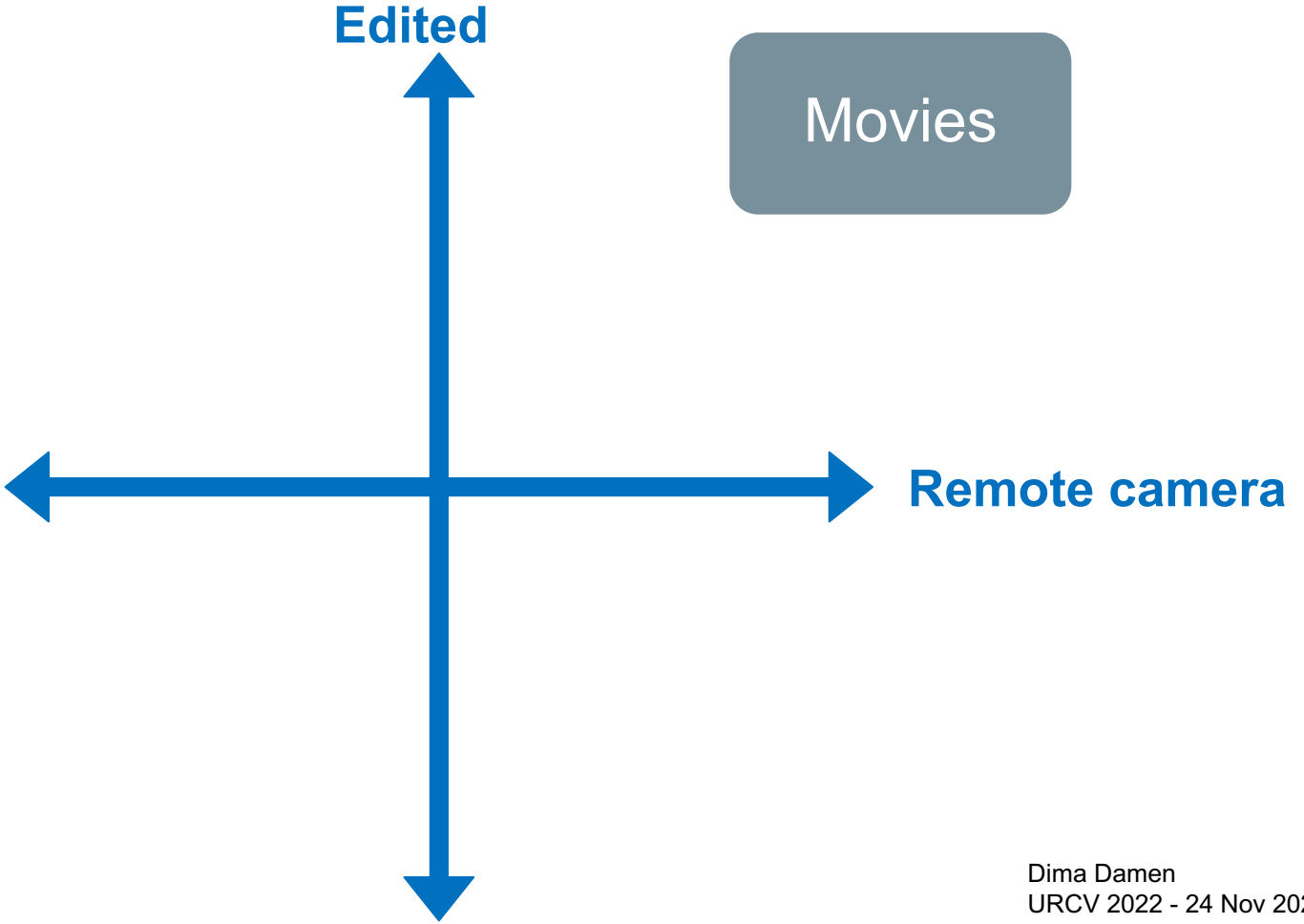
Representing Egocentric Videos

The history of VIDEO Understanding



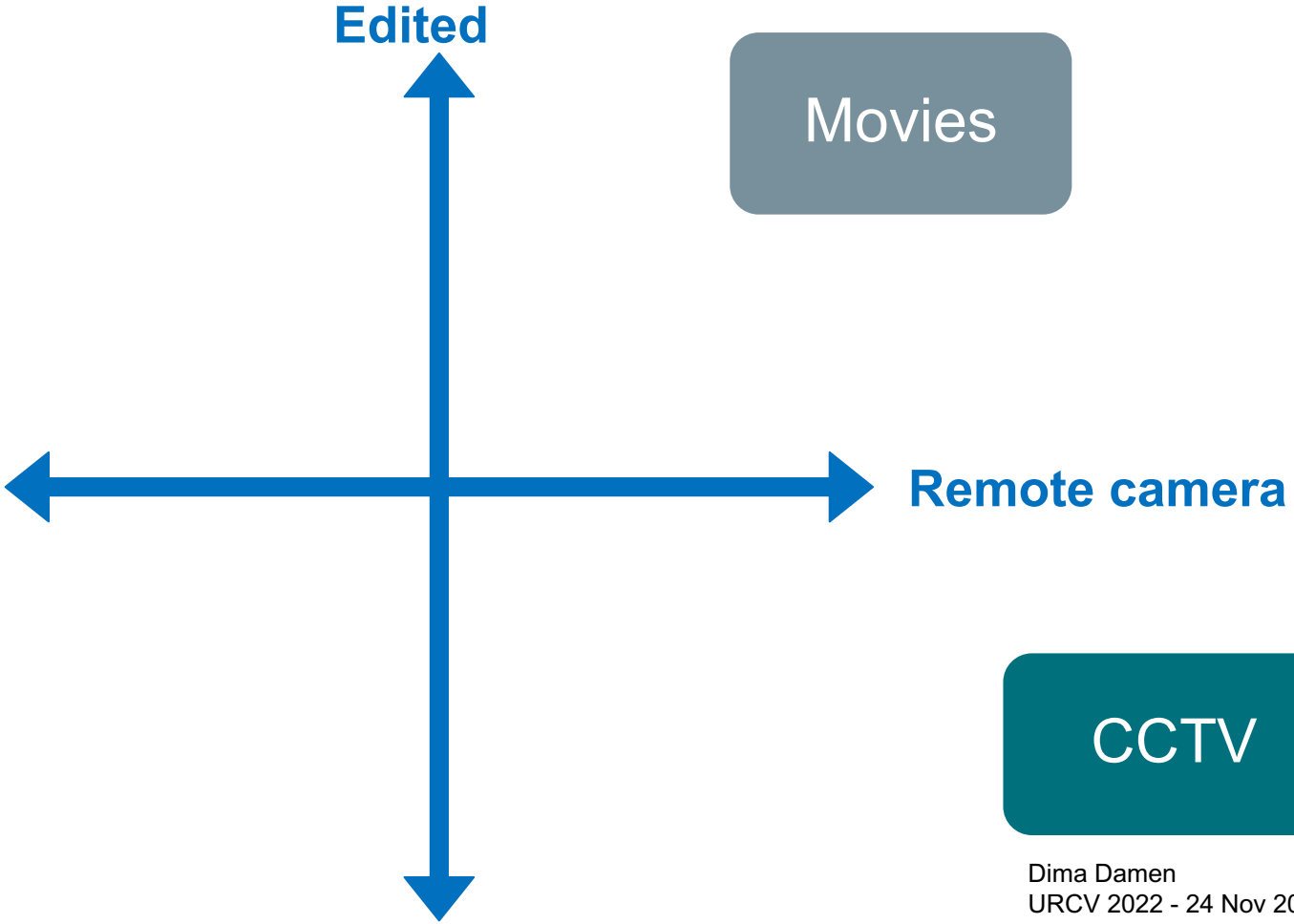
The history of VIDEO Understanding

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

Laptev and Perez (2007)

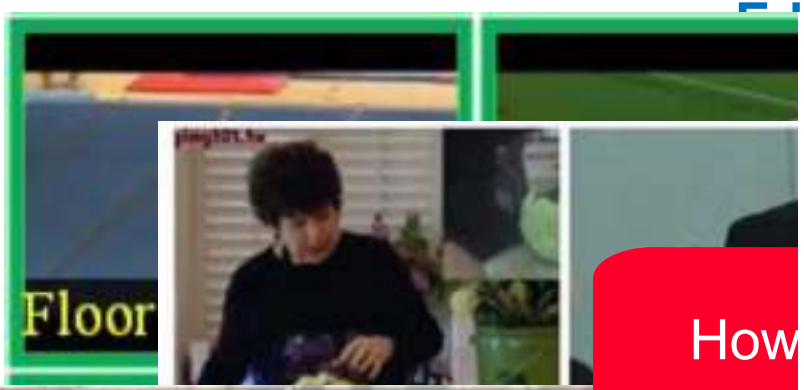
The history of VIDEO Understanding

The history of VIDEO Understanding



The history of VIDEO Understanding

The history of VIDEO Understanding



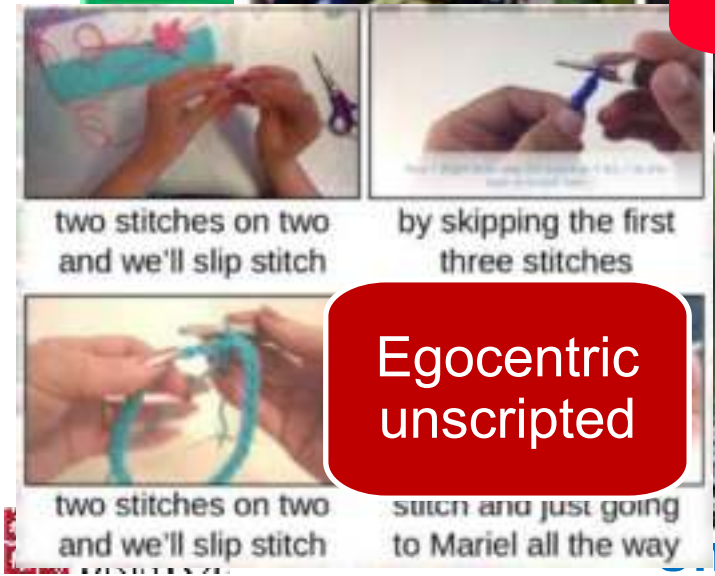
How

Templated, Multilingual Domain Queries:

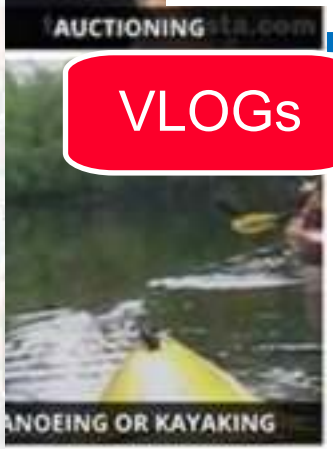
"Morning routine", "realistic ditl 2015", "mijn realistische routine", "Ma routine d'apres-midi", …

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

Remote camera



Egocentric unscripted



VLOGs

YouTube Videos

CCTV

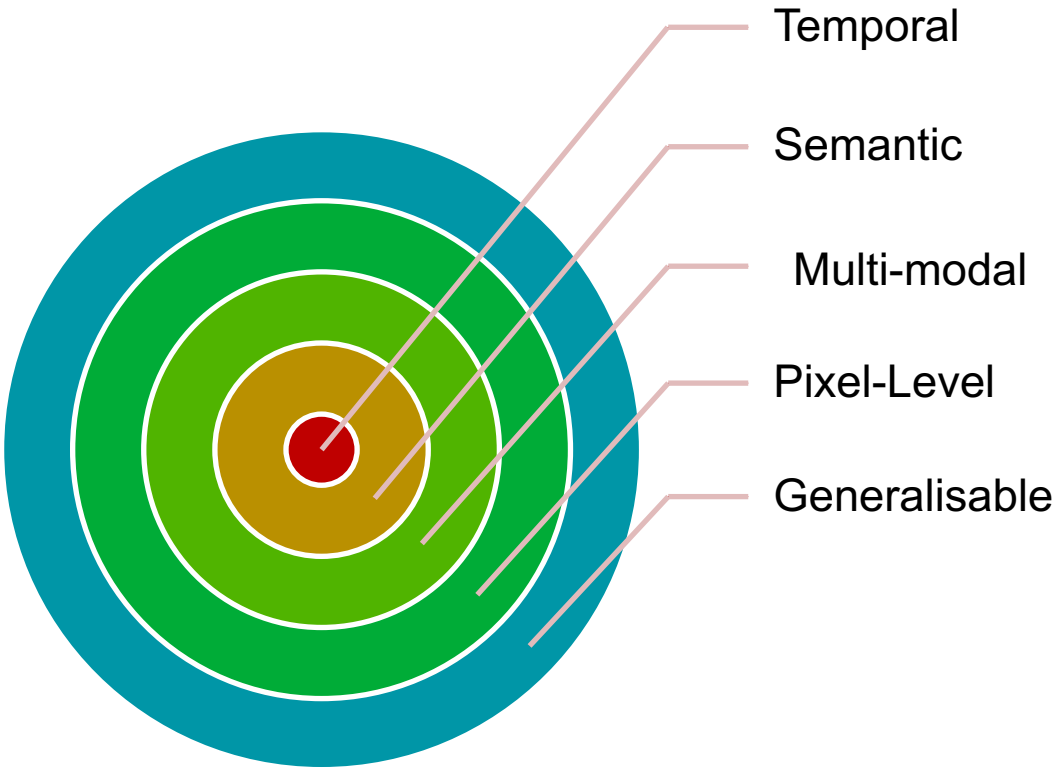
Unedited

Egocentric Videos?

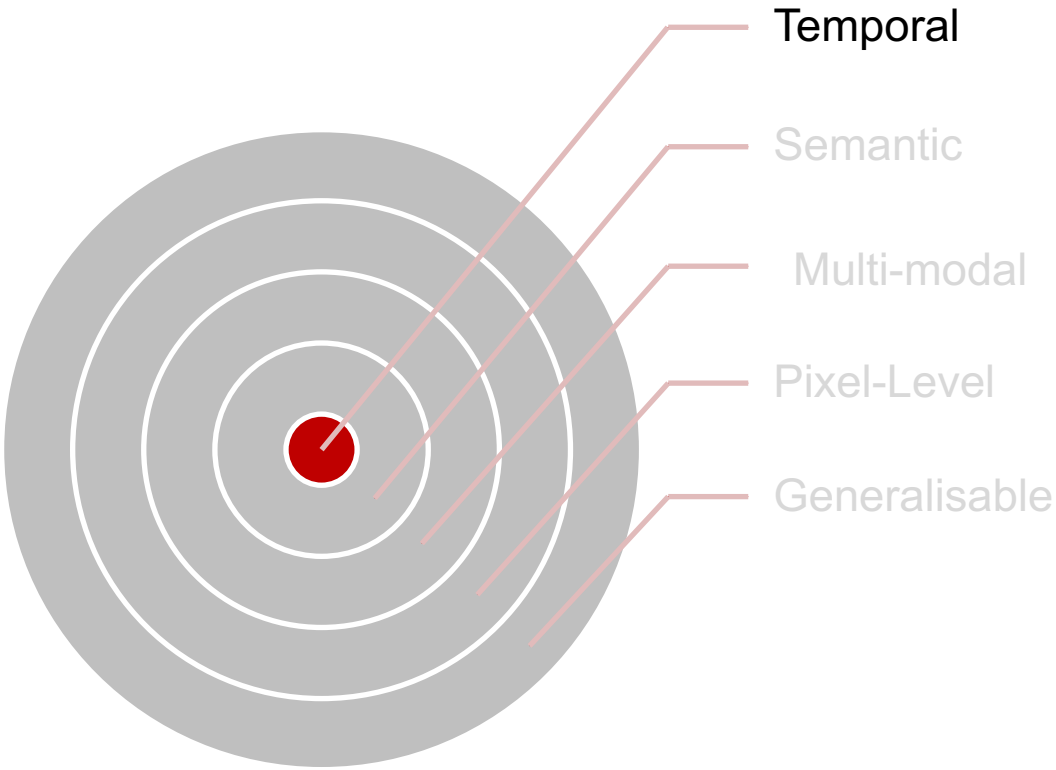
Egocentric Videos?


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

Representing Egocentric Actions



Representing Egocentric Actions




Do
It
Yourself


13
73
131
56
93
177
13
56
73
93
131
177


13
73
131
56
93
177
13
56
73
93
131
177

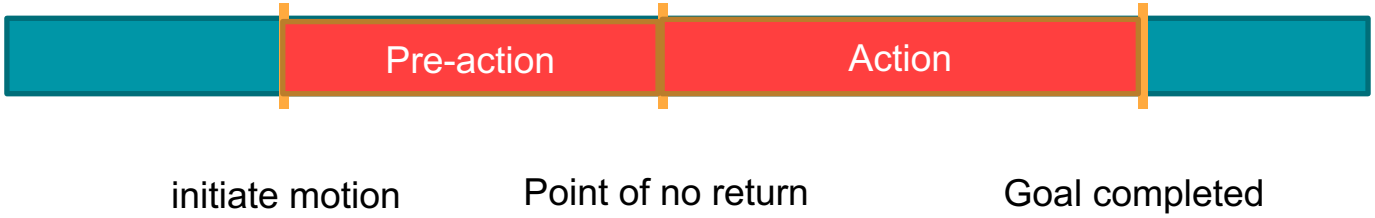

13
73
131
56
93
177
13
56
73
93
131
177


13
73
131
56
93
177
13
56
73
93
131
177


How to achieve consistent temporal labels?

The Rubicon Boundaries

- [A] There are two stages of an action, separated by three boundary points
  - Pre-action stage:
  - Action stage:



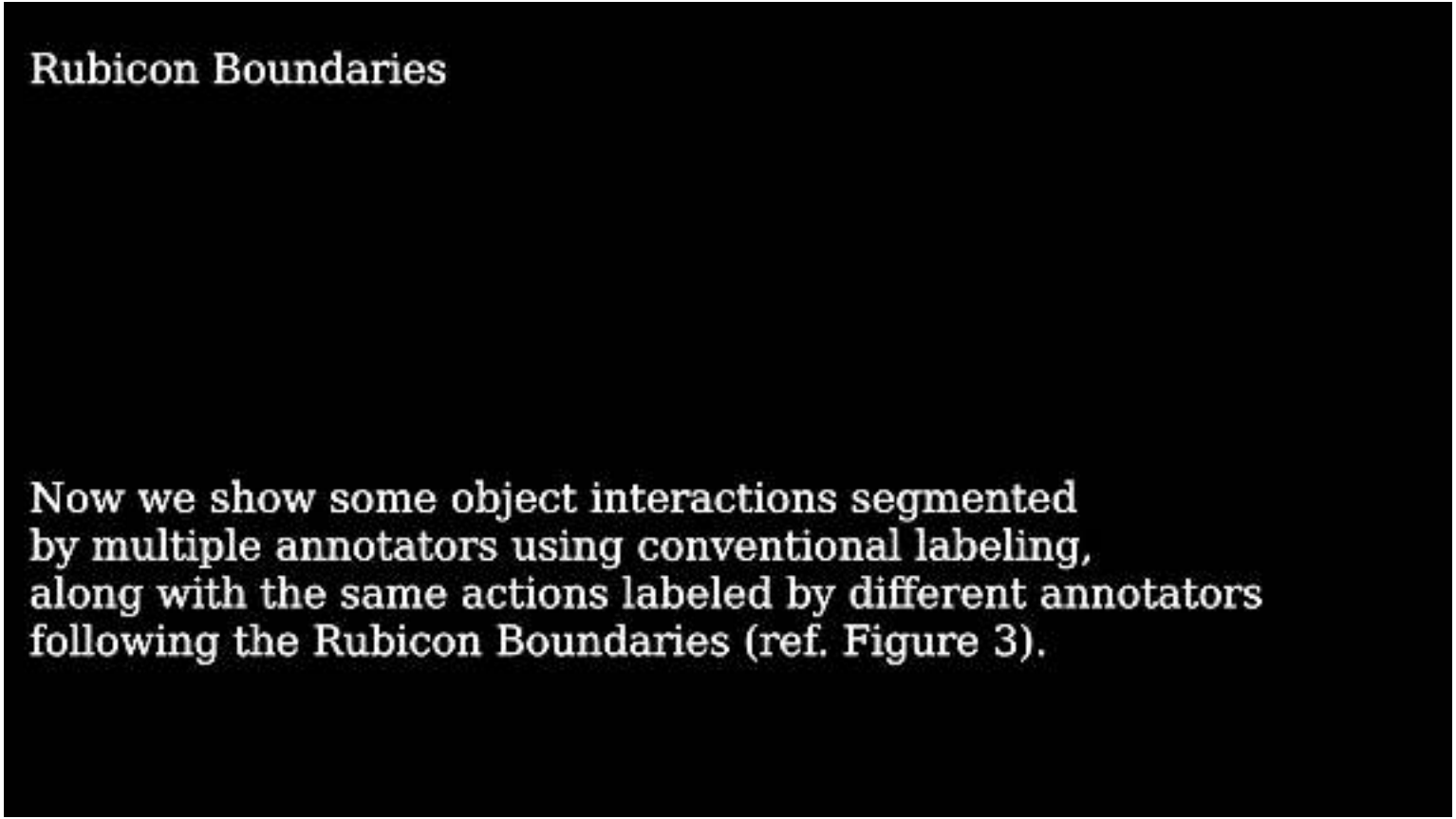
[A] P. M. Gollwitzer (1990). Action phases and mind-sets. Handbook of motivation and cognition.

The Rubicon Boundaries

D Moltisanti et al (2017). Trespassing the Boundaries: Labeling Temporal Bounds for Object Interactions in Egocentric Video. *International Conference on Computer Vision (ICCV).*

The Rubicon Boundaries

with: Davide Moltisanti

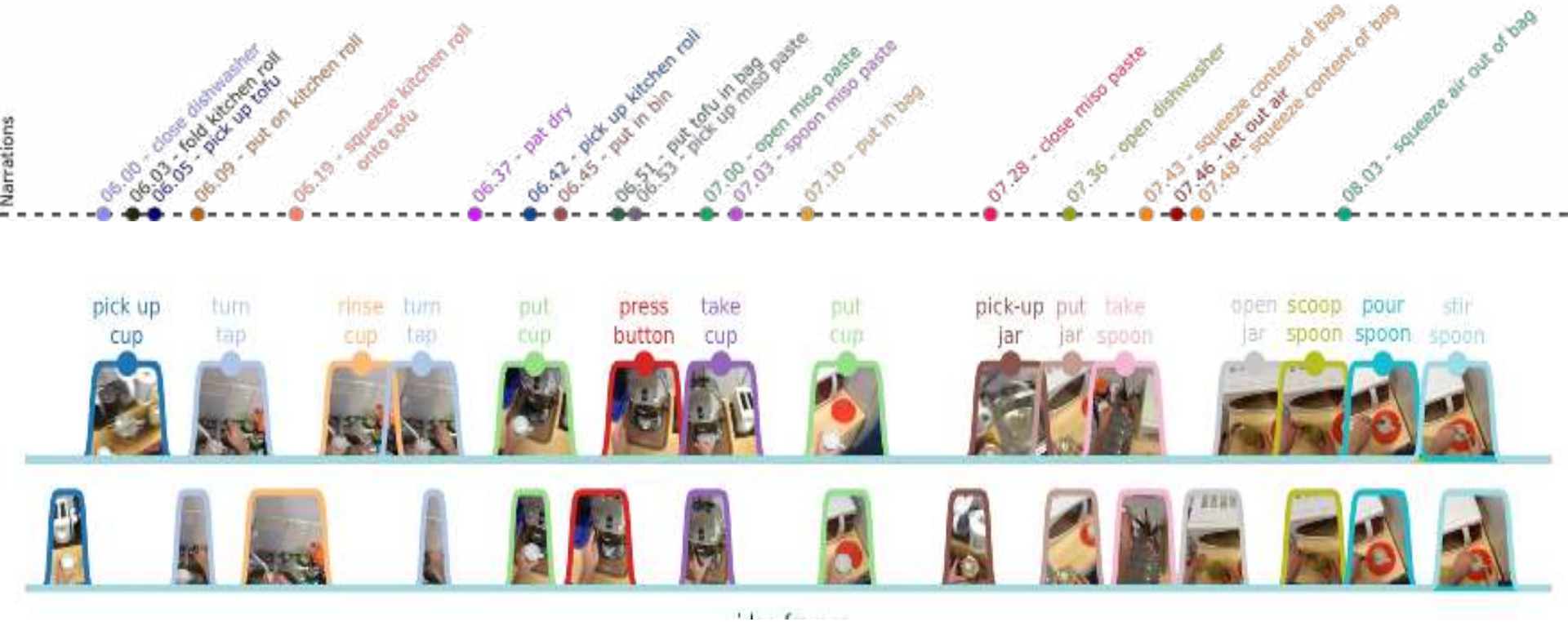


D Moltisanti et al (2017). Trespassing the Boundaries: Labeling Temporal Bounds for Object Interactions in Egocentric Video. *International Conference on Computer Vision (ICCV).*

The power of temporal labels

Learning from a Single Timestamp

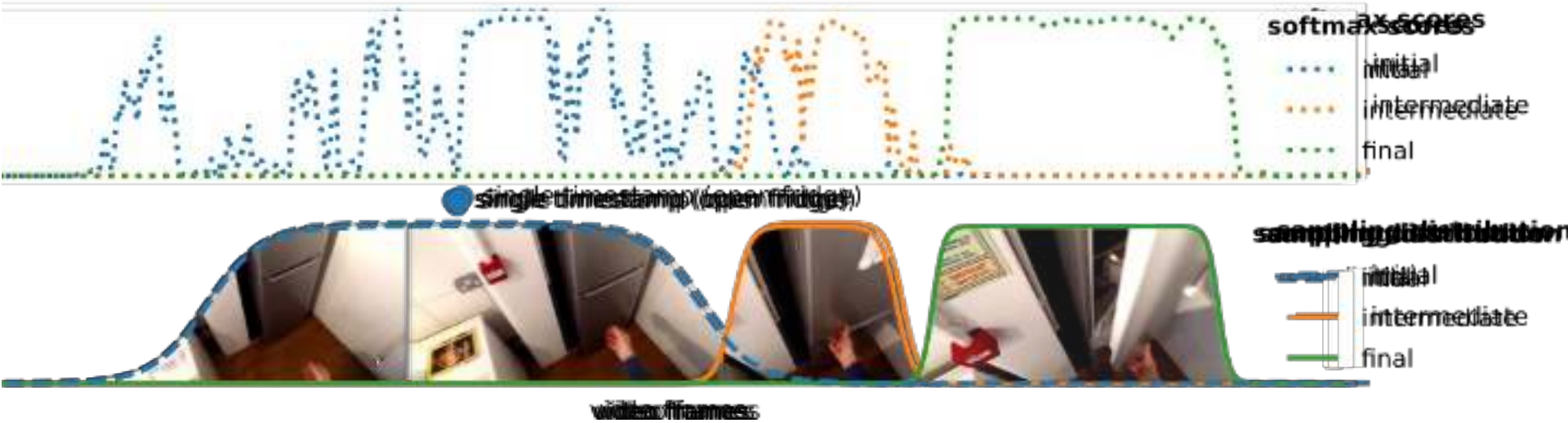
with: Davide Moltisanti
Sanja Fidler



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

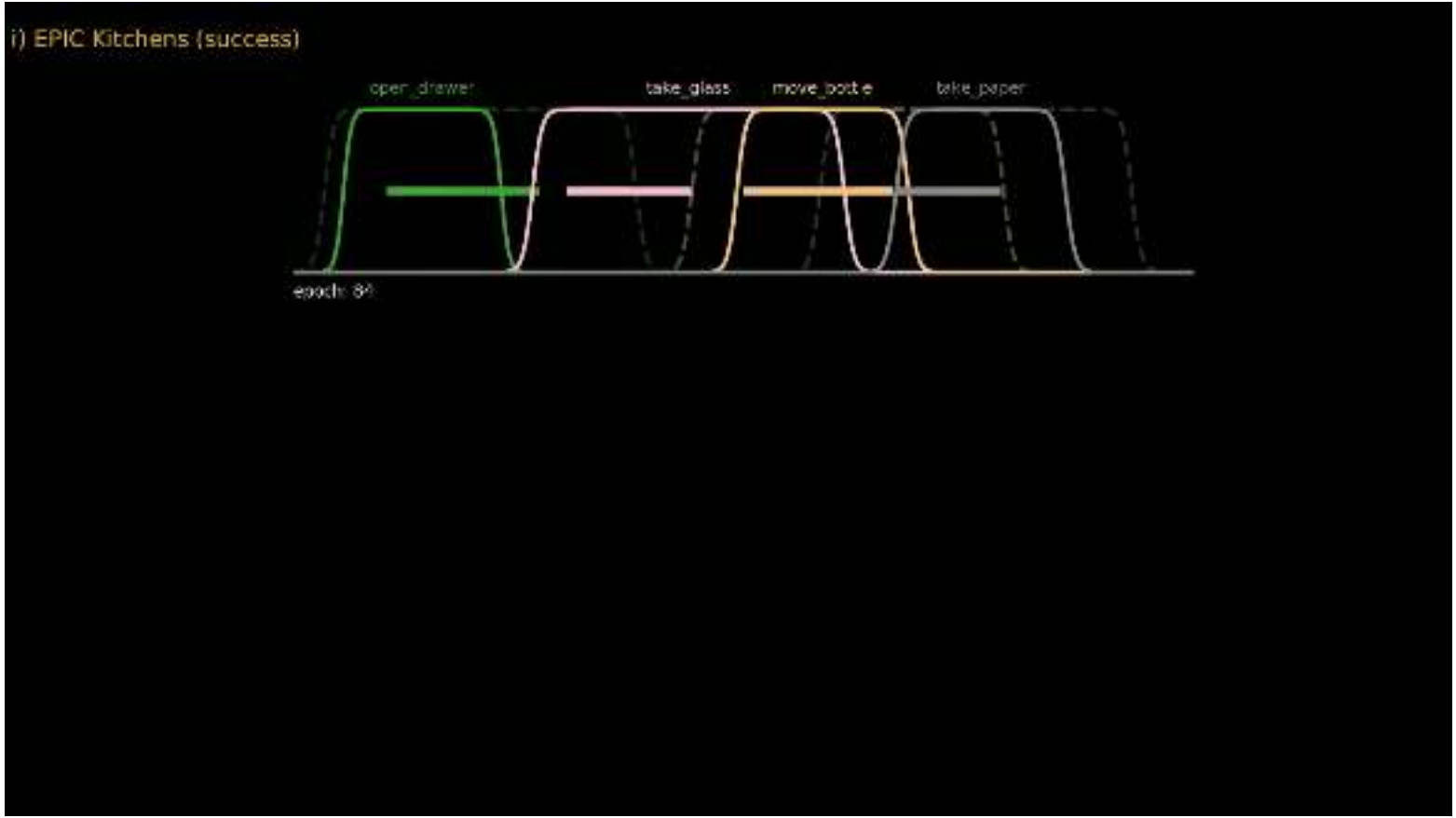
Learning from a Single Timestamp

with: Davide Moltisanti
Sanja Fidler



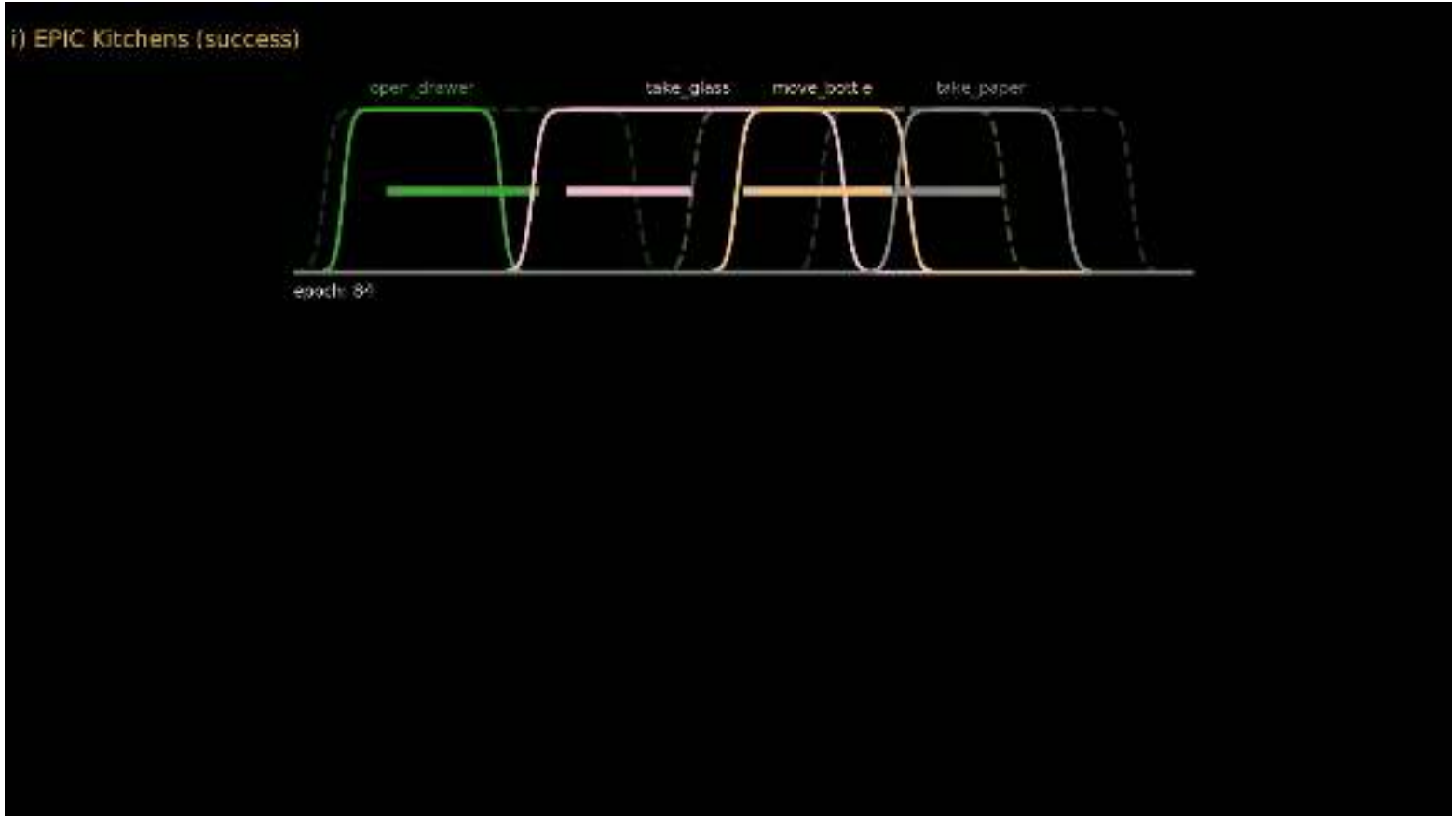
Learning from a Single Timestamp

with: Davide Moltisanti
Sanja Fidler



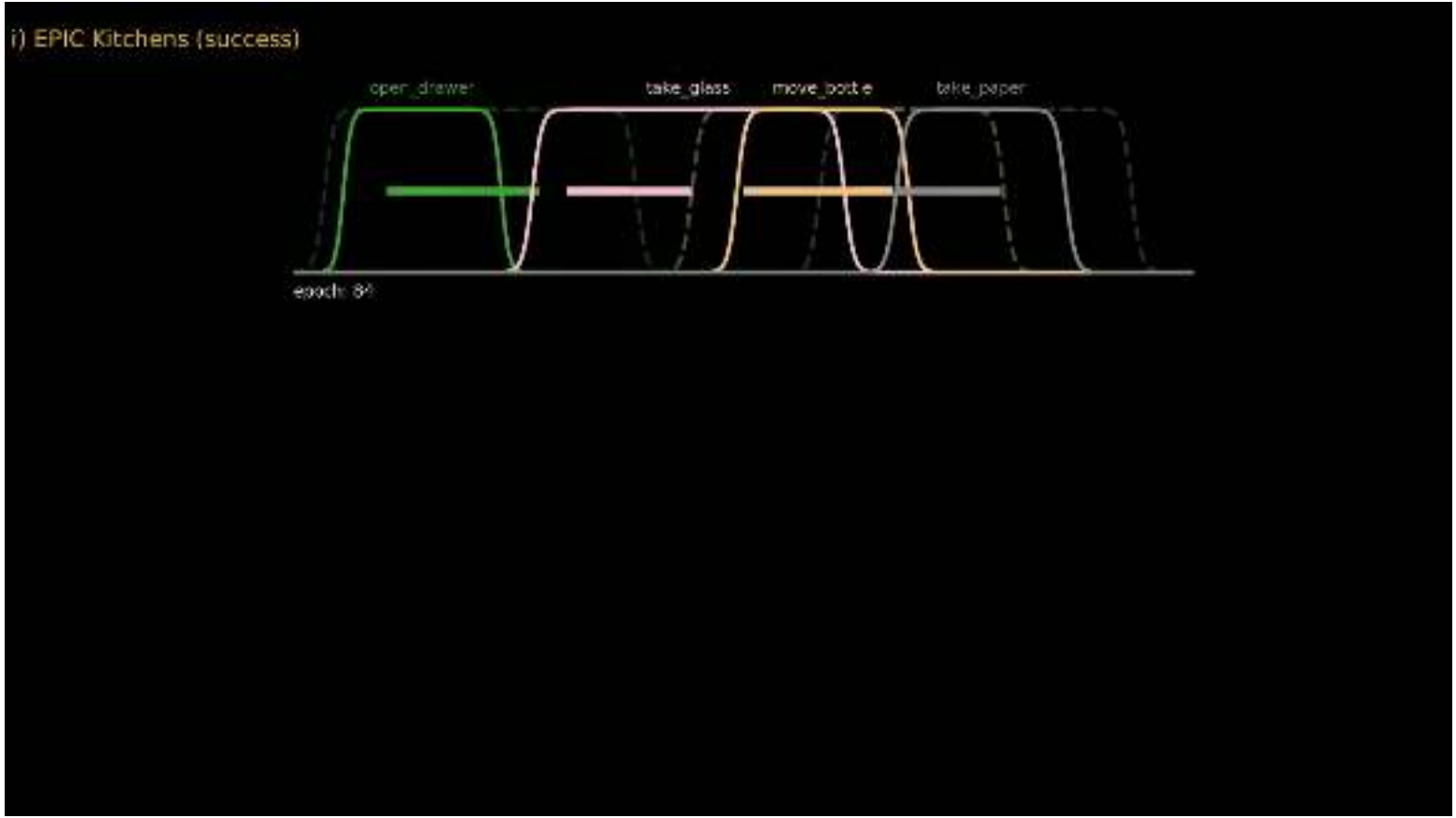
Learning from a Single Timestamp

with: Davide Moltisanti
Sanja Fidler



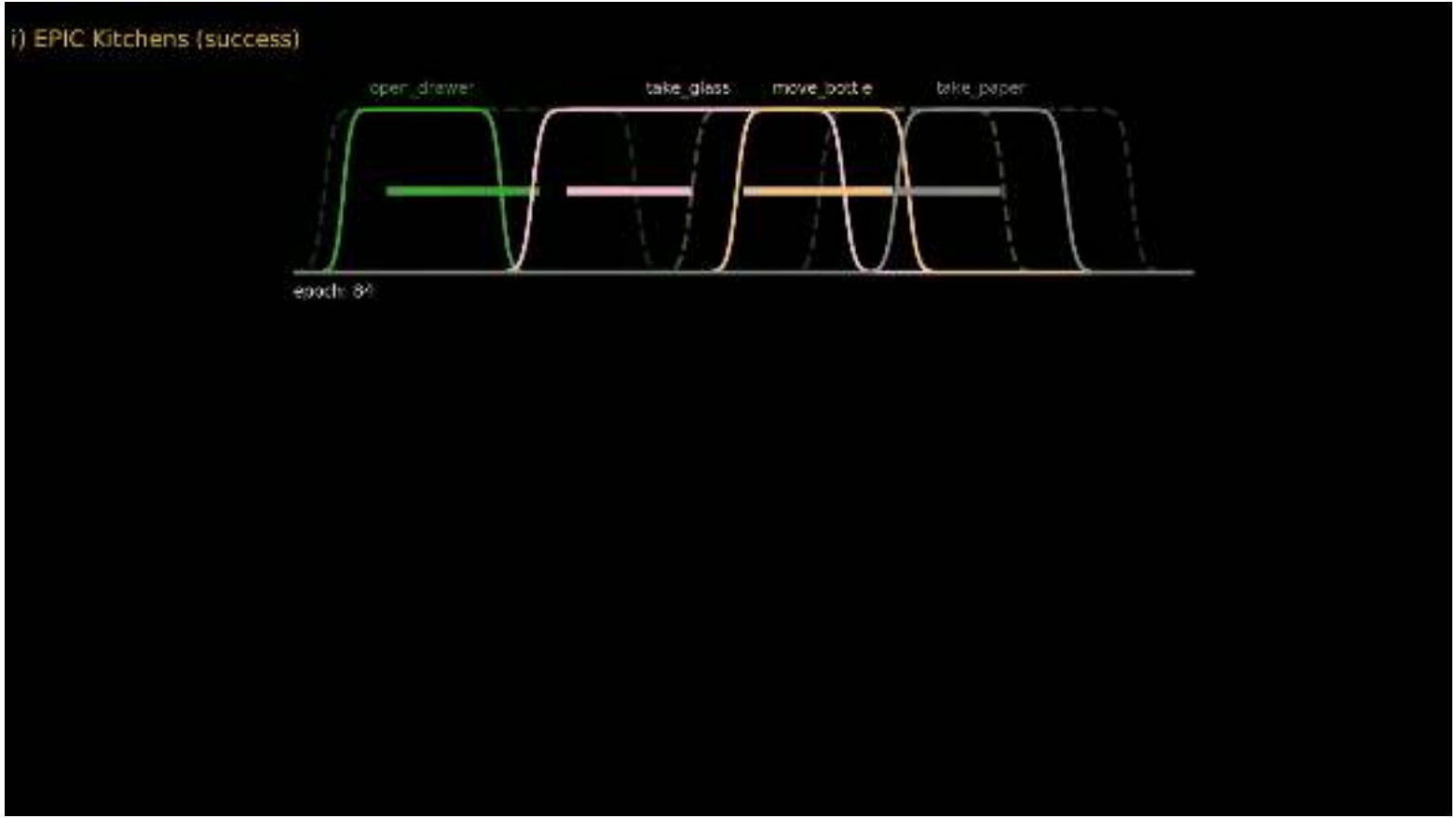
Learning from a Single Timestamp

with: Davide Moltisanti
Sanja Fidler



Learning from a Single Timestamp

with: Davide Moltisanti
Sanja Fidler



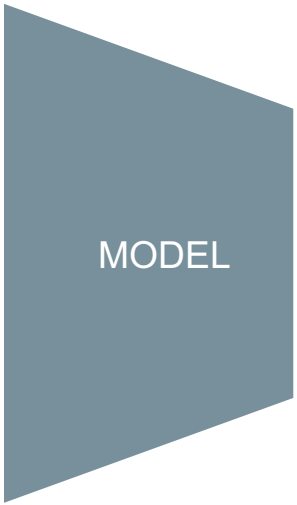
D Moltisanti et al (2019). Action Recognition from Single Timestamp Supervision in Untrimmed Videos. *Computer Vision and Pattern Recognition (CVPR)*

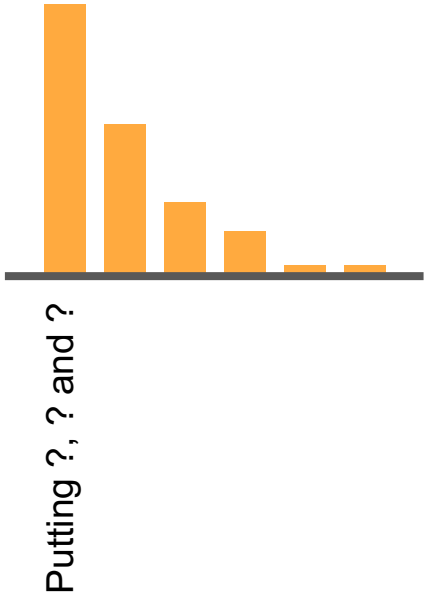
What about models?

How do models see frames?

Frame Attributions in Video Models

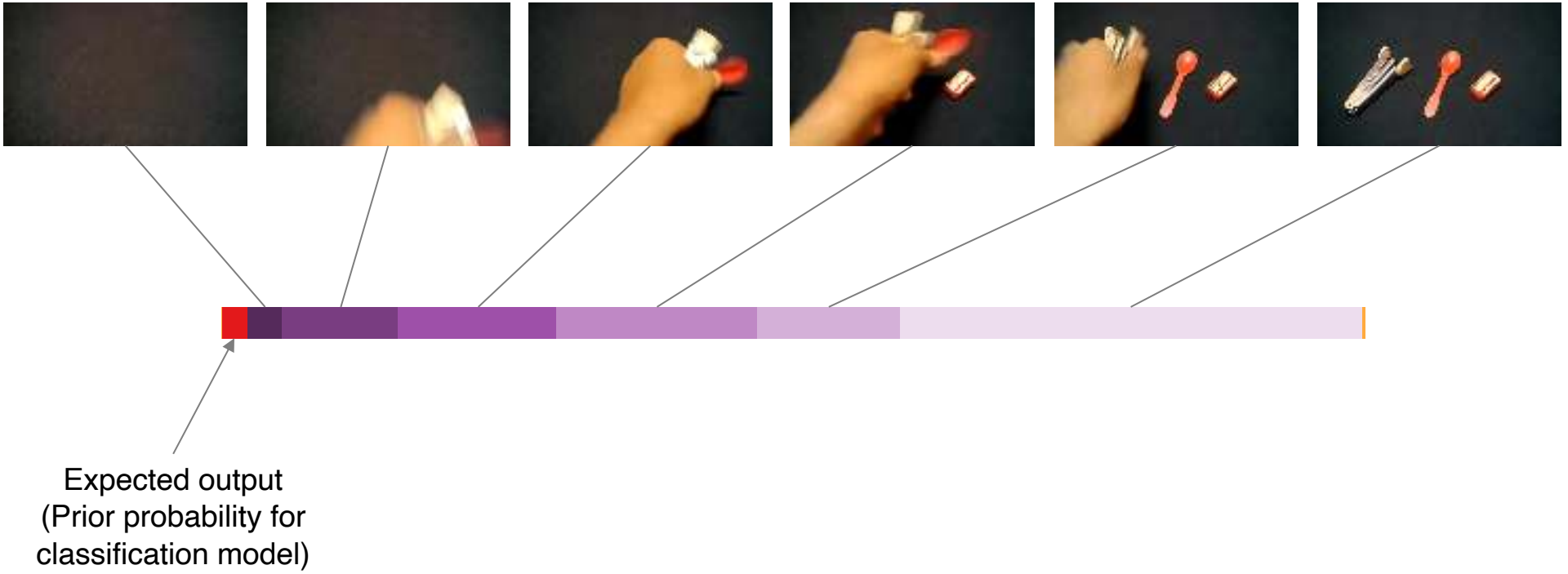
with: Will Price





Frame Attributions in Video Models

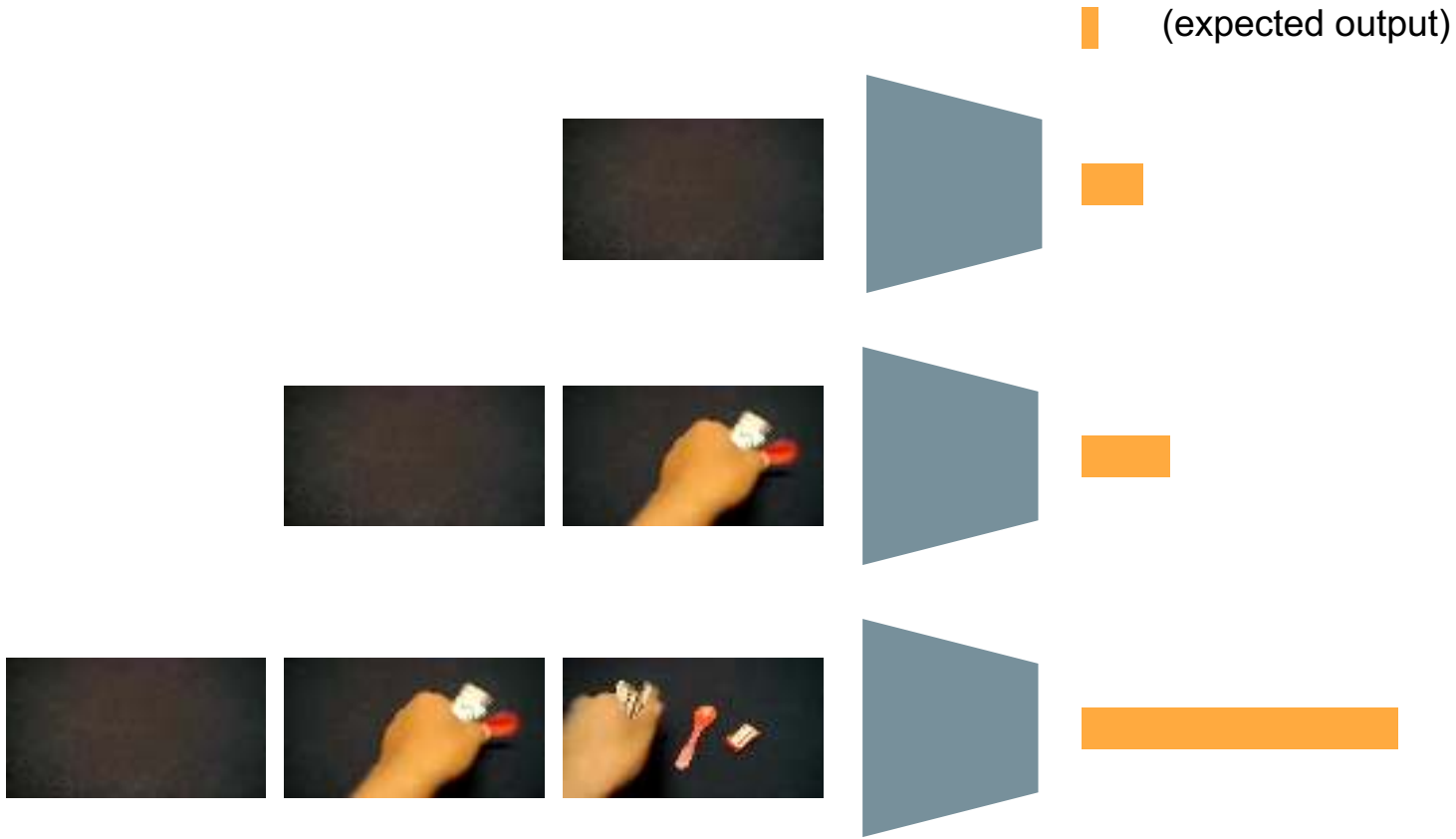
with: Will Price



W Price, D Damen (2020). Play Fair: Frame Attribution in Video Models. Asian Conference on Computer Vision (ACCV)

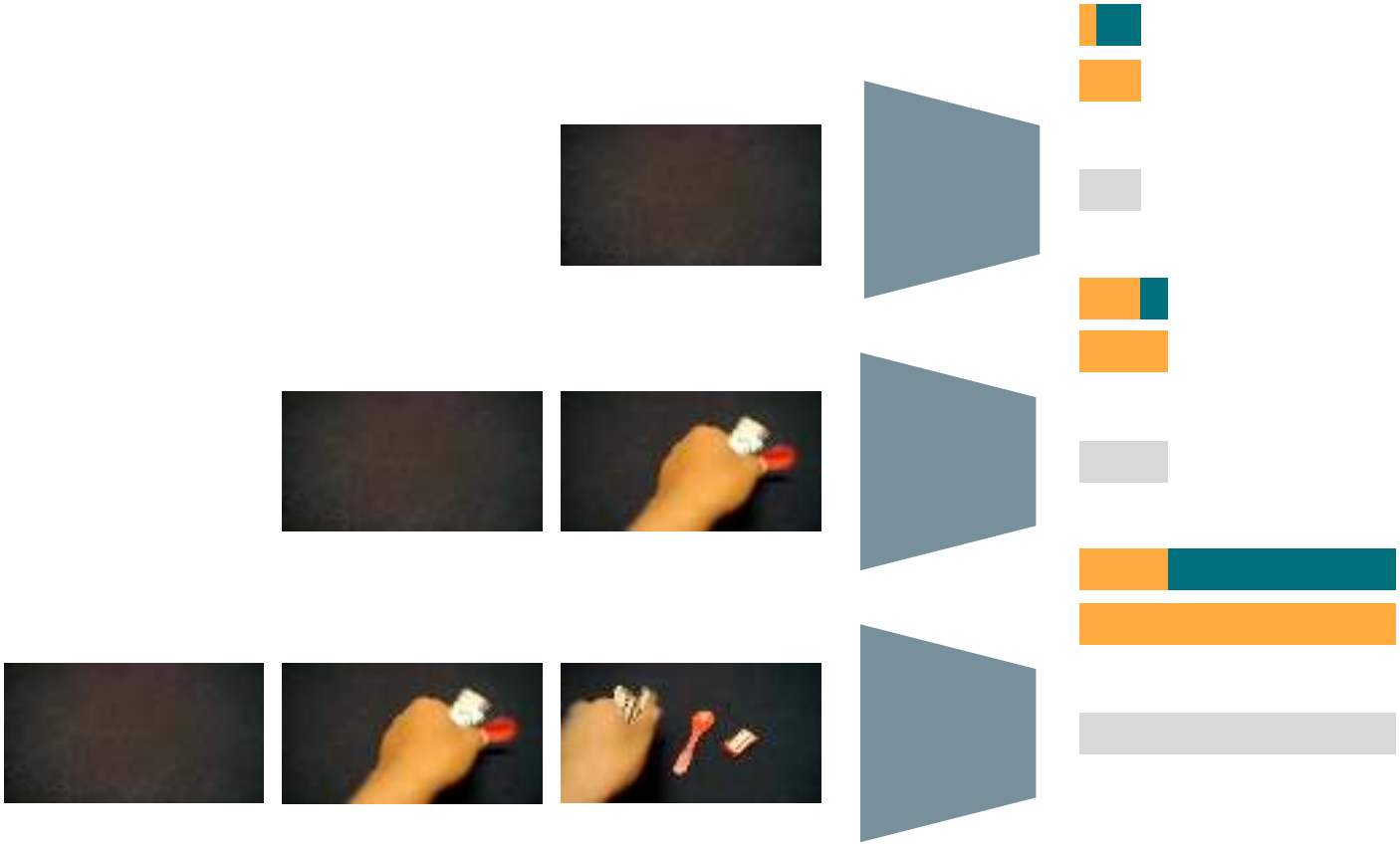
Frame Attributions in Video Models

with: Will Price



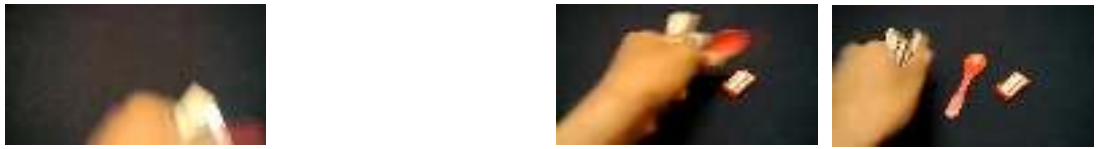
Frame Attributions in Video Models

with: Will Price



Frame Attributions in Video Models

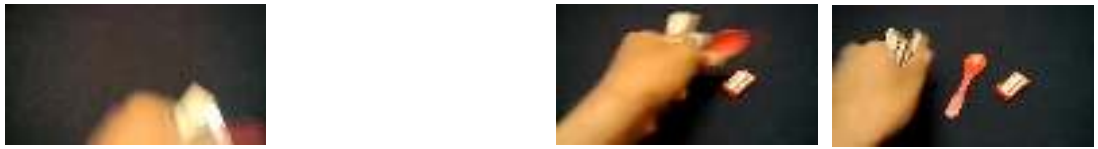
with: Will Price



Frame Attributions in Video Models

with: Will Price

MODEL



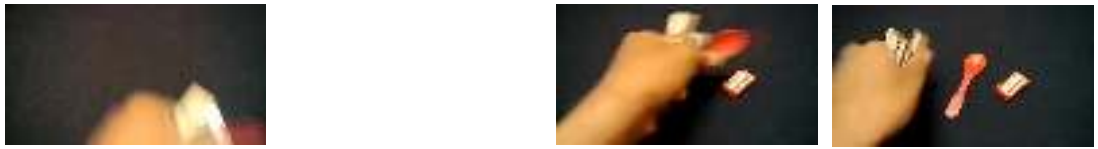
MODEL

Frame Attributions in Video Models

with: Will Price

MODEL

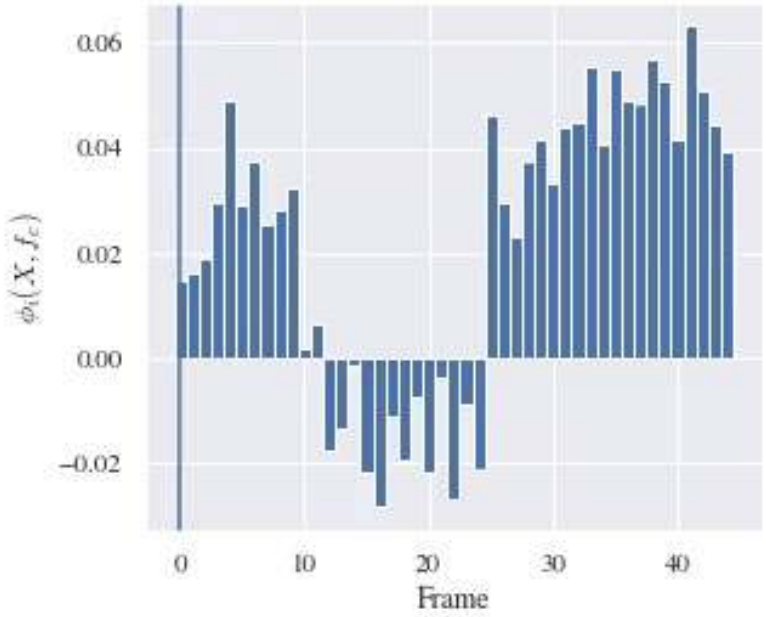
$\Delta_3(\{1,2,4,5\}) = -.2$



MODEL

Frame Attributions in Video Models

with: Will Price



Frame Attributions in Video Models

with: Will Price

Dashboard

Frame Attributions in Video Models

with: Will Price
Tom Stark

ESVs Dashboard for Epic

Select a verb

Select a noun

Select a video

Select number of frames

Model Predictions

ESV Predictions

W Price, D Damen (2020). Play Fair: Frame Attribution in Video Models. Asian Conference on Computer Vision (ACCV)

Frame Attributions in Video Models

with: Will Price
Tom Stark

ESVs Dashboard for Epic

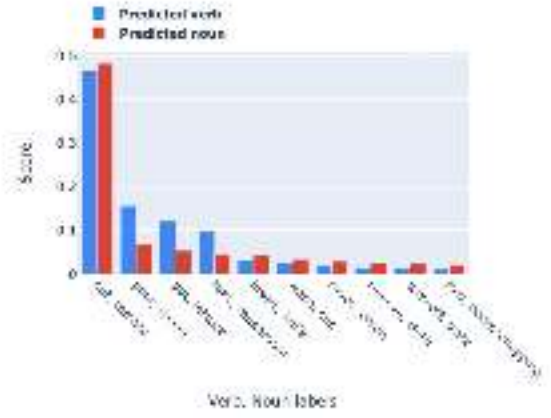
Select a verb

Select a noun

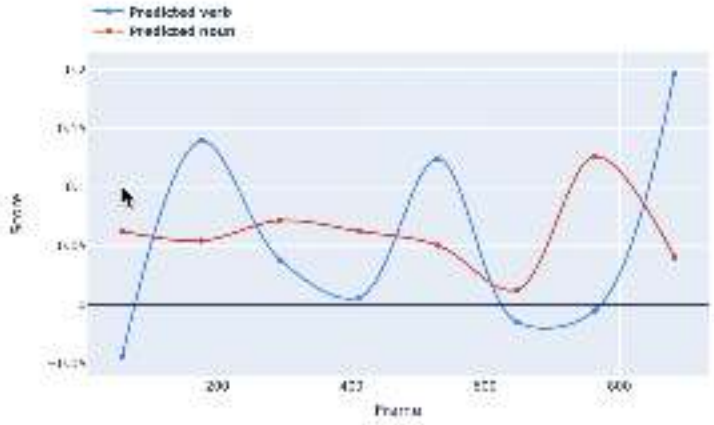
Select a video

Select number of frames

Model Predictions

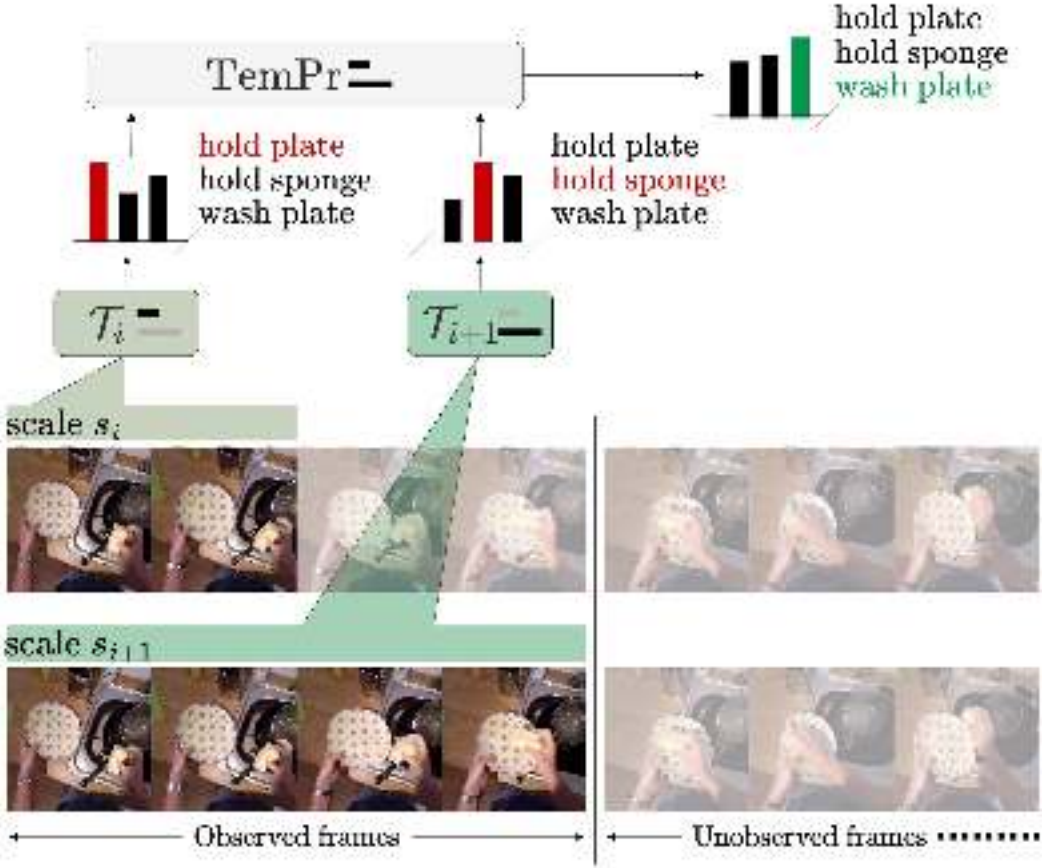


ESV Predictions



Early Action Prediction

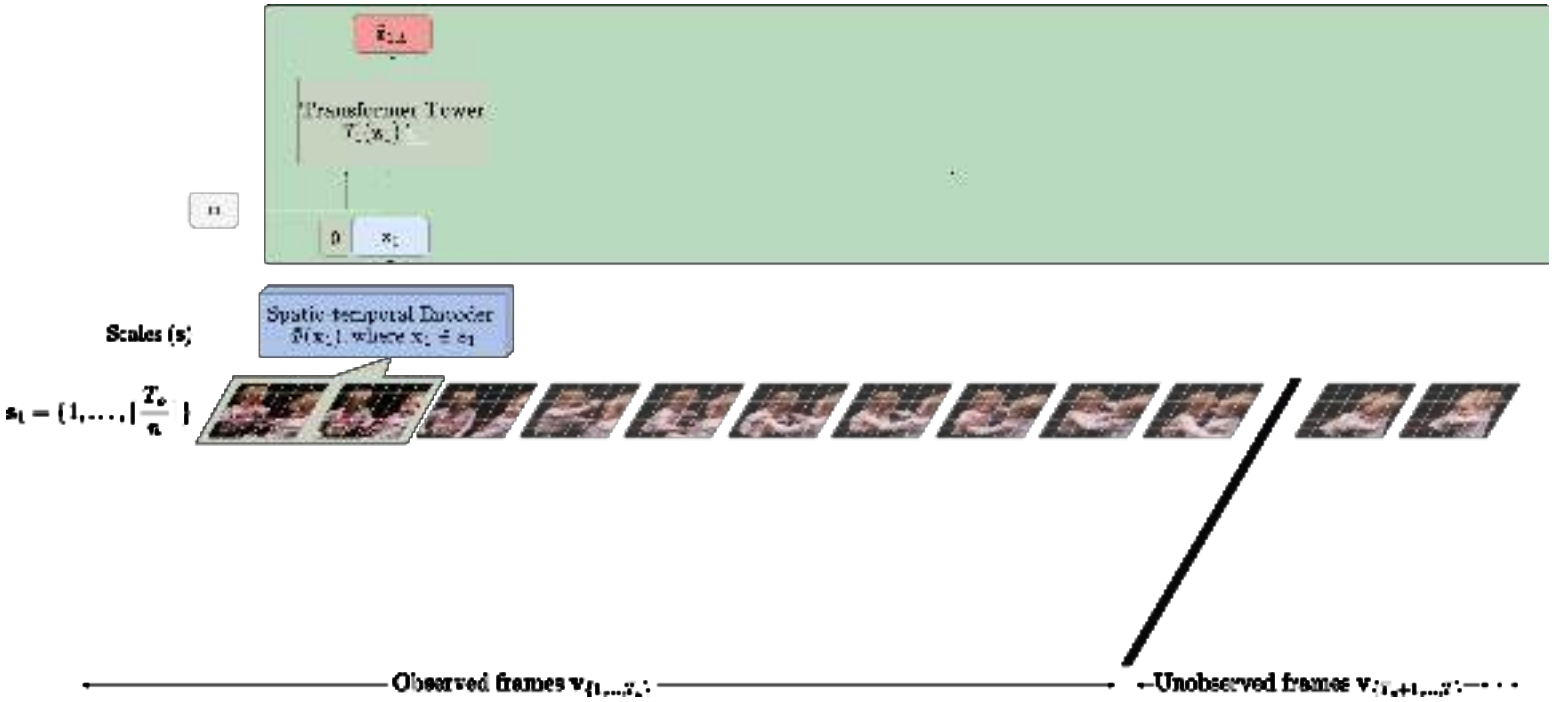
with: Alex Stergiou



A Stergiou and D Damen (2022). Temporal Progressive Attention for Early Action Prediction. ArXiv

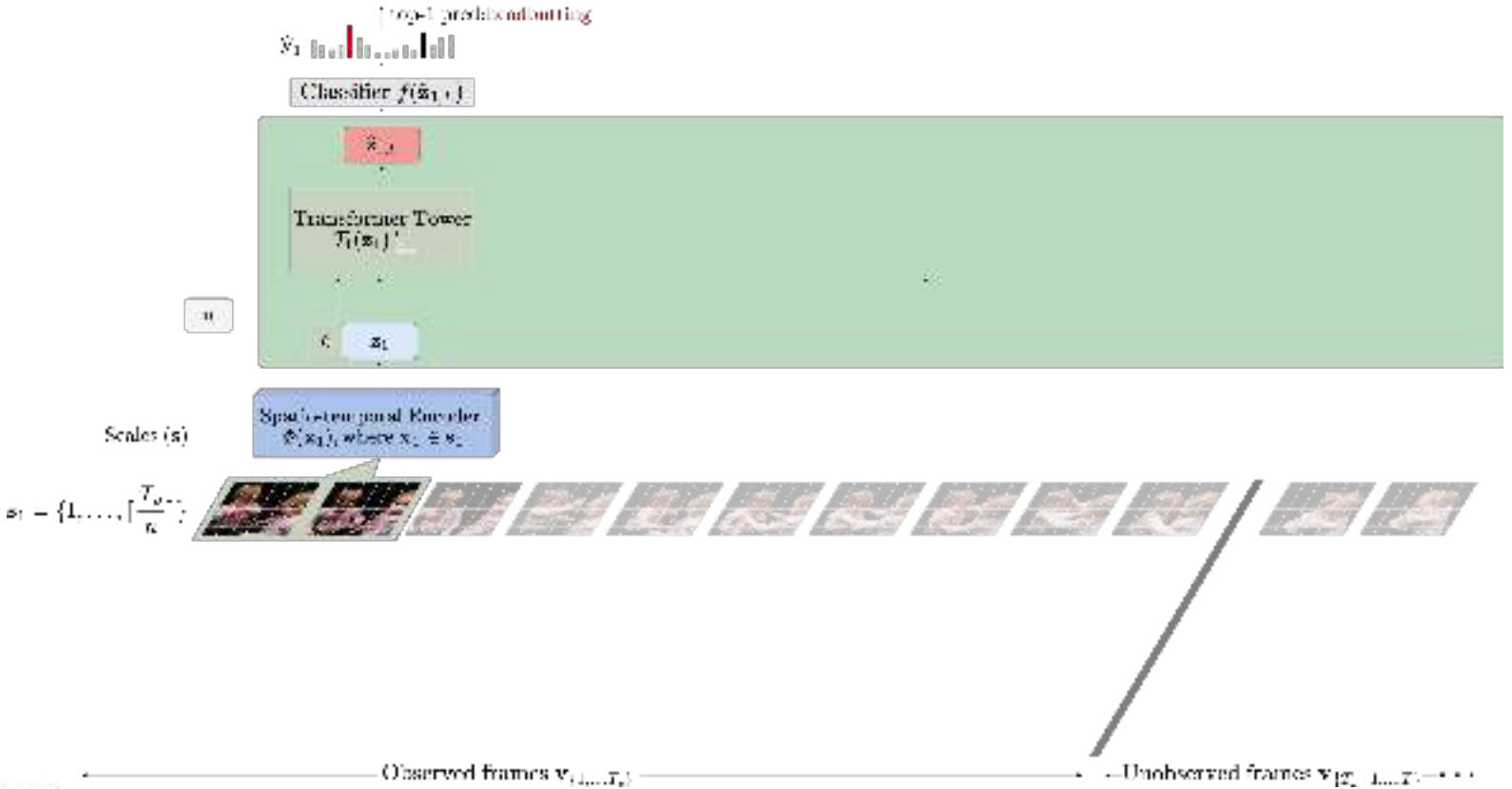
Early Action Prediction

with: Alex Stergiou



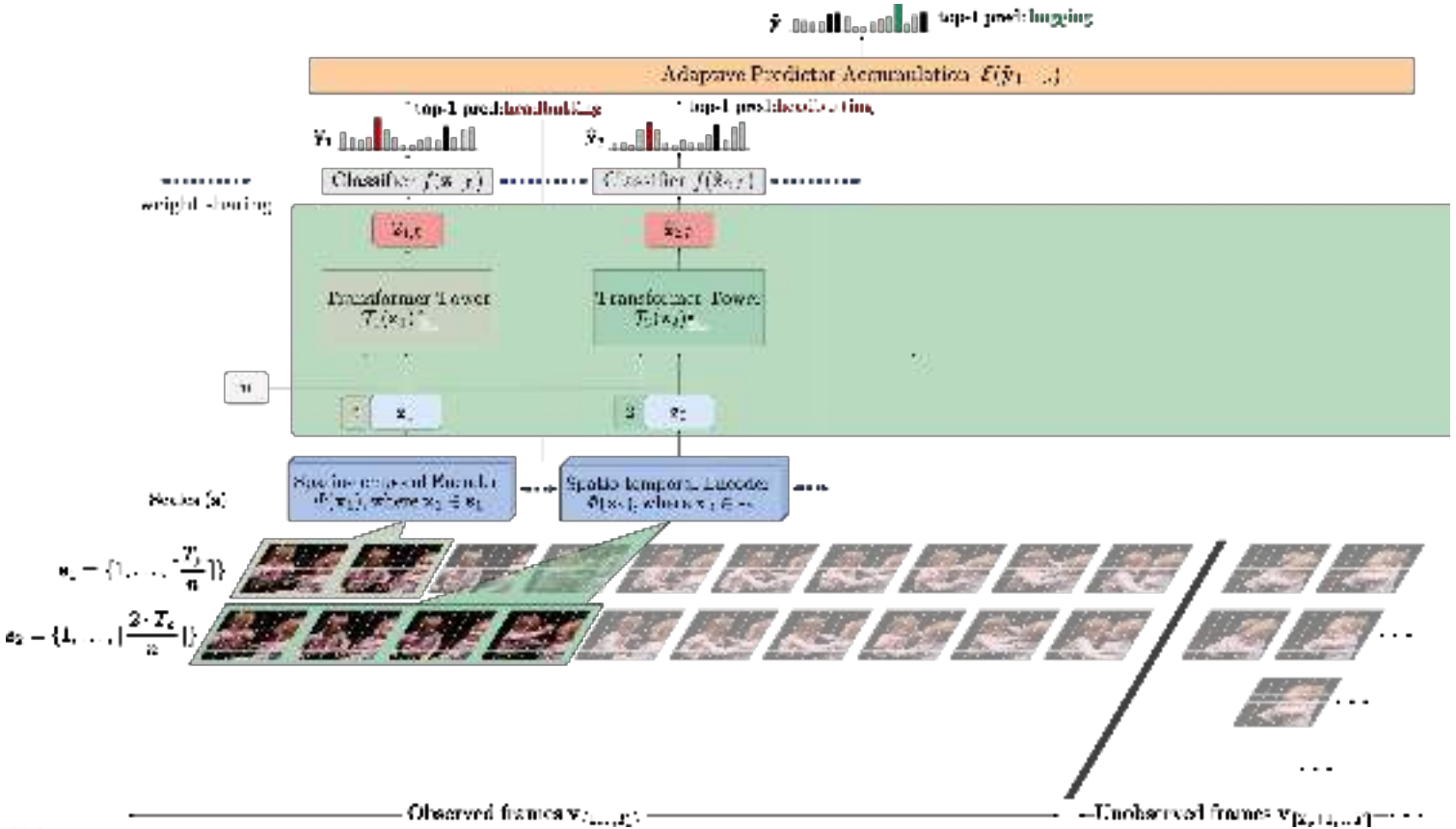
Early Action Prediction

with: Alex Stergiou



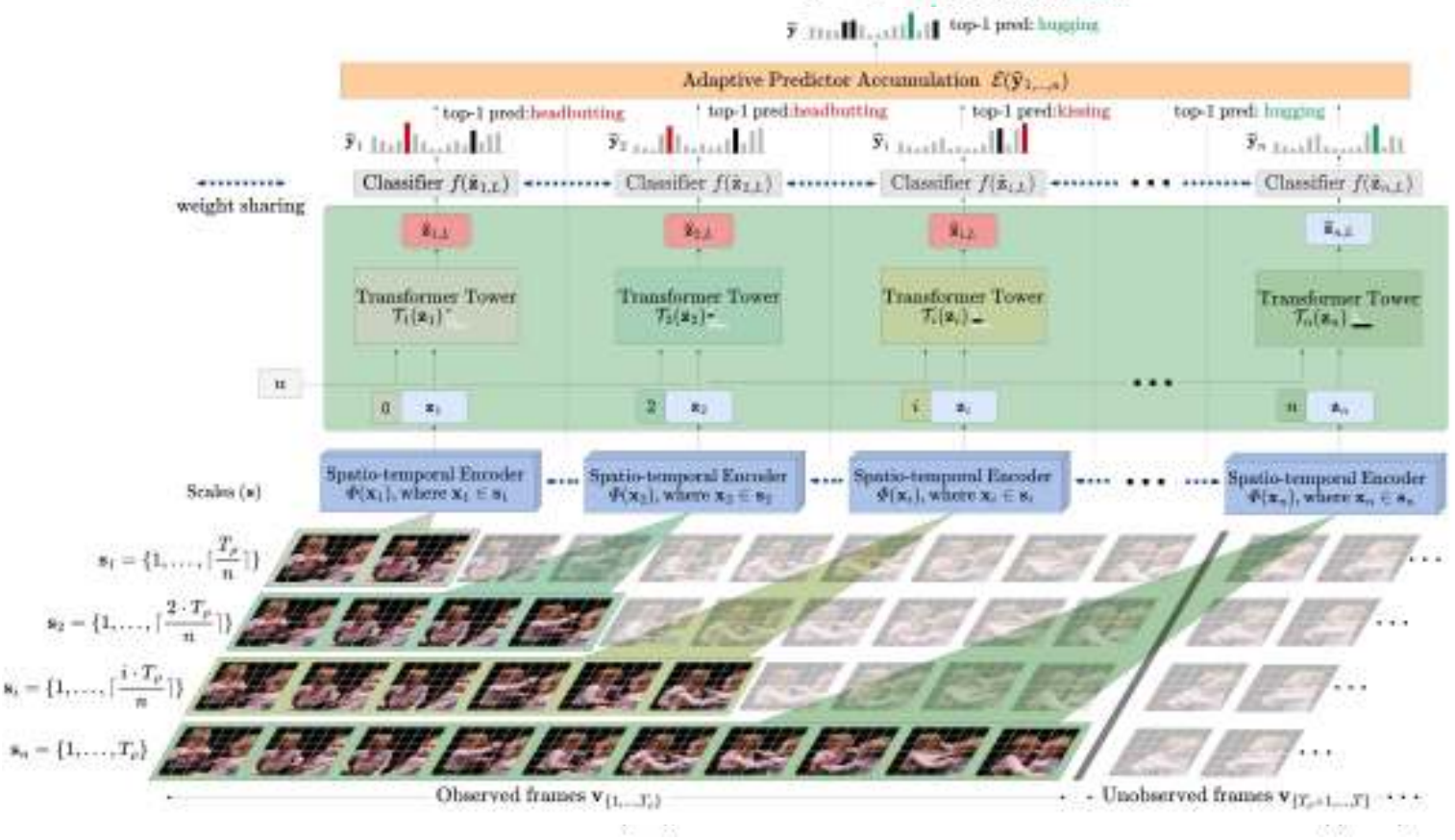
Early Action Prediction

with: Alex Stergiou



Early Action Prediction

with: Alex Stergiou



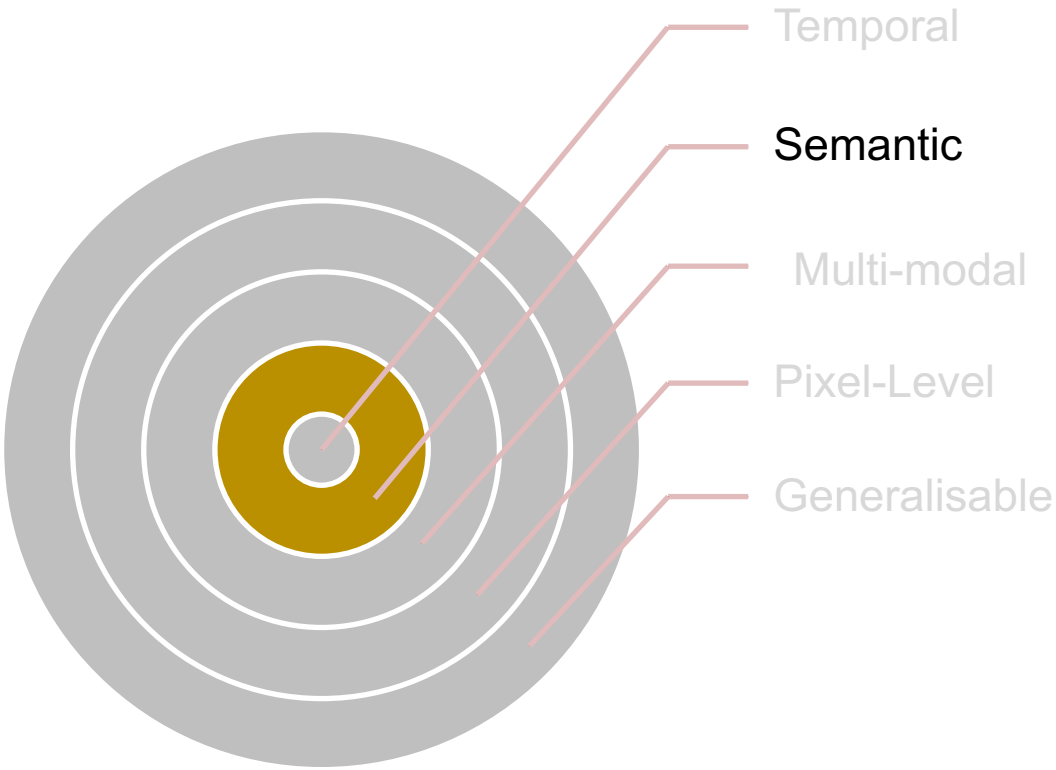
Early Action Prediction

(a) **Video Scales Strategy.**

Scale strategy	Observation ratios (ρ)			
	0.2	0.4	0.6	0.8
full	86.4	88.3	88.8	89.0
equal	83.7	84.6	86.3	87.1
random	88.8	89.7	90.2	90.6
decreasing	90.0	**90.9**	91.6	**92.6**
increasing	**90.2**	**90.9**	**91.8**	92.3

A Stergiou and D Damen (2022). Temporal Progressive Attention for Early Action Prediction. ArXiv

Representing Egocentric Actions



Verb?

Noun?

Verbs:

add

pour

sprinkle

salt

season

Nouns:

salt

sea salt

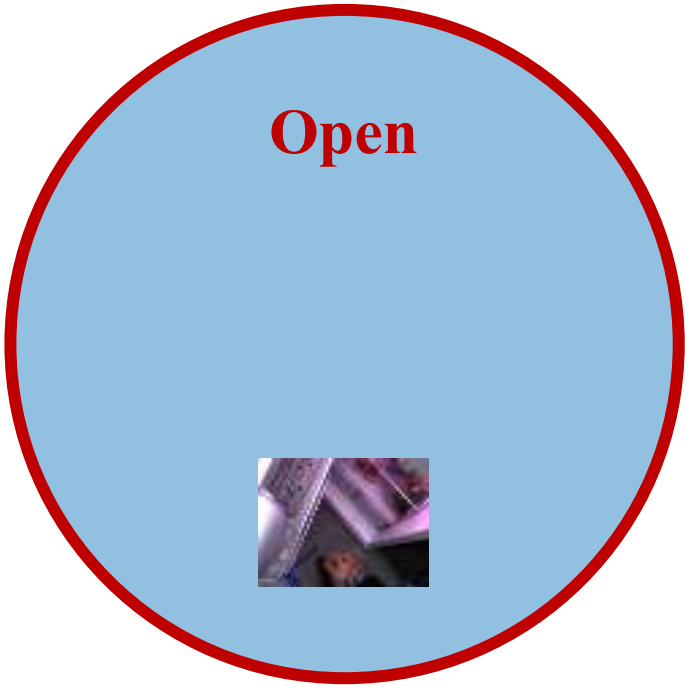
seasoning

salt granules

sprinkle salt

season meat

M Wray and D Damen (2019). Learning Visual Actions Using Multiple Verb-Only Labels. British Machine Vision Conference (BMVC).



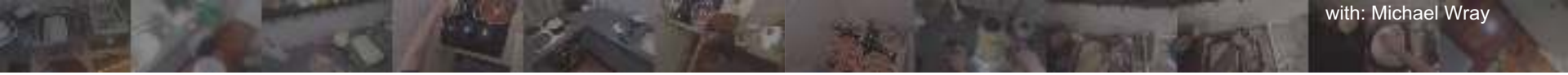
M Wray and D Damen (2019). Learning Visual Actions Using Multiple Verb-Only Labels. British Machine Vision Conference (BMVC).

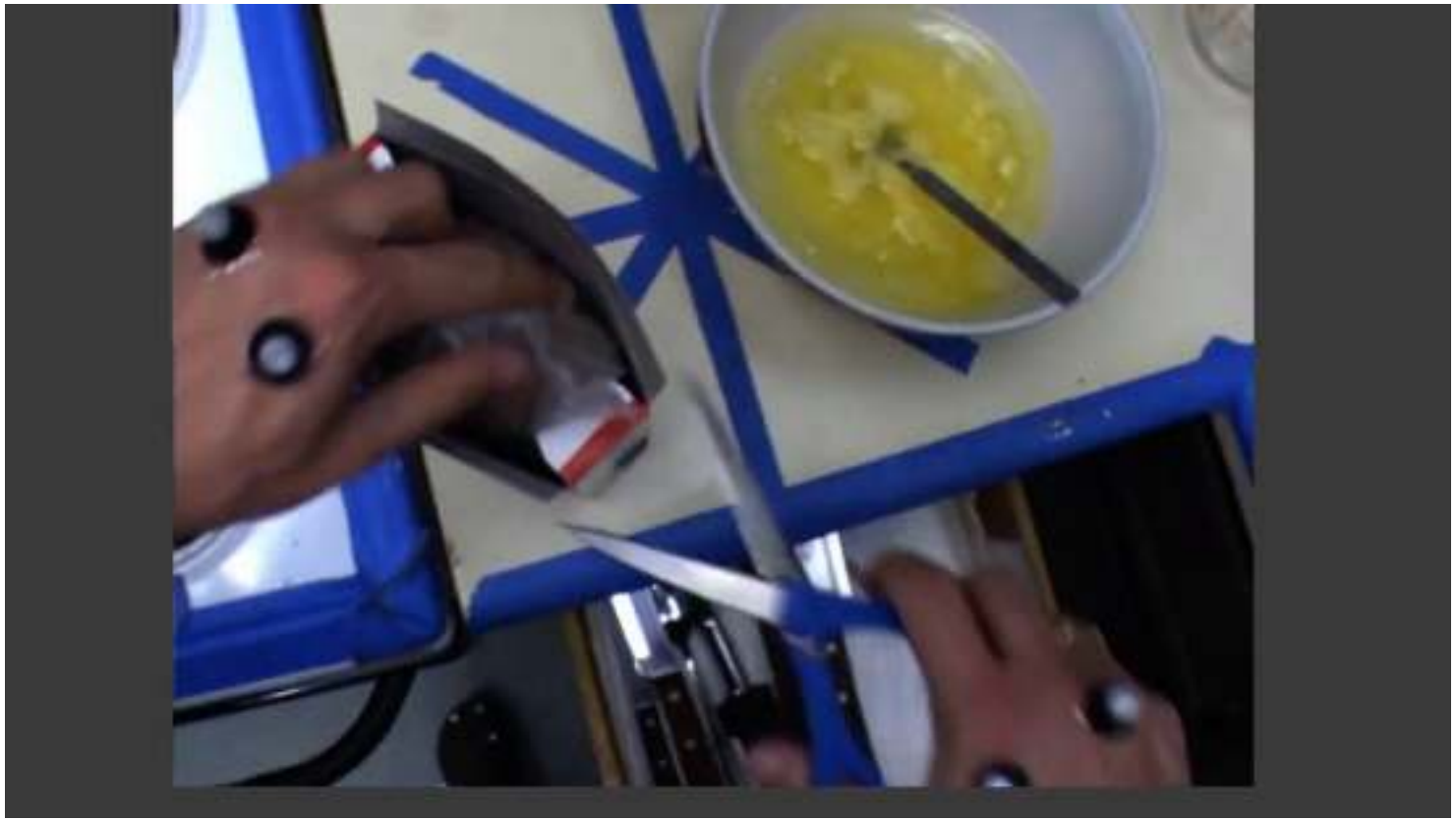
M Wray and D Damen (2019). Learning Visual Actions Using Multiple Verb-Only Labels. British Machine Vision Conference (BMVC).

Open

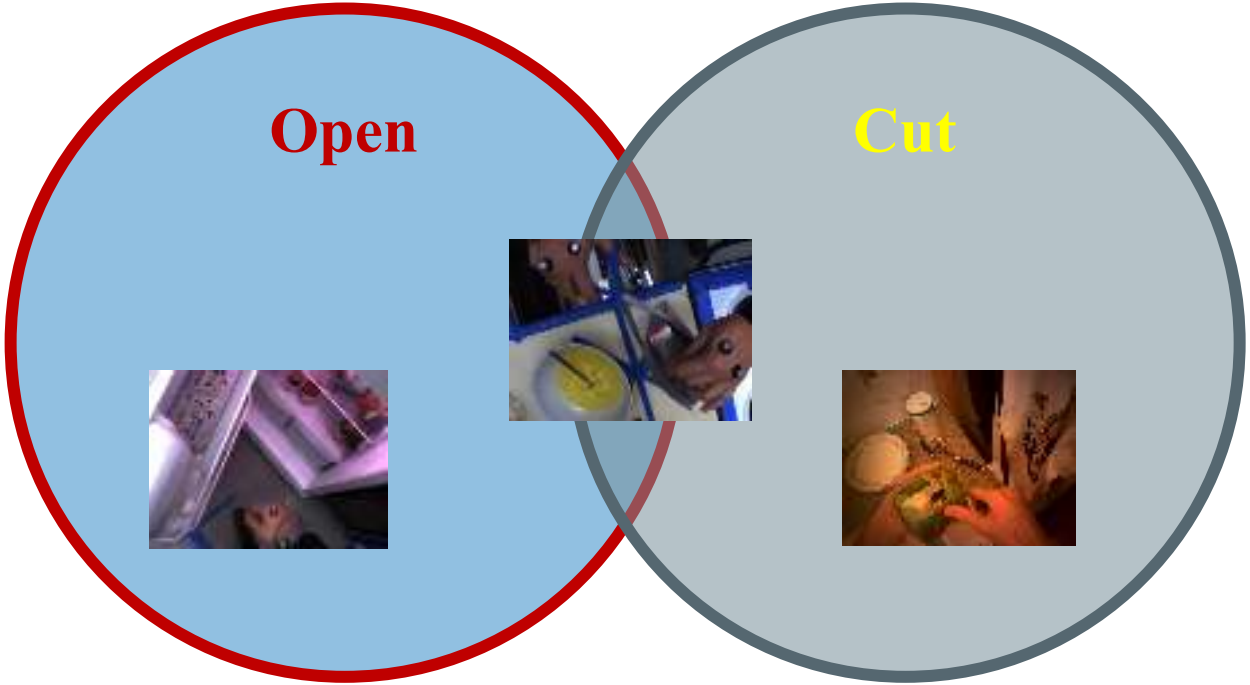
Cut

M Wray and D Damen (2019). Learning Visual Actions Using Multiple Verb-Only Labels. British Machine Vision Conference (BMVC).





M Wray and D Damen (2019). Learning Visual Actions Using Multiple Verb-Only Labels. British Machine Vision Conference (BMVC).



M Wray and D Damen (2019). Learning Visual Actions Using Multiple Verb-Only Labels. British Machine Vision Conference (BMVC).

On Semantic Similarity in Video Retrieval

with: Michael Wray
Hazel Doughty

- Which of these captions correspond to the following video?

A band is performing for the crowd

A man is peeling fruit carefully and neatly.

A girl is sitting in a chair

Add prawns to the pan and mix

On Semantic Similarity in Video Retrieval

with: Michael Wray
Hazel Doughty

- Which of these captions correspond to the following video?

A man performing an Origami tutorial

A demonstration in Origami

A guy explains the steps of folding paper

A man folding a piece of paper into a paper airplane

On Semantic Similarity in Video Retrieval

with: Michael Wray
Hazel Doughty

- Previous methods have made the following assumption
  - *"There exists only one corresponding caption for a given video and vice versa"*

Peel and chop the potatoes
Peel and cut up the potato
Peel the potatoes and cut them
Peel and cut the potatoes into chunks
Peel the potatoes and cut them into halves

YouCook2

Put fork and spoon in drying rack
Put spoons in drying rack
Put spoon in drying rack
Put bowl in drying rack
Put plate in drying rack

EPIC-KITCHENS

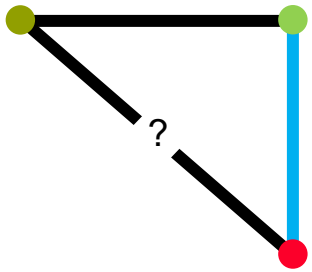
MSR-VTT

A band is performing for the crowd
A band is performing on a brightly lit stage
A band is playing a show
A band and singers perform
3 guys singing and playing instruments on a stage

On Semantic Similarity in Video Retrieval

with: Michael Wray
Hazel Doughty

- Want to relate two items semantically.



A band is performing for a crowd

Two men competing in a ping pong match

- Assume that a caption sufficiently describes a video.

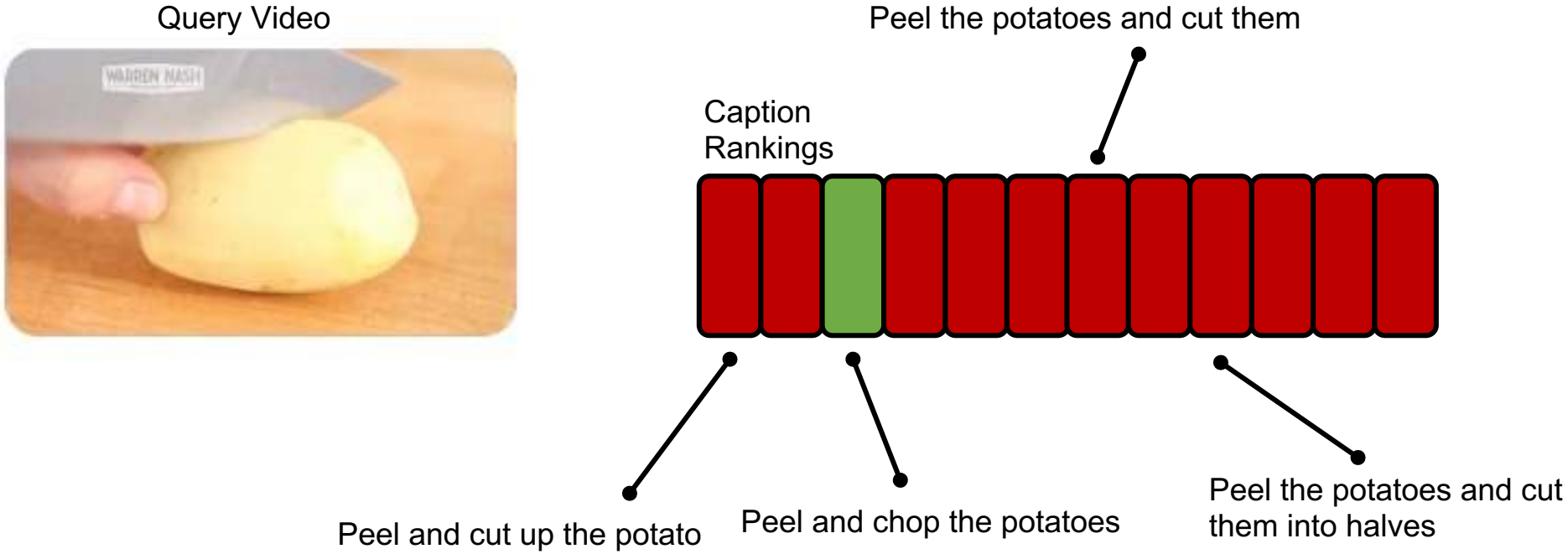
- Define a proxy function that relates captions

$$S(x_i, y_j) = S'(y_i, y_j)$$

On Semantic Similarity in Video Retrieval

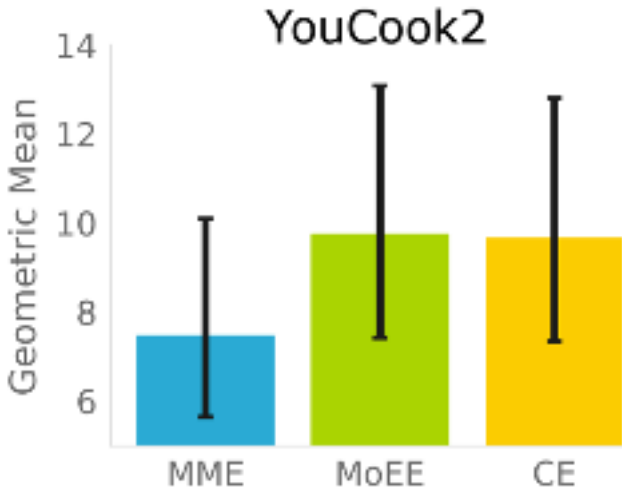
with: Michael Wray
Hazel Doughty

- When evaluating with a single caption, the correct caption can be arbitrary.



On Semantic Similarity in Video Retrieval

with: Michael Wray
Hazel Doughty

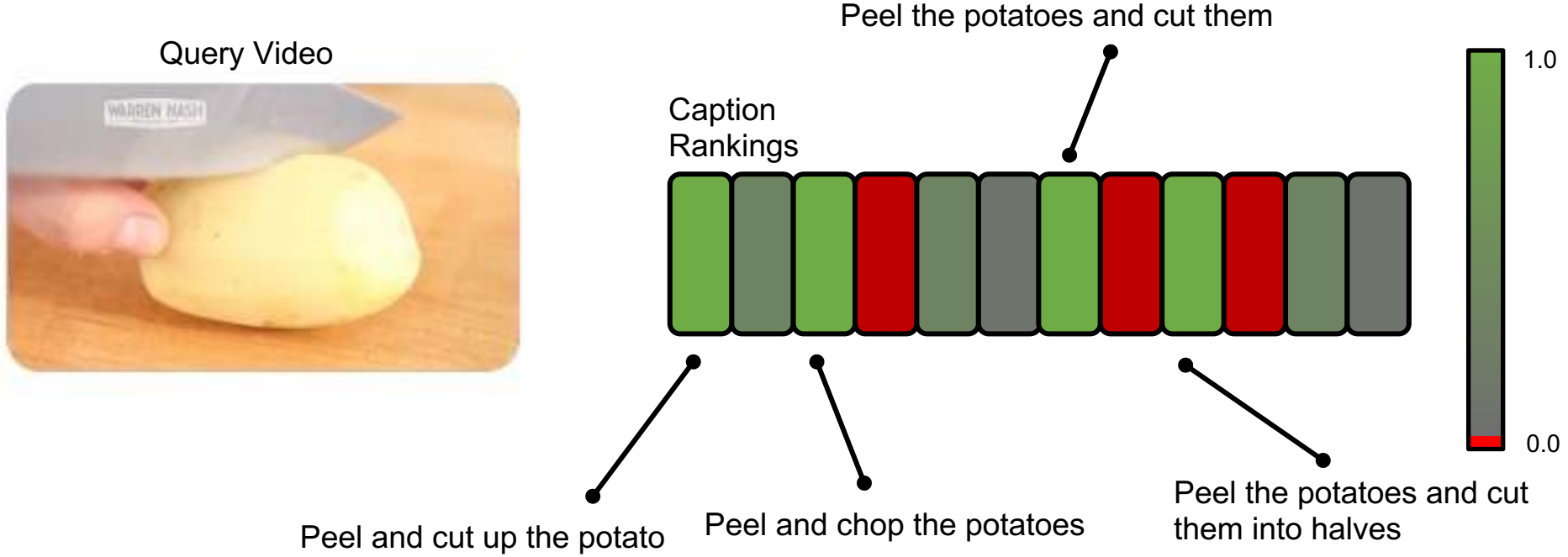


MoEE: Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. CoRR, abs/1804.02516, 2018
CE: Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In BMVC, 2019
JPoSE: Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts of-speech embeddings. In ICCV, 2019

On Semantic Similarity in Video Retrieval

with: Michael Wray
Hazel Doughty

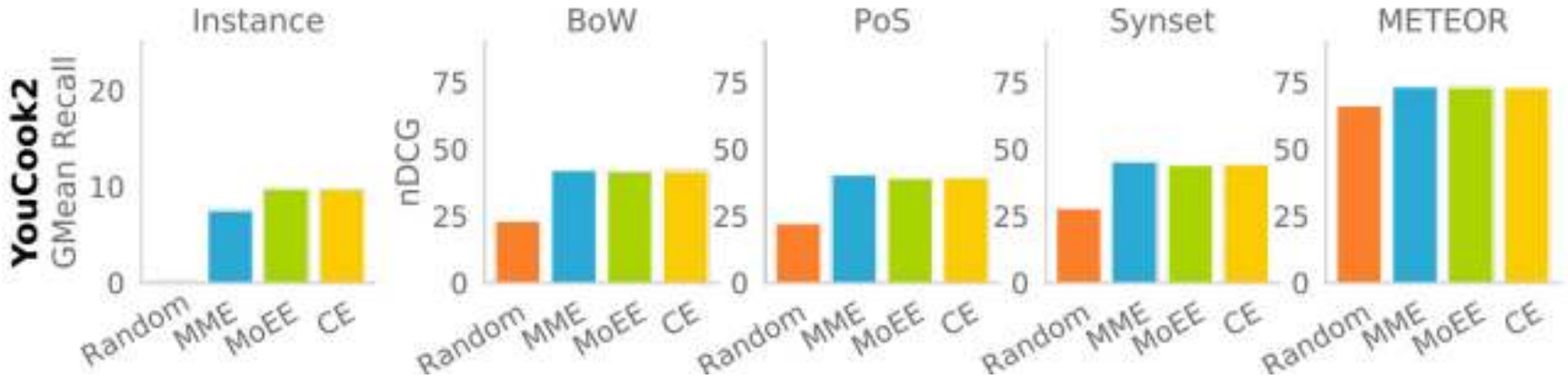
- We instead, propose to use normalised Discounted Cumulative Gain to evaluate multiple items with differing relevance.



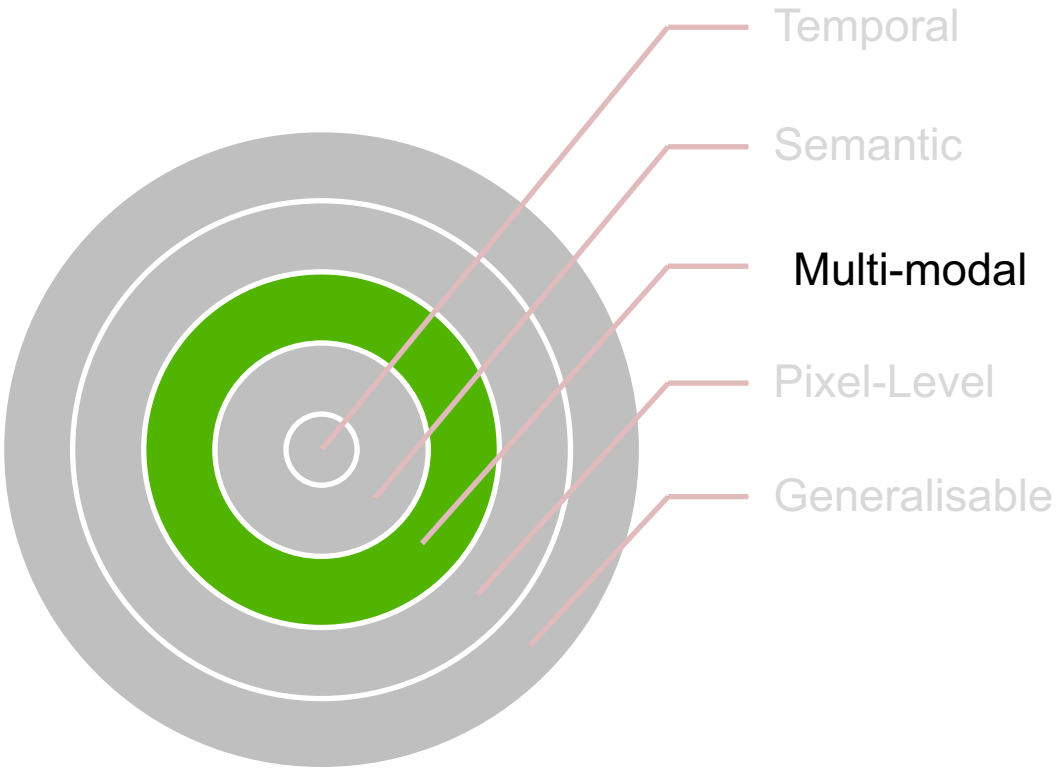
On Semantic Similarity in Video Retrieval

with: Michael Wray
Hazel Doughty

- Whilst models outperform the MLP baseline (MME) for Instance Video Retrieval, this isn't the case when Semantic Similarity is used.



Representing Egocentric Actions



Multi-modal learning…

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

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

Multi-modal learning…

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

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

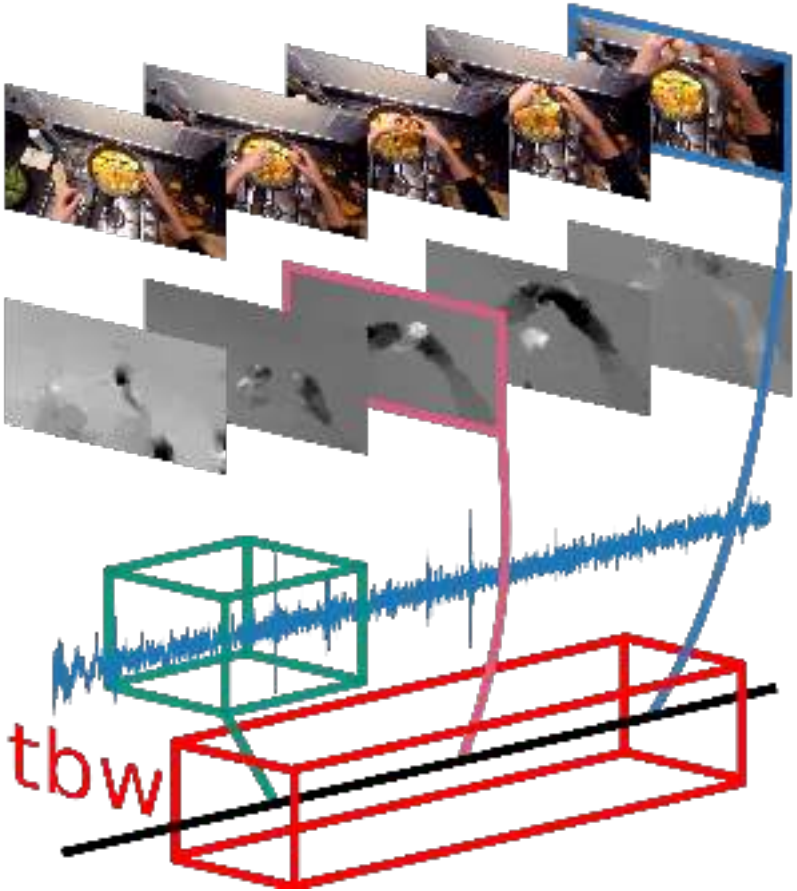
Multi-modal learning…

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

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

Audio-Visual Temporal Binding

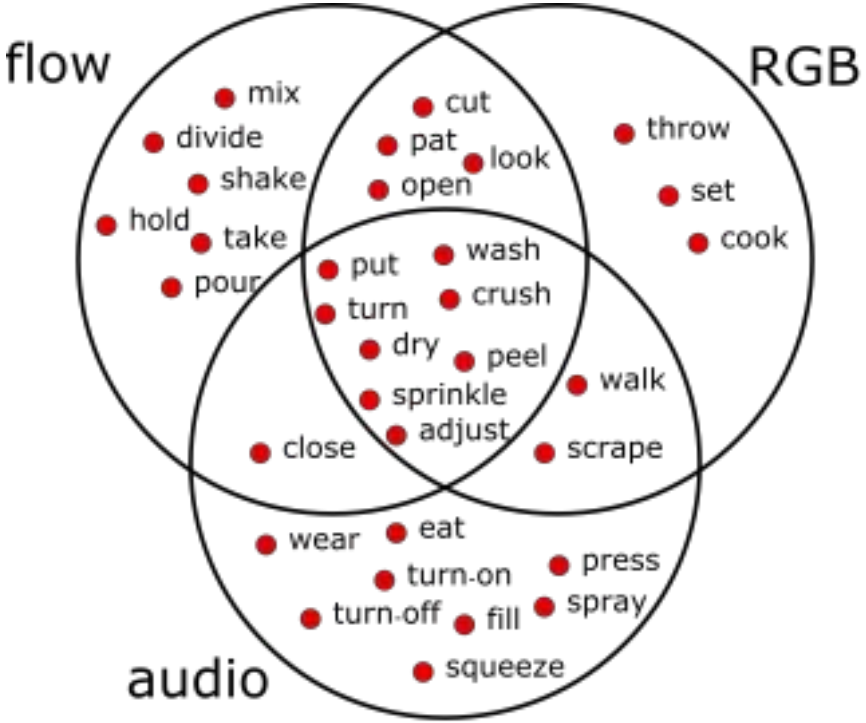
with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman



E Kazakos, A Nagrani, A Zisserman, D Damen (2019). EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition. ICCV

Audio-Visual Temporal Binding

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman



Audio-Visual Temporal Binding

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman

		Top-1 Accuracy			Top-5 Accuracy			Avg Class Precision			Avg Class Recall		
		VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION
S1	RGB	45.68	36.80	19.86	85.56	64.19	41.89	61.64	34.32	09.96	23.81	31.62	08.81
	Flow	55.65	31.17	20.10	85.99	56.00	39.30	48.83	26.84	09.02	27.58	24.15	07.89
	Audio	43.56	22.35	14.21	79.66	43.68	27.82	32.28	19.10	07.27	25.33	18.16	06.17
	TBN (RGB+Flow)	60.87	42.93	30.31	89.68	68.63	51.81	**61.93**	39.68	18.11	39.99	38.37	16.90
	TBN (All)	**64.75**	**46.03**	**34.80**	**90.70**	**71.34**	**56.65**	55.67	**43.65**	**22.07**	**45.55**	**42.30**	**21.31**
S2	RGB	34.89	21.82	10.11	74.56	45.34	25.33	19.48	14.67	04.77	11.22	17.24	05.67
	Flow	48.21	22.98	14.48	77.85	45.55	29.33	23.00	13.29	05.63	19.61	16.09	07.61
	Audio	35.43	11.98	06.45	69.20	29.49	16.18	22.46	09.41	04.59	18.02	09.79	04.19
	TBN (RGB+Flow)	49.61	25.68	16.80	78.36	50.94	32.61	30.54	20.56	09.89	21.90	20.62	11.21
	TBN (All)	**52.69**	**27.86**	**19.06**	**79.93**	**53.78**	**36.54**	**31.44**	**21.48**	**12.00**	**28.21**	**23.53**	**12.69**

E Kazakos, A Nagrani, A Zisserman, D Damen (2019). EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition. ICCV

Harmonic vs Percussive

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman

EPIC-KITCHENS

Harmonic Sounds

Percussive Sounds

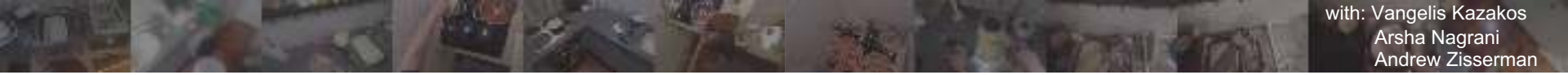
Harmonic vs Percussive

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman

VGG-Sound

Harmonic Sounds

Percussive Sounds

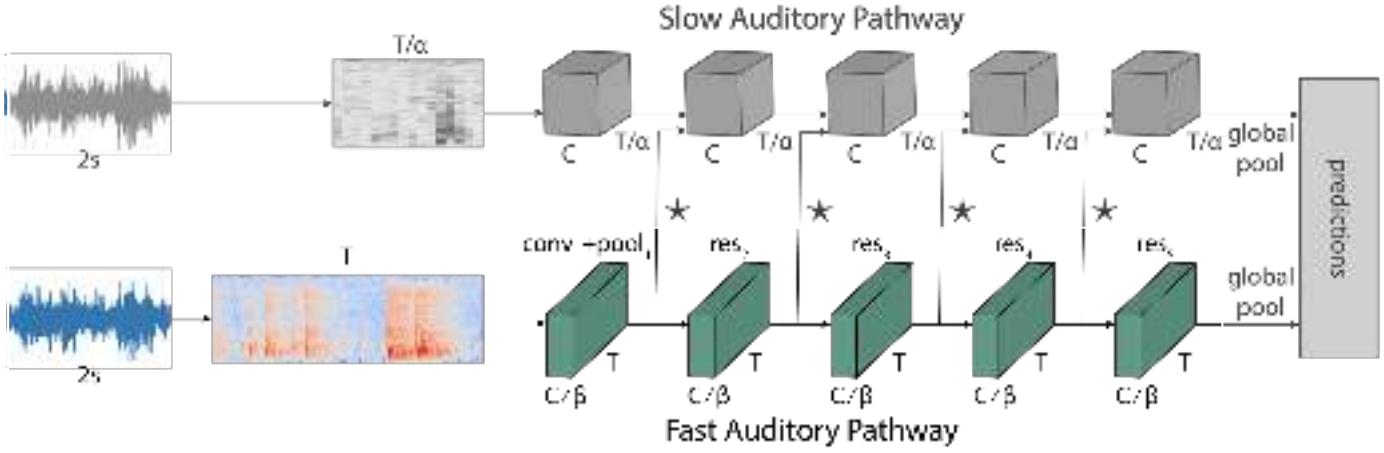


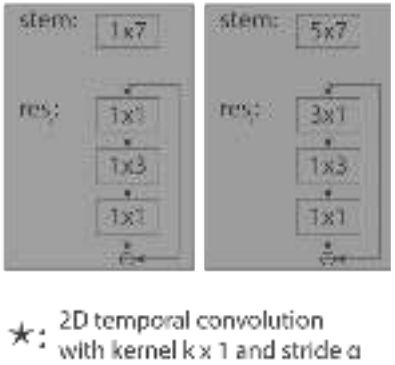
Auditory Slow-Fast

Outstanding Paper Award – ICASSP 2021

Audio Slow-Fast

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman

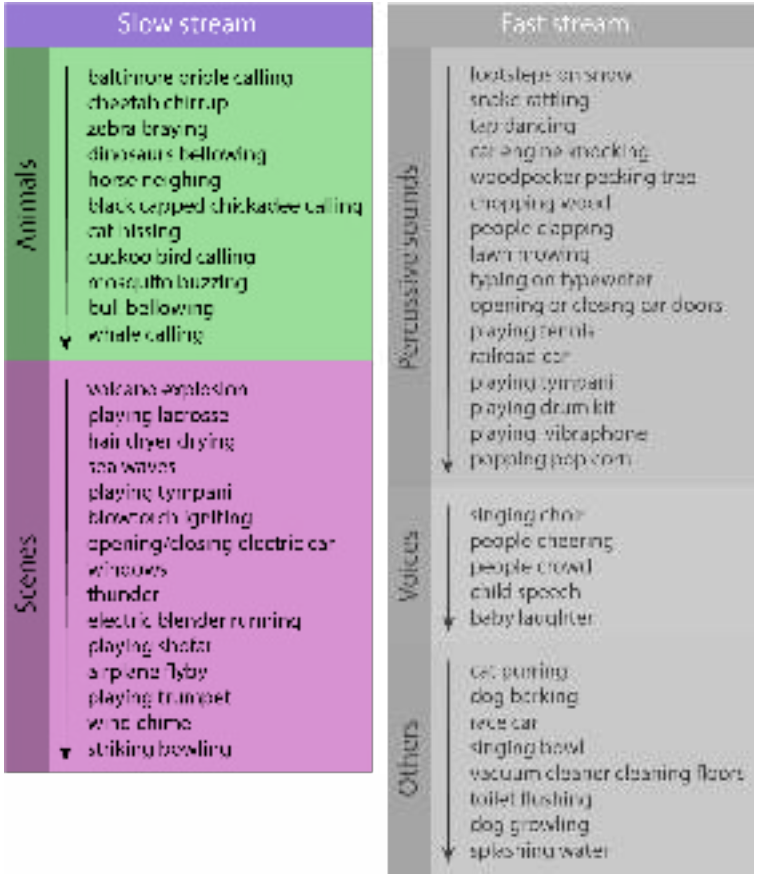




- Slow has low temporal precision and large amount of channels
- Fast has fewer channels but high temporal resolution
- Multi-level lateral connections
- Separable convolutions

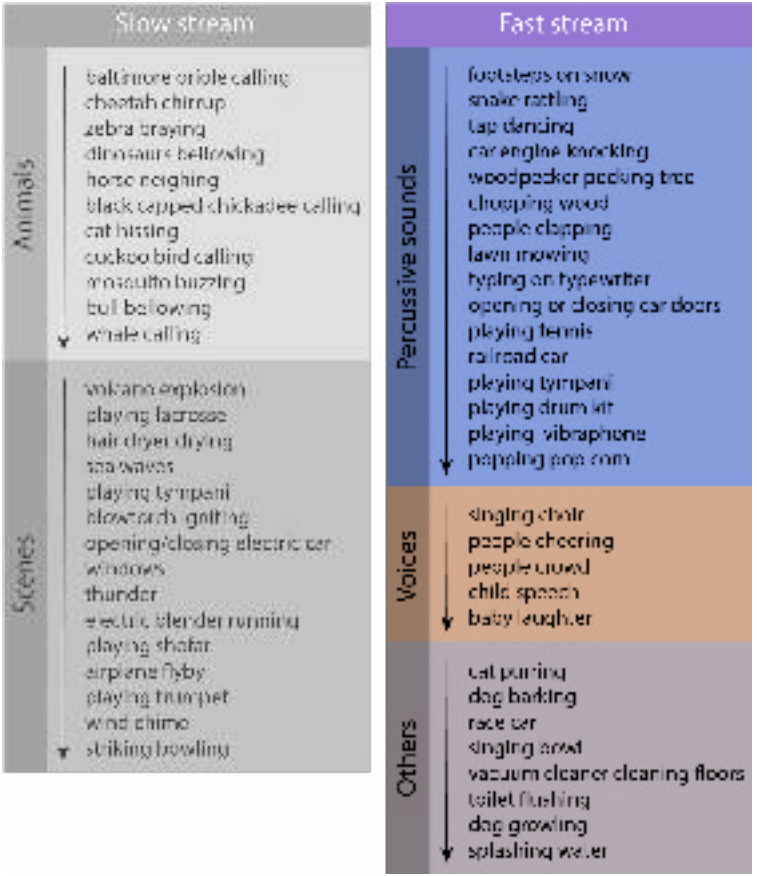
Audio Slow-Fast

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman



Audio Slow-Fast

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman



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

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	38.4/36.2
SF ResNet50	26m	Spectrogram	[17]	74.0/74.3	56.9/73.4	59.6/65.2	63.3/71.7	37.2/36.6
NFNet-F0	68m	Spectrogram	Ours	**76.1**/76.0	59.0/65.9	61.8/65.5	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.

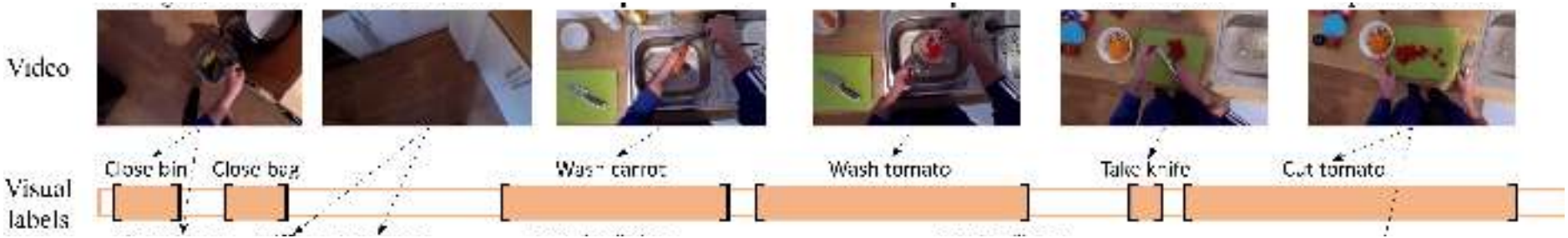
Index Terms— audio representations, representation evaluation, speech, music, acoustic scenes

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

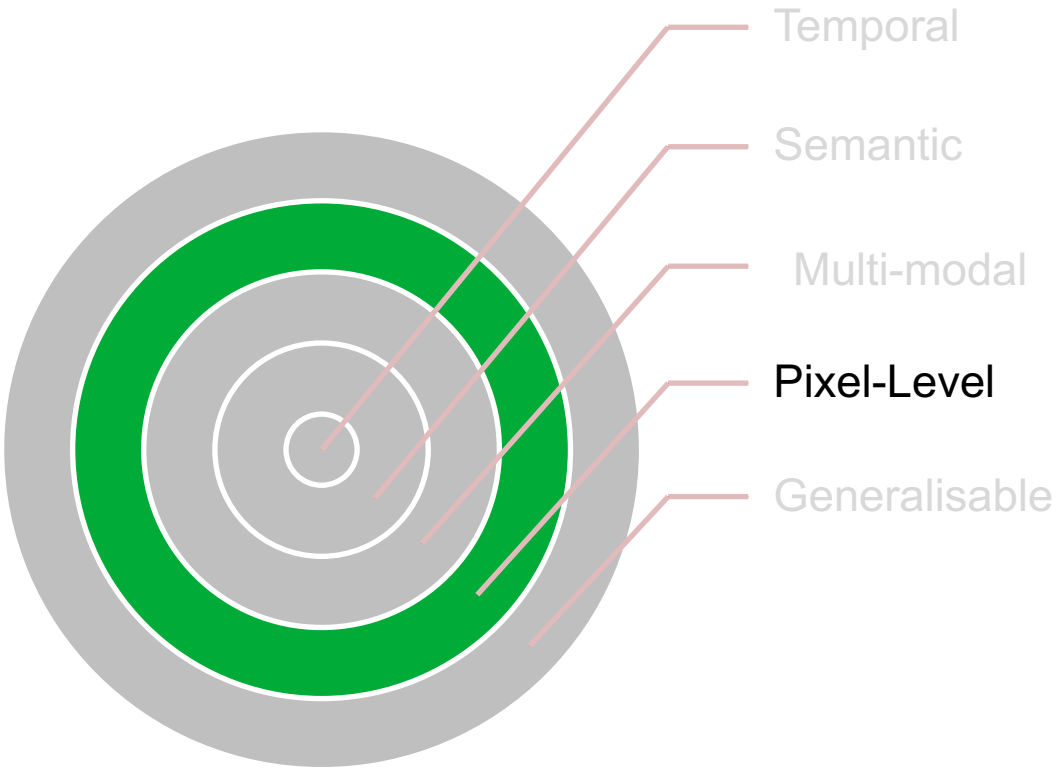
E Kazakos, A Nagrani, A Zisserman, D Damen (2021). Slow-Fast Auditory Streams For Audio Recognition. ICASSP

EPIC-SOUNDS

with: Jaesung Huh Jacob Chalk
Evangelos Kazakos Andrew Zisserman



Representing Egocentric Actions

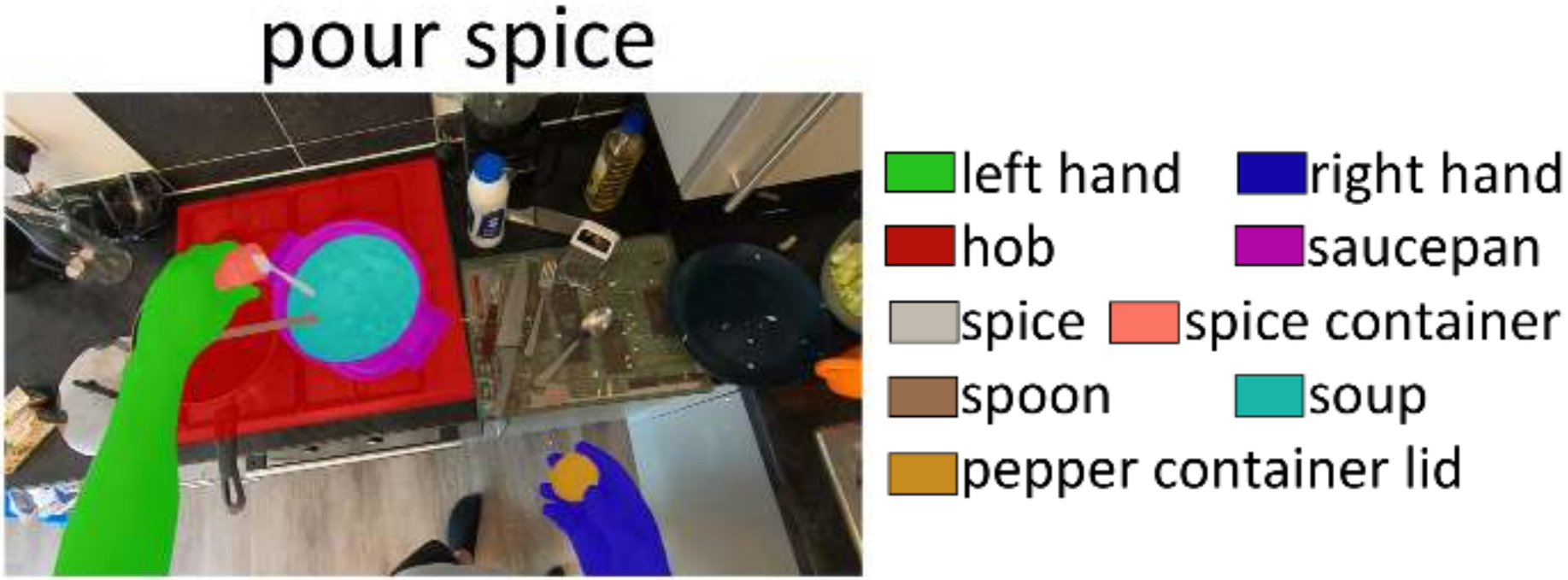


EPIC-KITCHENS VISOR

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

VISOR is….

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

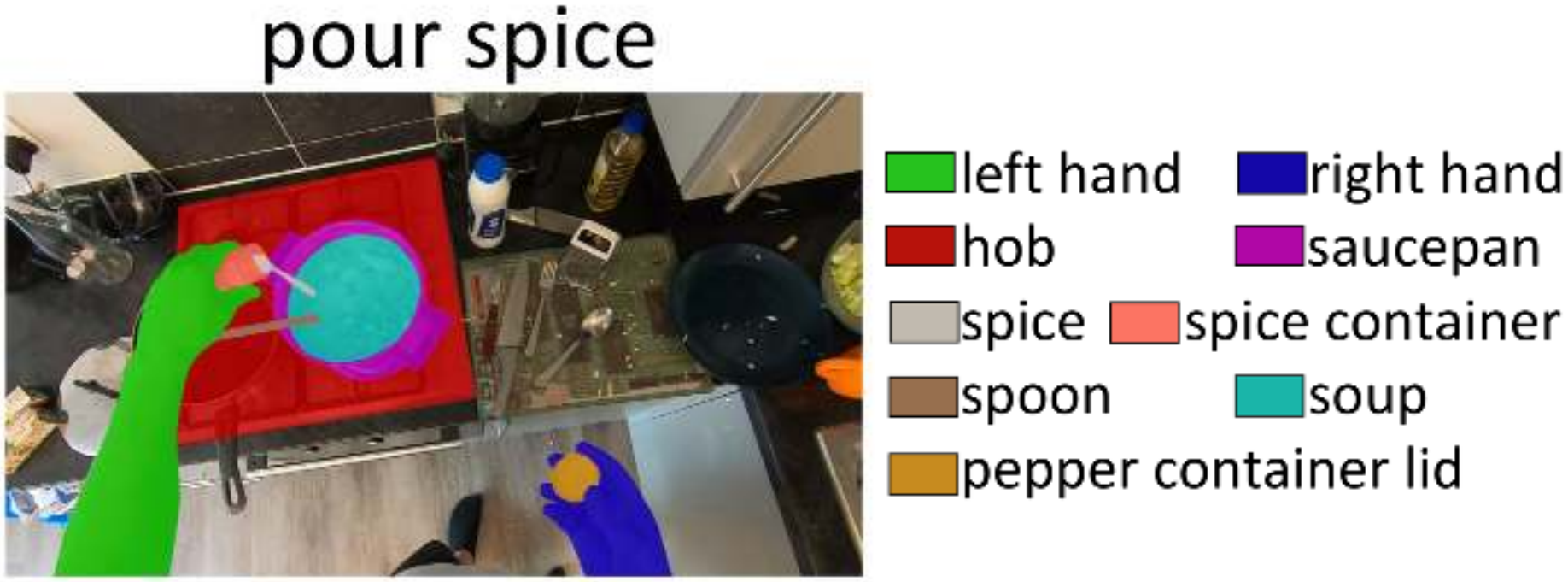


VISOR is….

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

VISOR is….

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



saucepan → pan → cookware

spoon → spoon → cutlery

VISOR is….

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

pour spice ← action

left hand
right hand
hob
saucepan
spice
spice container
spoon
soup
pepper container lid

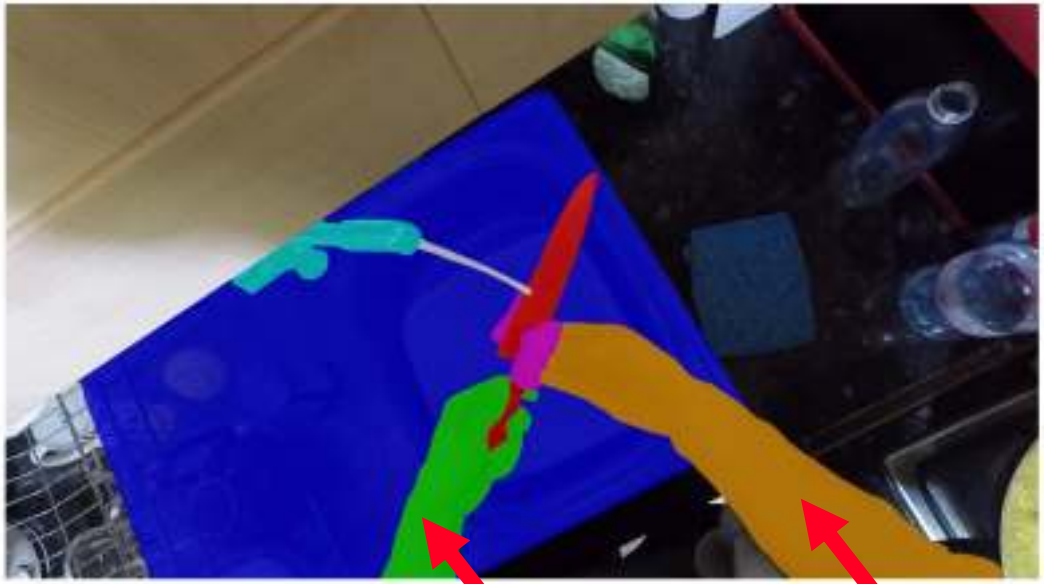
in-contact (spice container)

in-contact (container lid)

VISOR is….

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

wash knife



left hand
right hand
knife
sponge
sink
tap
water

in-contact (knife)

in-contact (sponge)

Comparative Stats

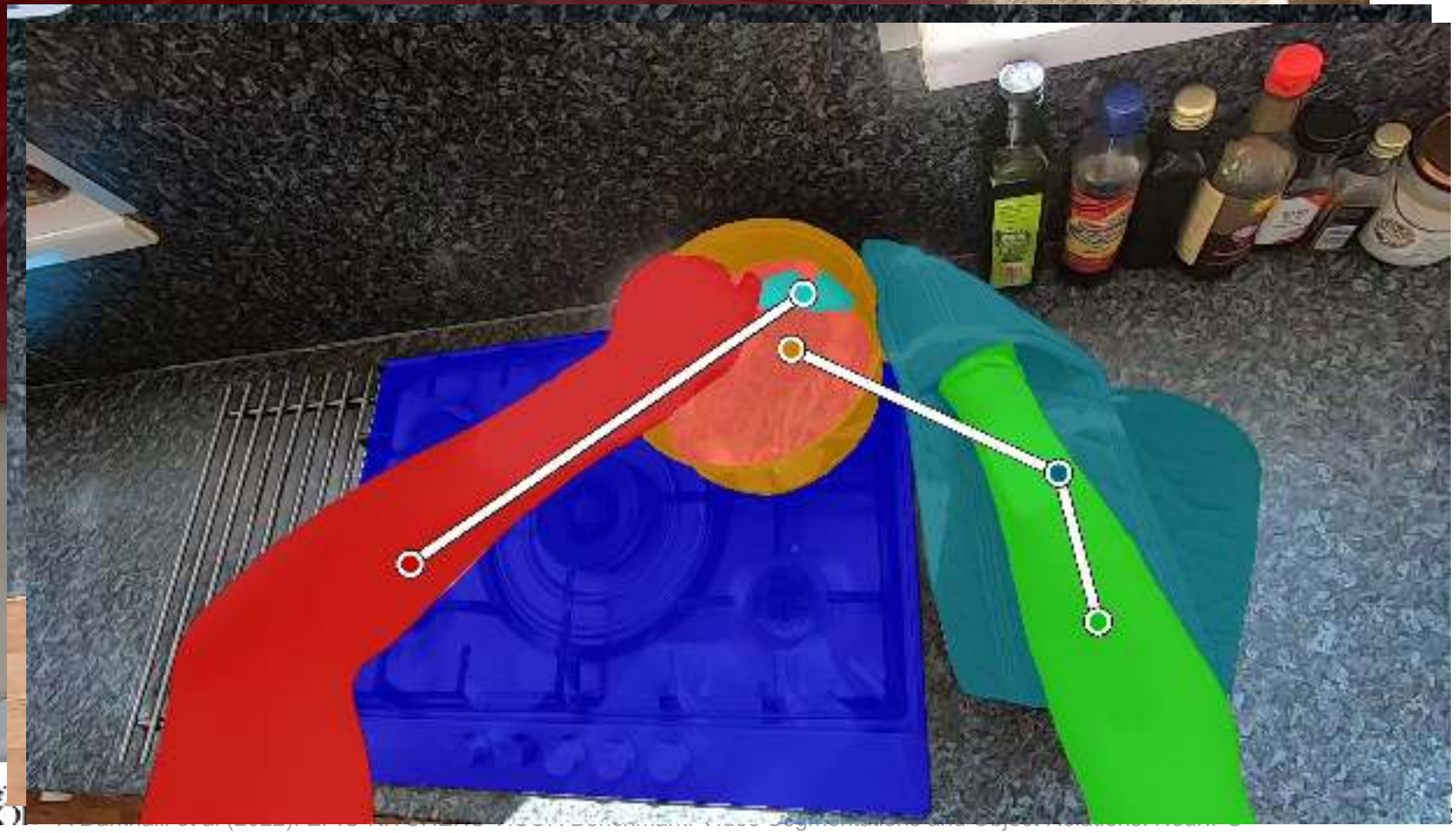
with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler, Dima Damen

Dataset	Basic Statistics		Pixel-Level Annotations	Action Annotations		
	Total Mins	Avg Seq Ln	Total Masks	Actions	#Action Classes	#Entity Classes
EgoHand [3]	72	-	15.1K	-	-	2
DAVIS [6]	8	3s	32.0K	-	-	-
YTVOS [43]	335	5s	197.2K	-	-	94
UVOv0.5 (Sparse) [41]	511	3s	*200.6K	10,213	300	-
VISOR (Ours)	**2,180**	**12s**†	**271.6K**	**27,961**	**2,594**	**257**

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

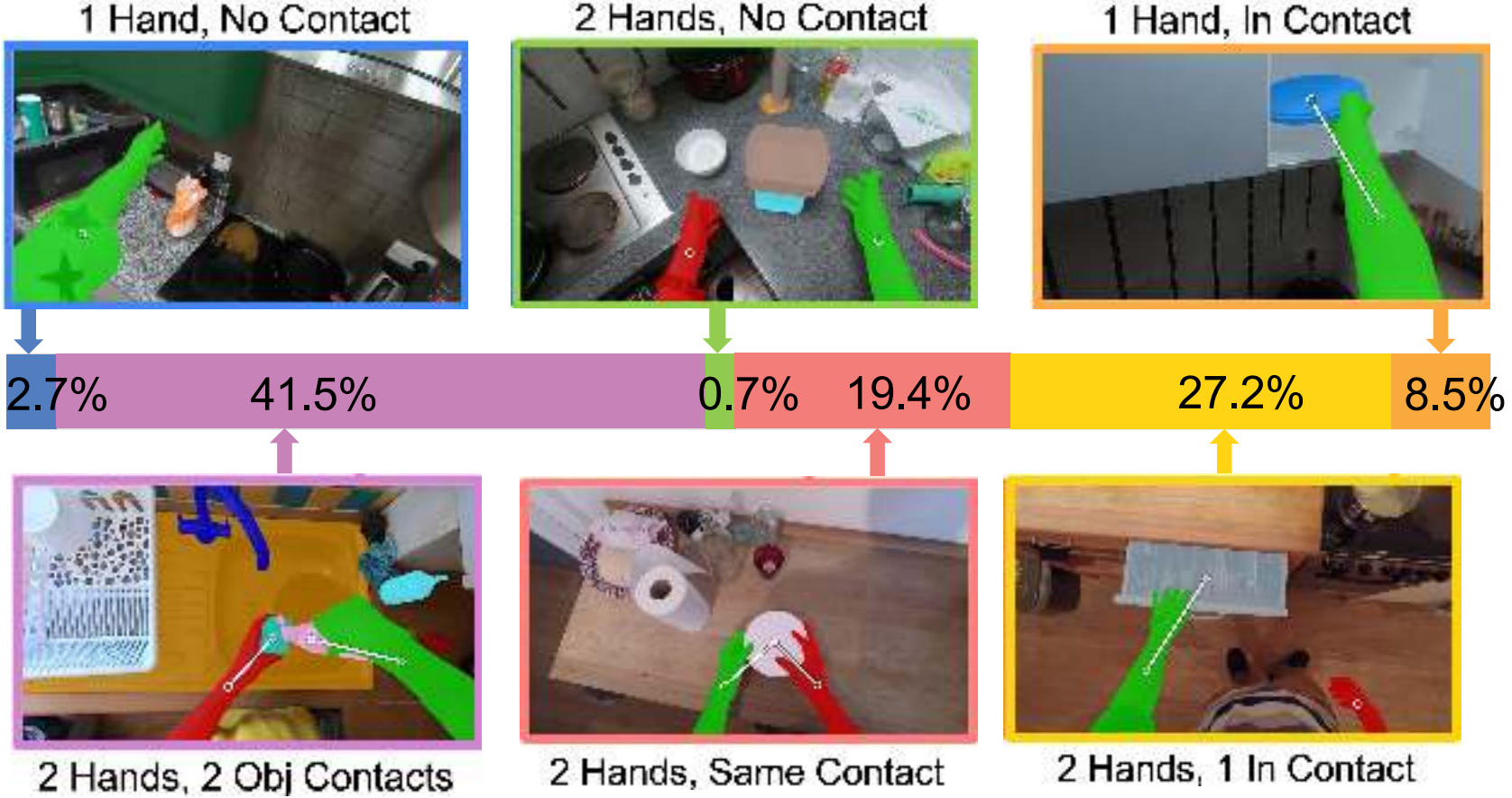
VISOR Relations

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



Object relation stats

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

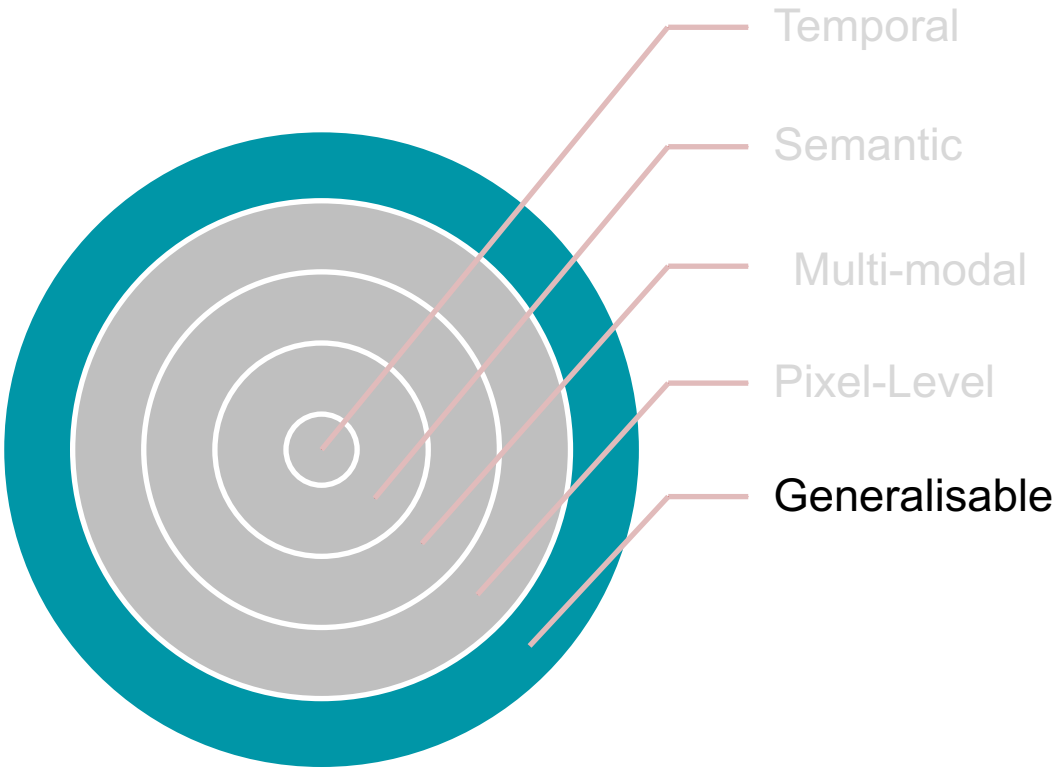


EPIC-KITTCHENS VISOR

Dataset Released: 16th of Aug 2022…
Code and Models: 27th of Sep 2022…

http://epic-kitchens.github.io/VISORFurther

Representing Egocentric Actions



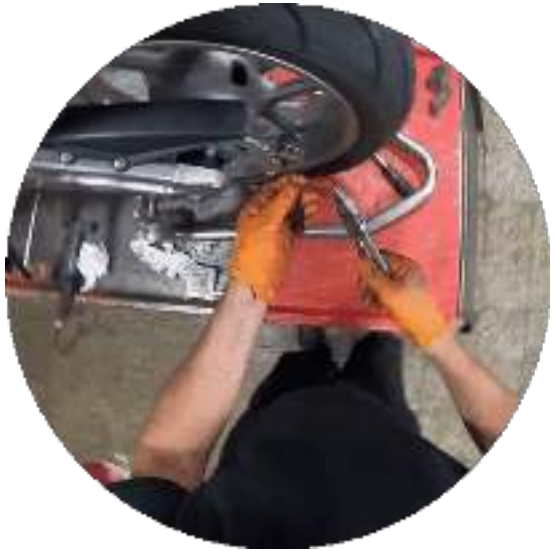
Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett

Ongoing work….

Generalisation across Scenarios and Locations

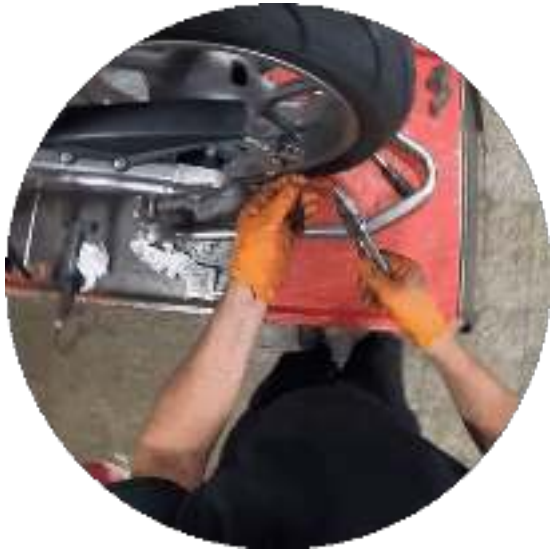
with: Chiara Plizzari
Toby Perrett



Ongoing work….

Generalisation across Scenarios and Locations

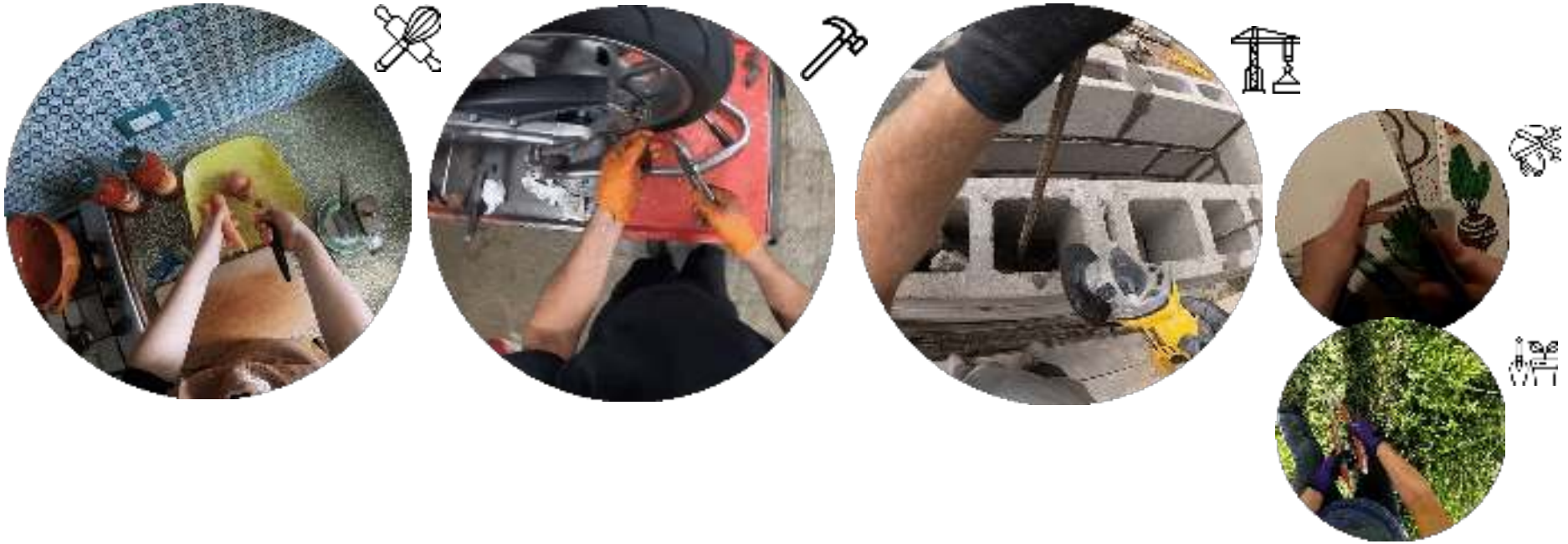
with: Chiara Plizzari
Toby Perrett



Ongoing work….

Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett



Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett

Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett

Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett

Ongoing work….

Generalisation across Scenarios and Locations

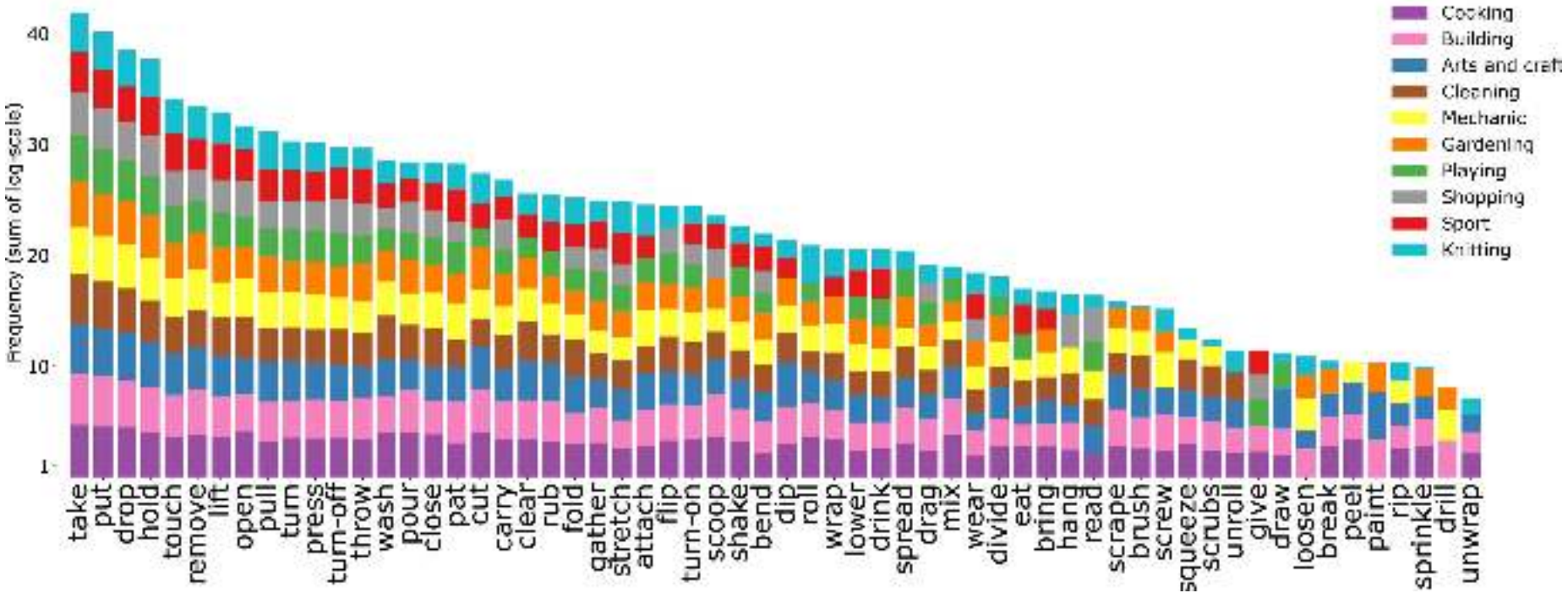
with: Chiara Plizzari
Toby Perrett

Ongoing work….

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



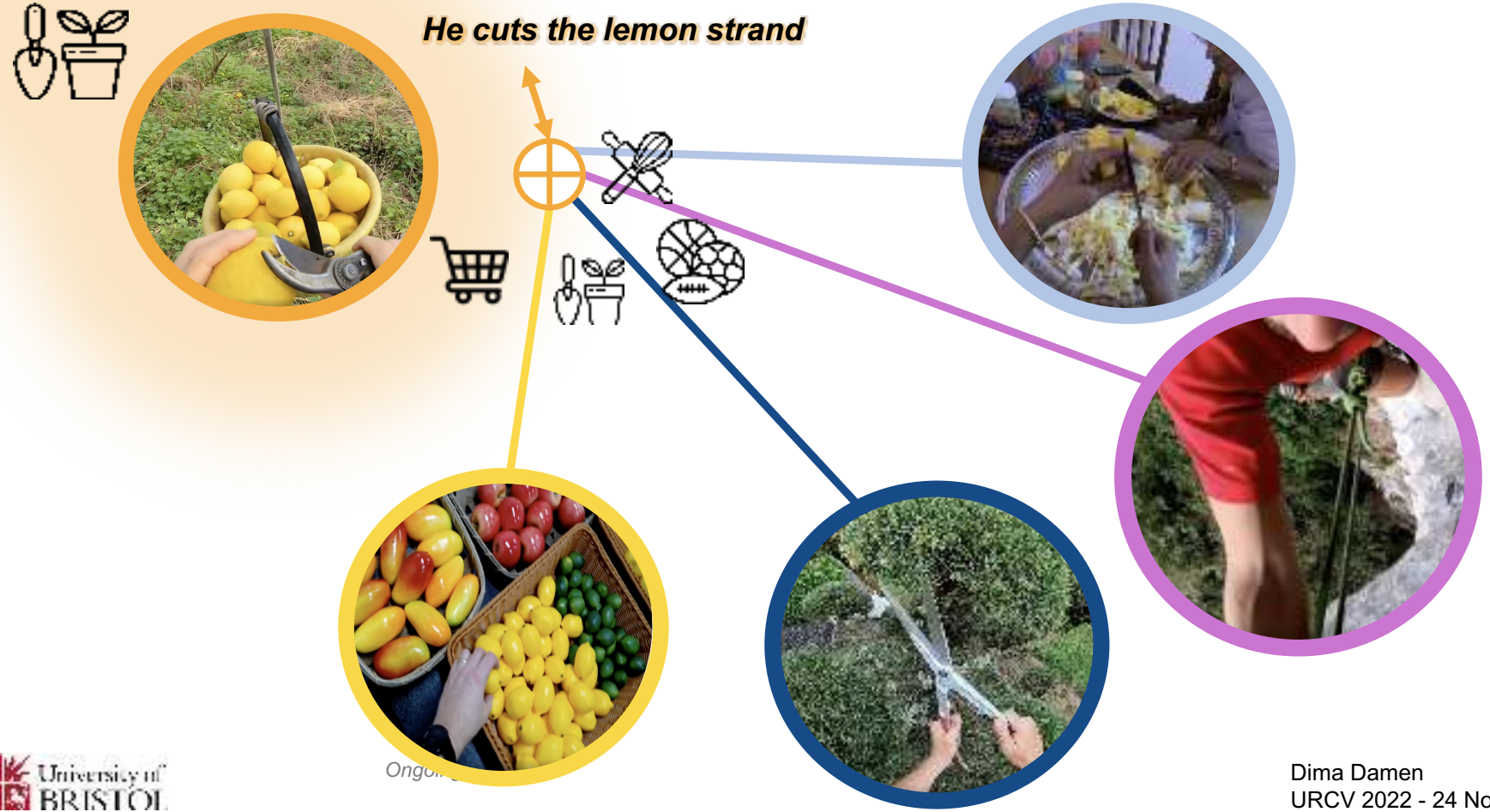
Ongoing work….

Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett

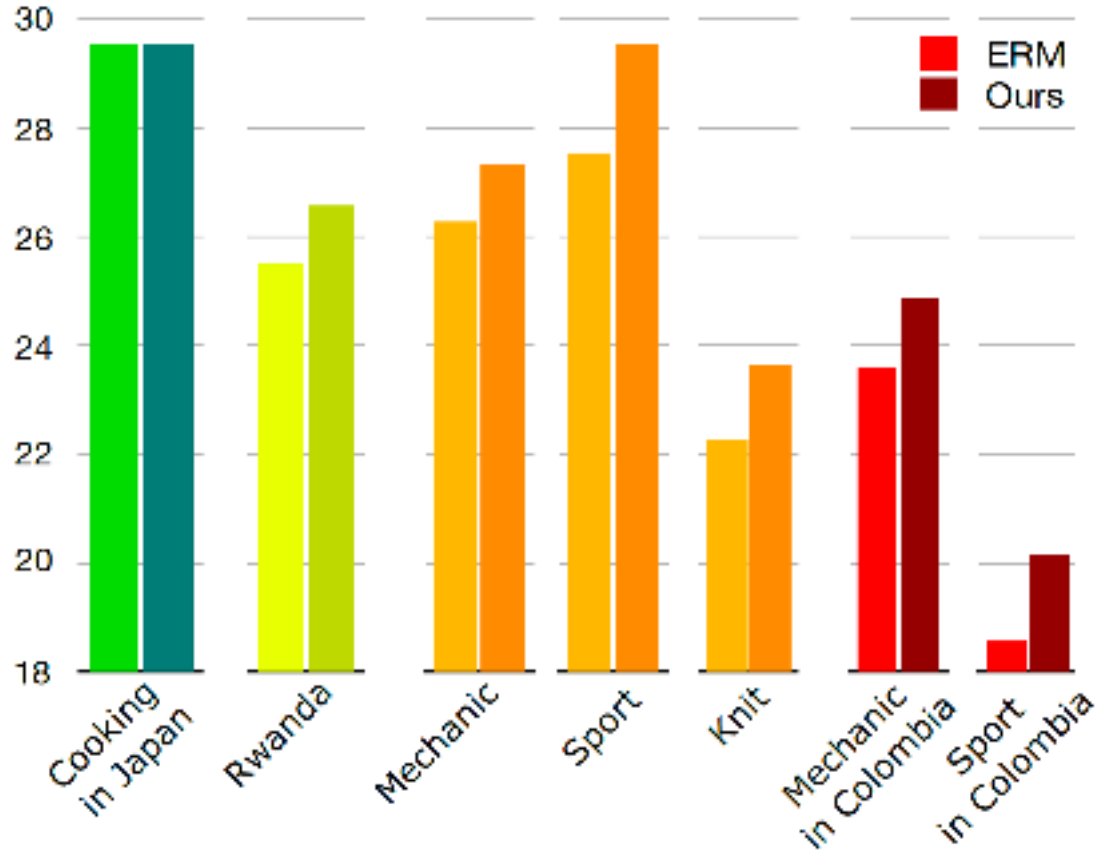
Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett



Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett



Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett

support 1 support 2 support 3 support 4 support 5

The Team

Thank you

For further info, datasets, code, publications…

http://dimadamen.github.io


@dimadamen


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

Q&A