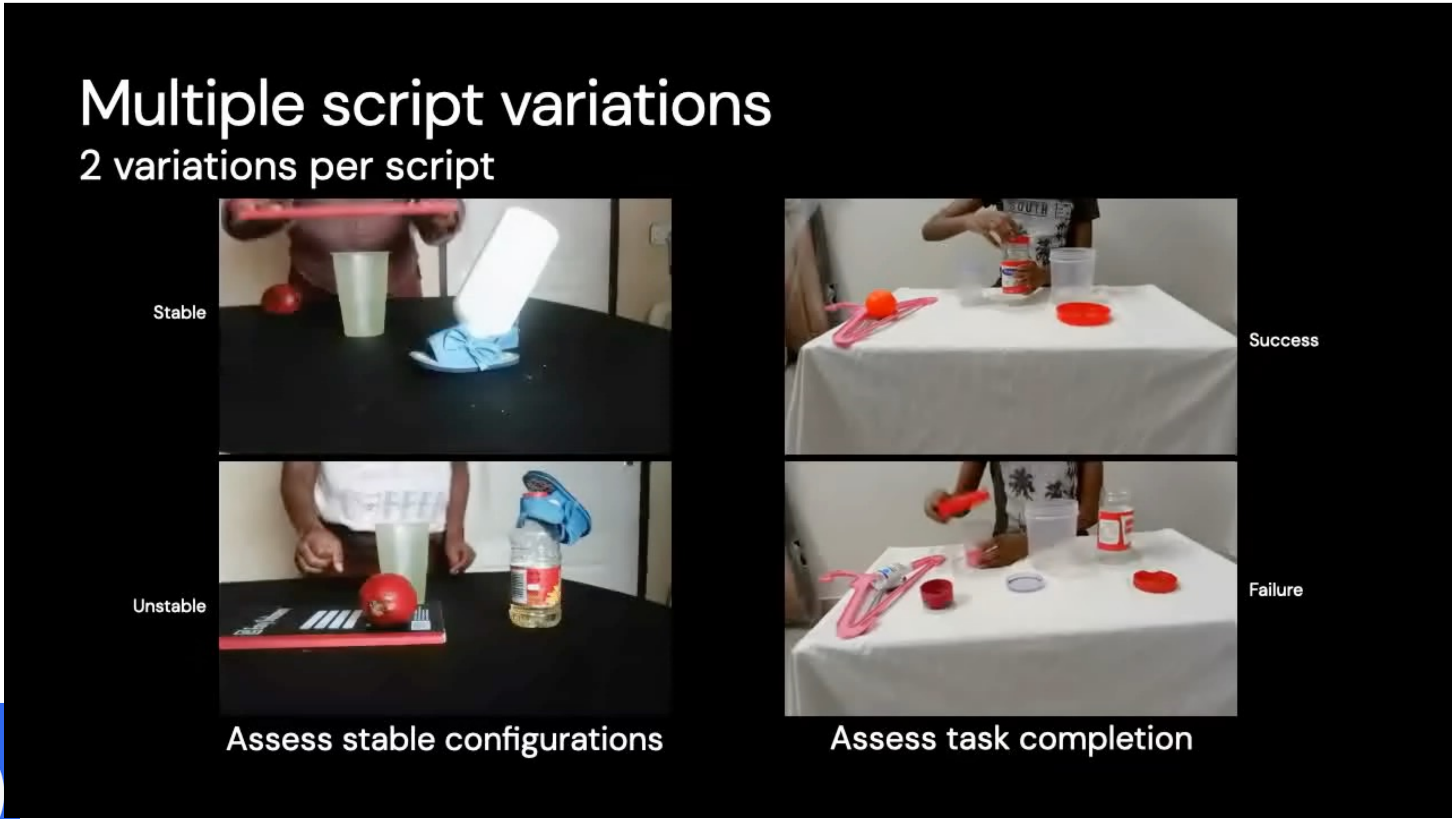
with: Viorica Patraucean, Joao Carreira, Andrew Zisserman
+ DeepMind Team



V Patraucean et al (2022). Perception Test: A Diagnostic Benchmark for Multimodal Video Models.

The Perception Test

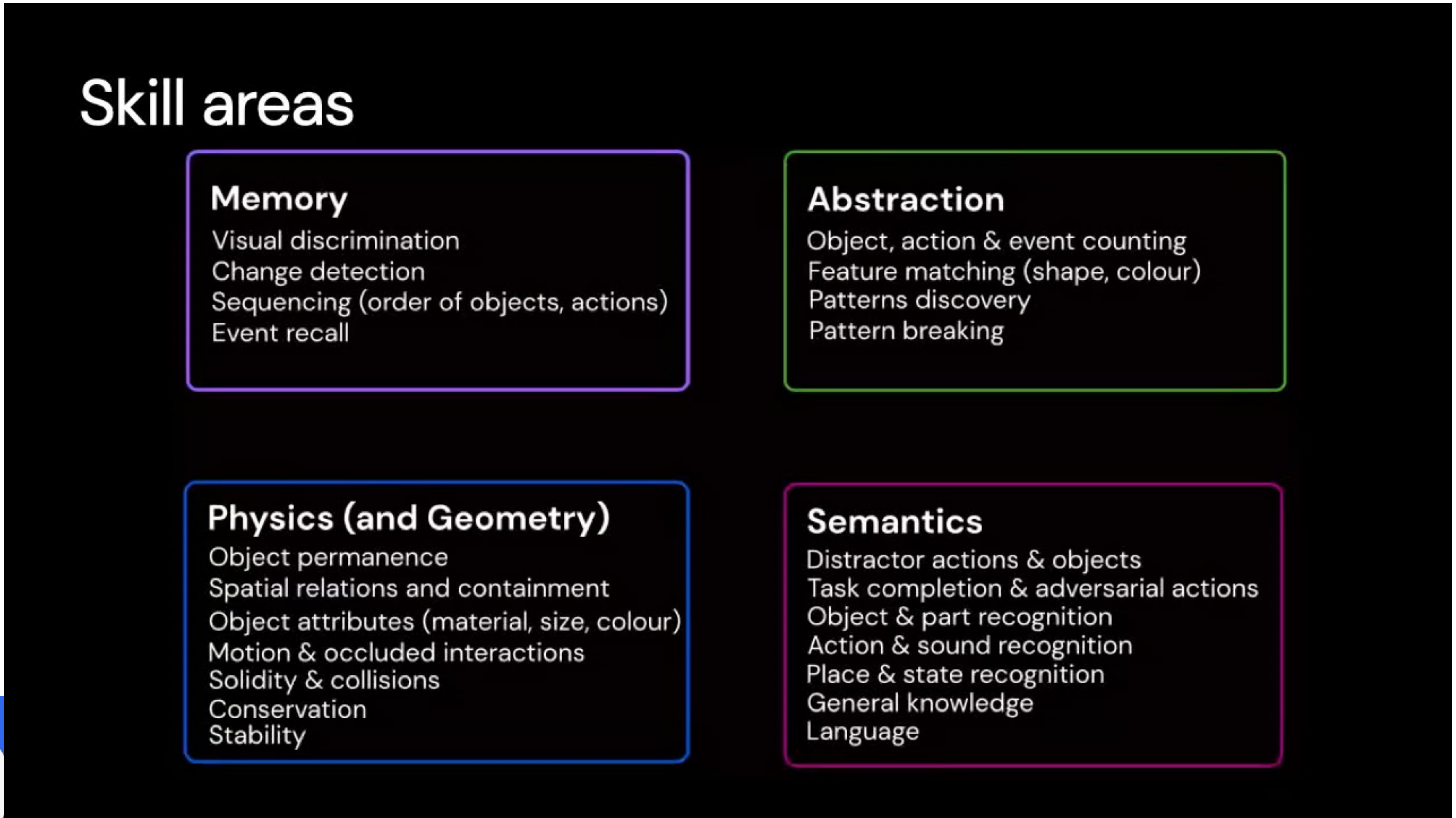
with: Viorica Patraucean, Joao Carreira, Andrew Zisserman
+ DeepMind Team



V Patraucean et al (2022). Perception Test: A Diagnostic Benchmark for Multimodal Video Models.

The Perception Test

with: Viorica Patraucean, Joao Carreira, Andrew Zisserman
+ DeepMind Team



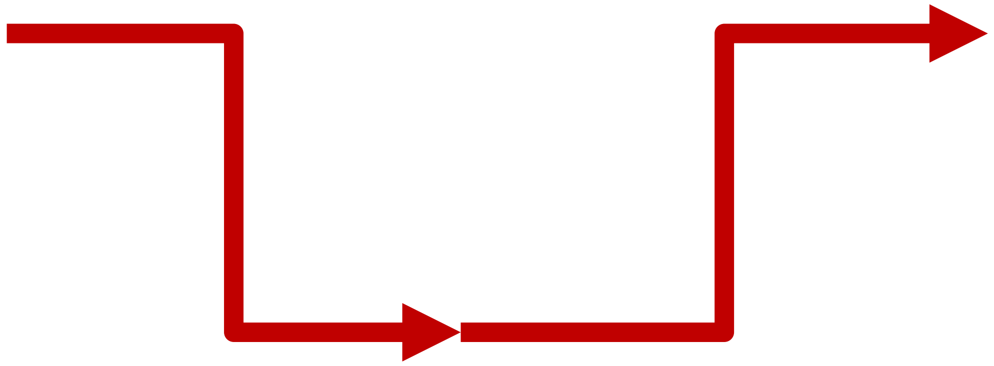
V Patraucean et al (2022). Perception Test: A Diagnostic Benchmark for Multimodal Video Models.

The Perception Test

with: Viorica Patraucean, Joao Carreira, Andrew Zisserman
+ DeepMind Team

V Patraucean et al (2022). Perception Test: A Diagnostic Benchmark for Multimodal Video Models.

And back…



Multi-Modality in Egocentric Data

V — High frame-rate RGB footage from the camera wearer's perspective

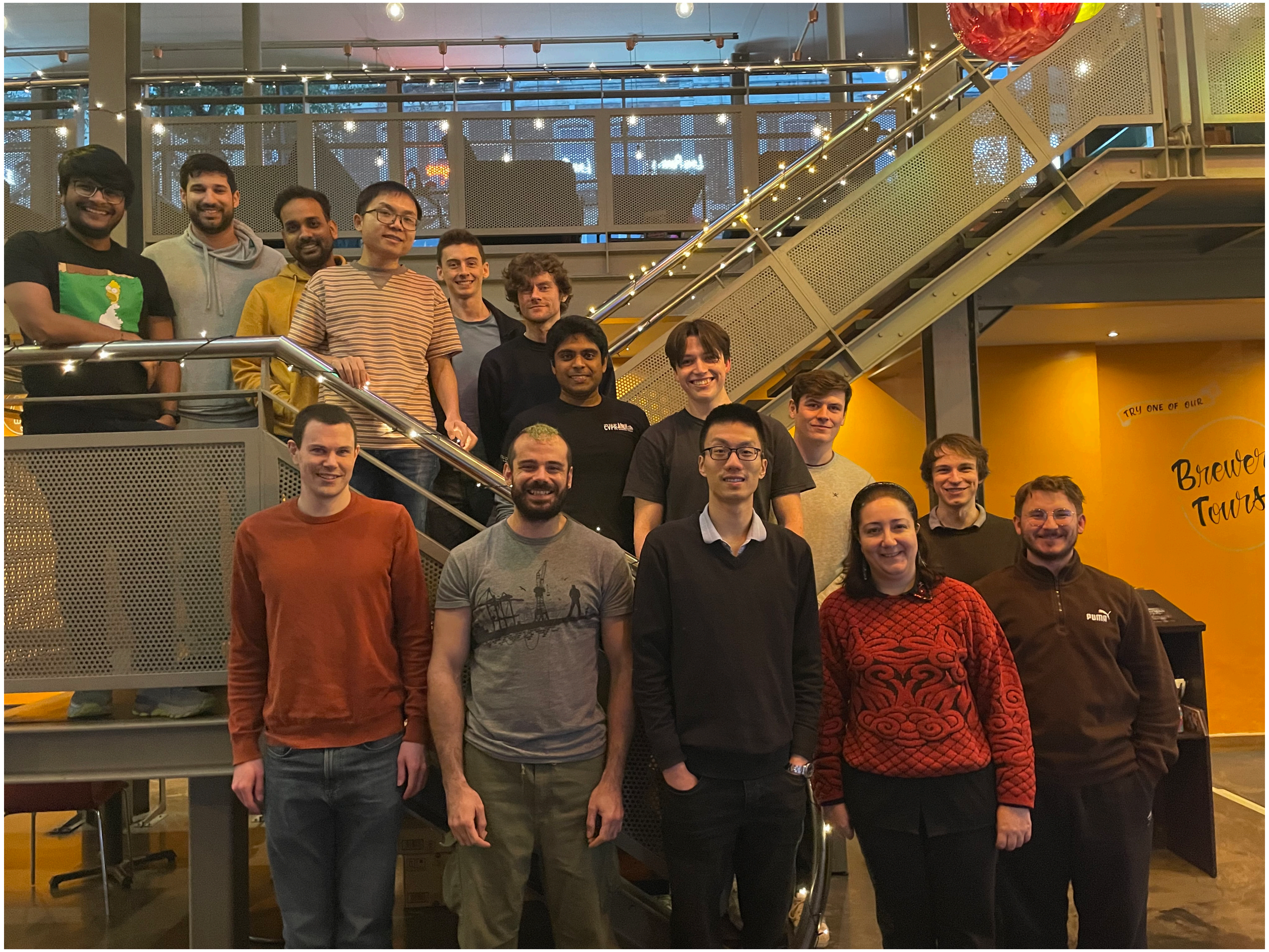
A — One or many microphones, on the wearable device, best positioned to capture the sounds of actions and interactions

L — Speech in the video... or
Narrations/Captions added to index the videos

University of BRISTOL

Dima Damen
MULA@CVPR2024

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