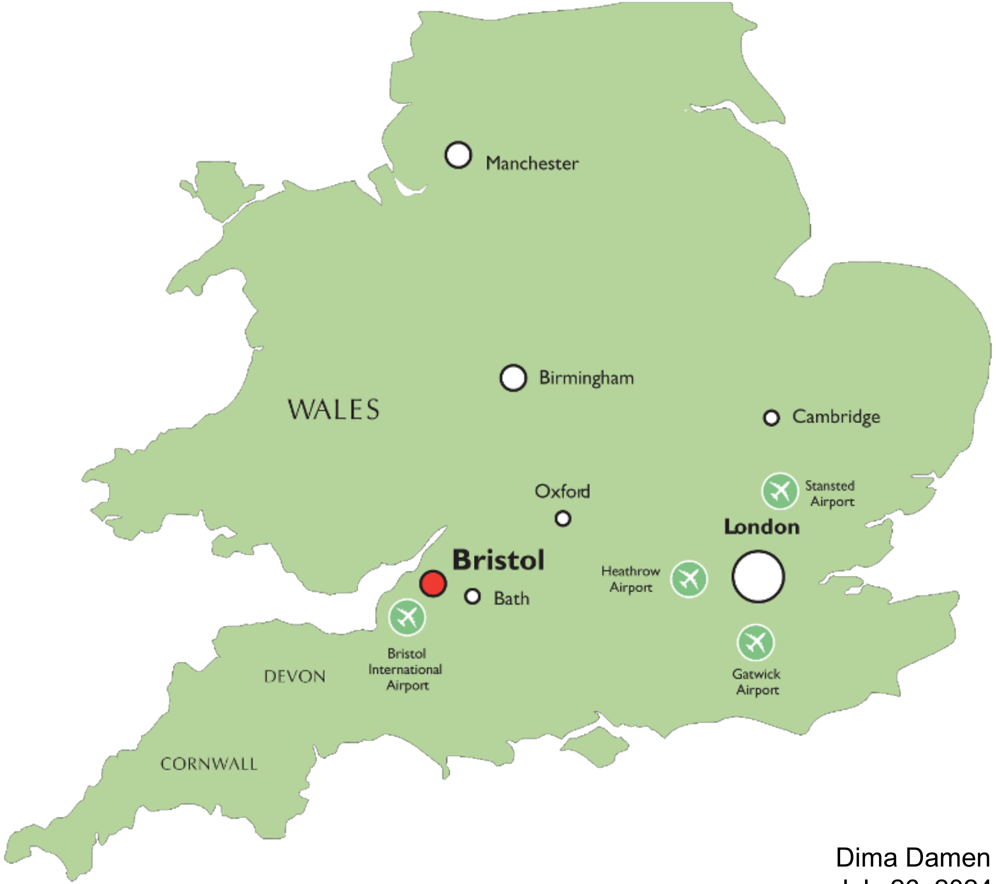
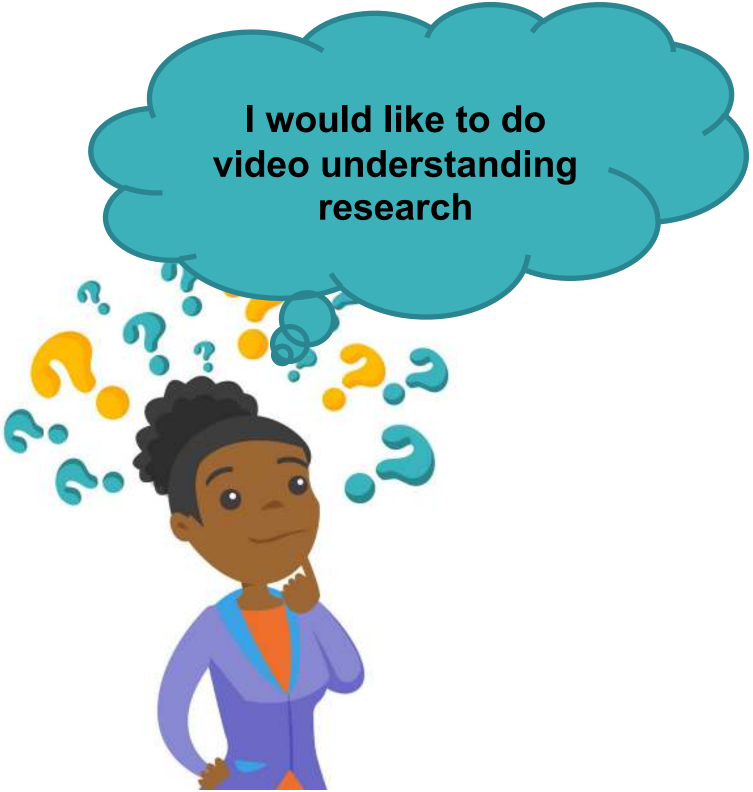
Video Understanding

An Egocentric Perspective

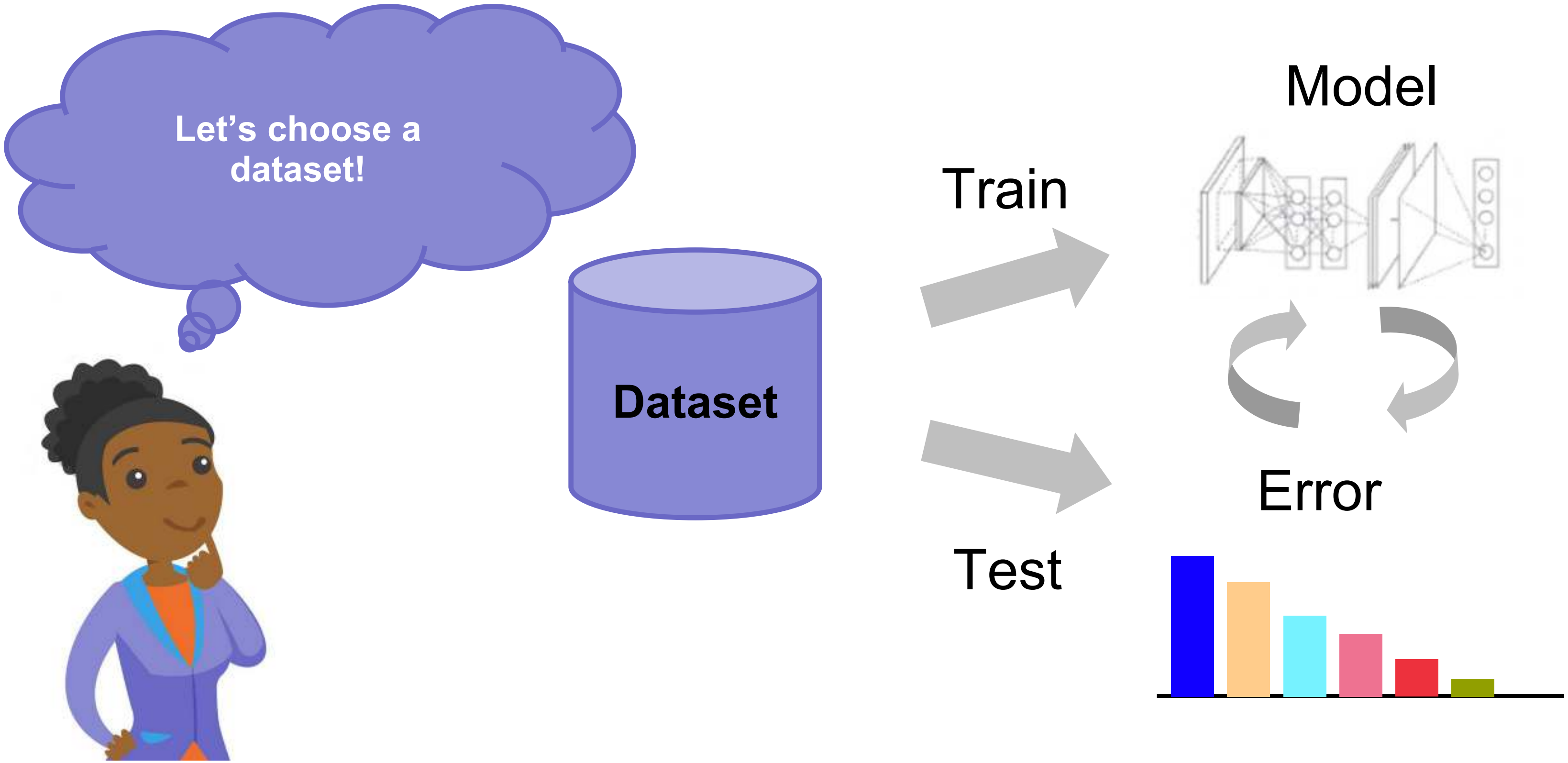
Introduction…



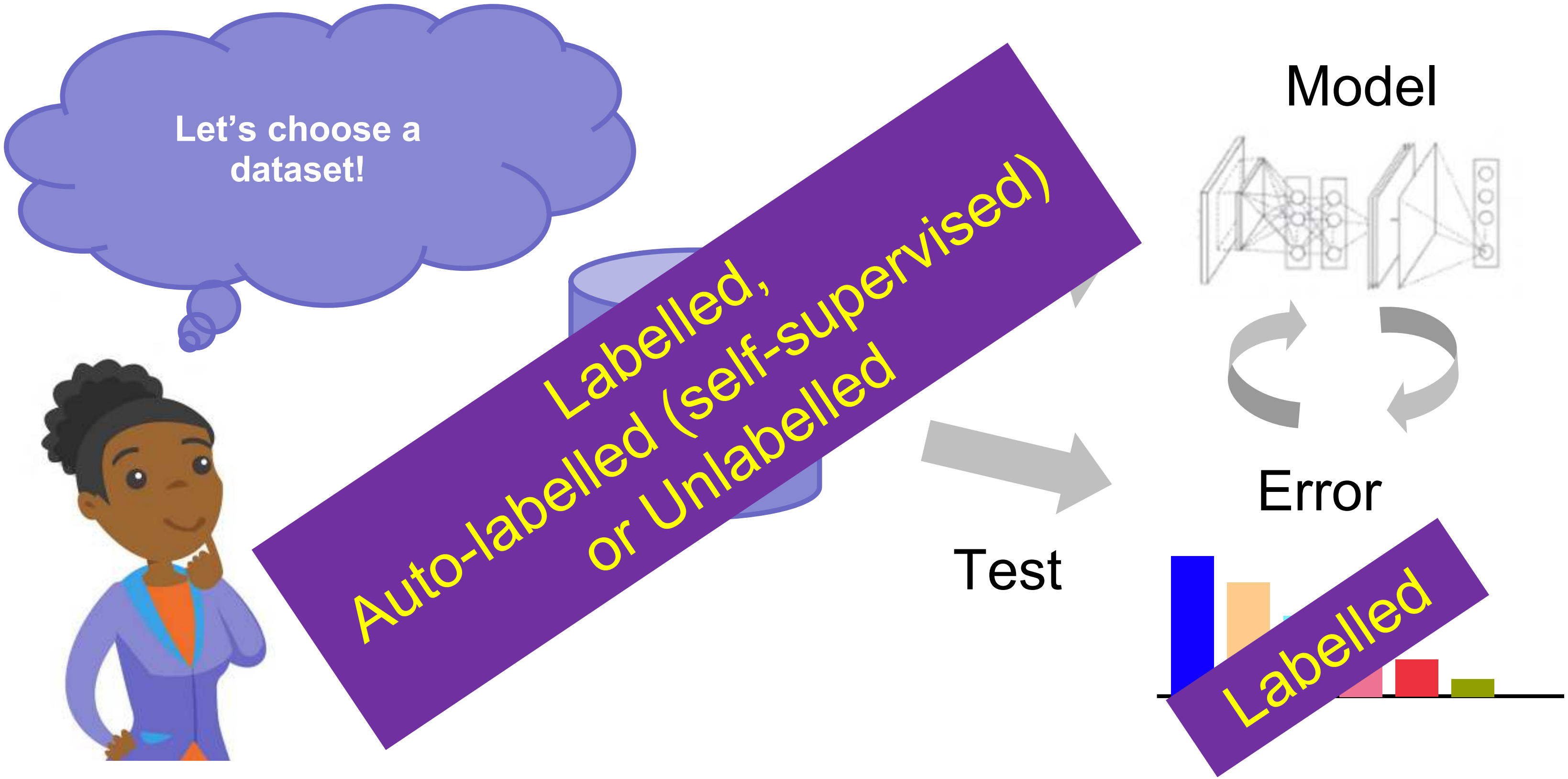
The current paradigm of Computer Vision Research



The current paradigm of Computer Vision Research



The current paradigm of Computer Vision Research



In this talk… on Video Understanding

Dataset

+

Labels

→

Tasks and Methods

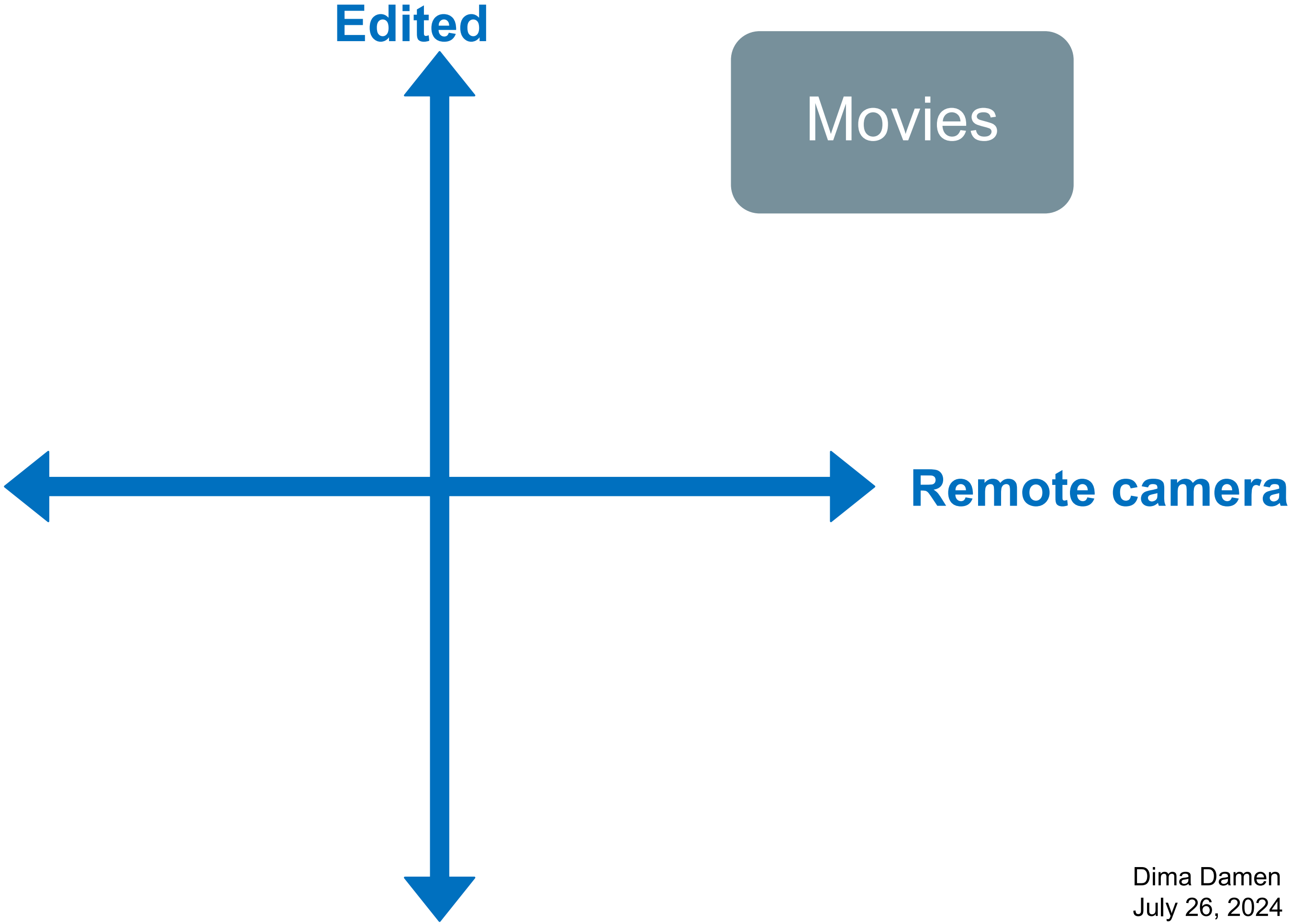
Dataset

Labels

Tasks and Methods

Part I: Collecting a Dataset

Our dataset is made up of… *videos*


Edited
Movies
Remote camera

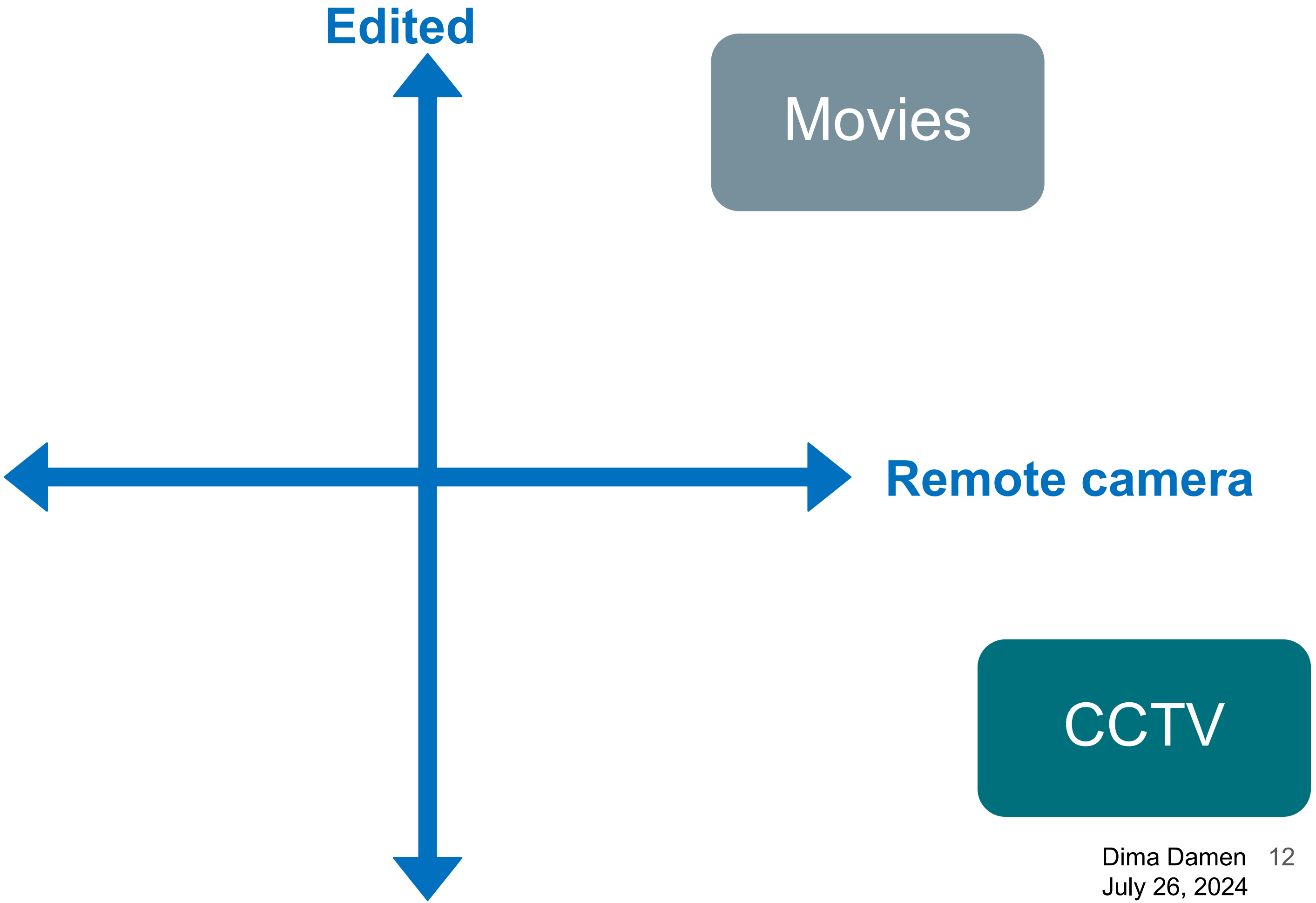
The history of VIDEO understanding

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

Laptev and Perez (2007)

The history of VIDEO understanding

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

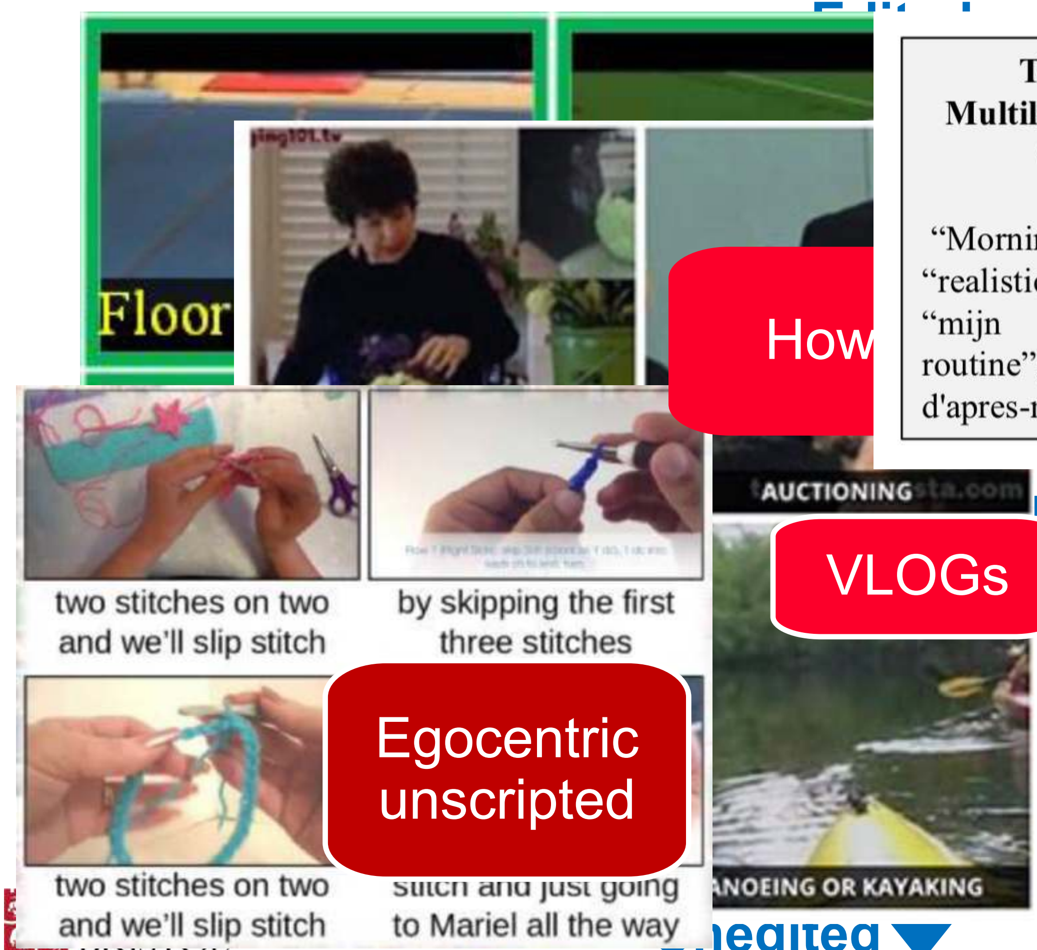


University of BRISTOL

The history of VIDEO understanding

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

The history of VIDEO understanding



How

VLOGs

Egocentric unscripted

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

YouTube Videos

CCTV

Unedited

The history of Video Understanding

Movies

TV (sports)

HowTo

VLOGs

YouTube Videos

Egocentric unscripted

CCTV

Video Understanding

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

The Egocentric Perspective

with: Kristen Grauman
+83 authors

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

Egocentric Videos?

Data Collection Exercises

EPIC KITCHENS

2017 - now

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

EGO 4D

2020 - now

6730 hours
923 participants
74 locations
9 countries
Short-term recording
All daily activities

Data Collection Exercises

EGO-EXO4D

2022 - now

Released Dec 2023
1422 hours
8 skilled activities
839 camera wearers
Ego-Exo recordings

2024 – [coming]

[new recordings]

Ego-Exo4D

with: Kristen Grauman
+102 authors

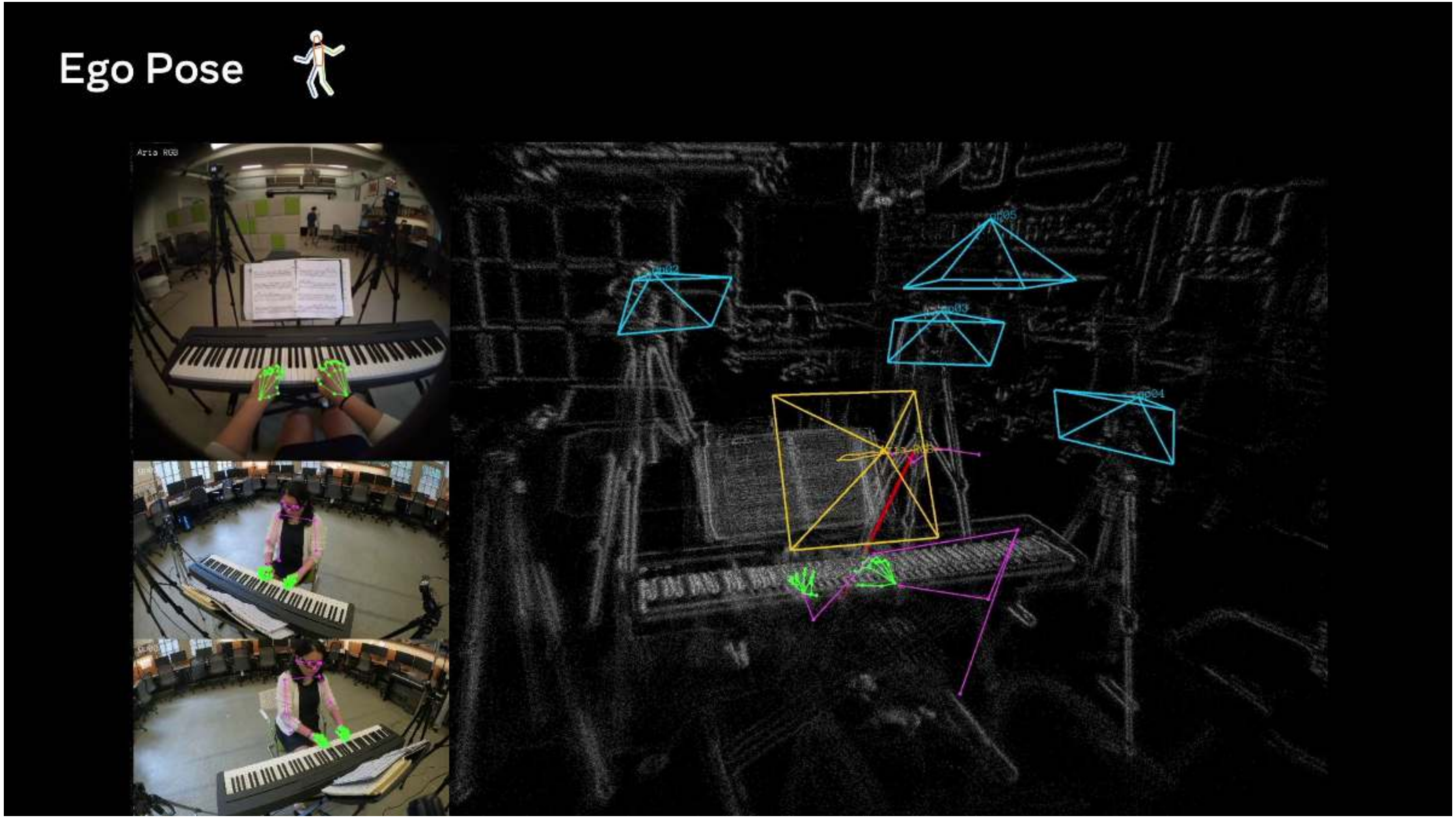
K Grauman et al (2024). Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives.. *CVPR*

with: Kristen Grauman
+102 authors

Ego-Exo Relation

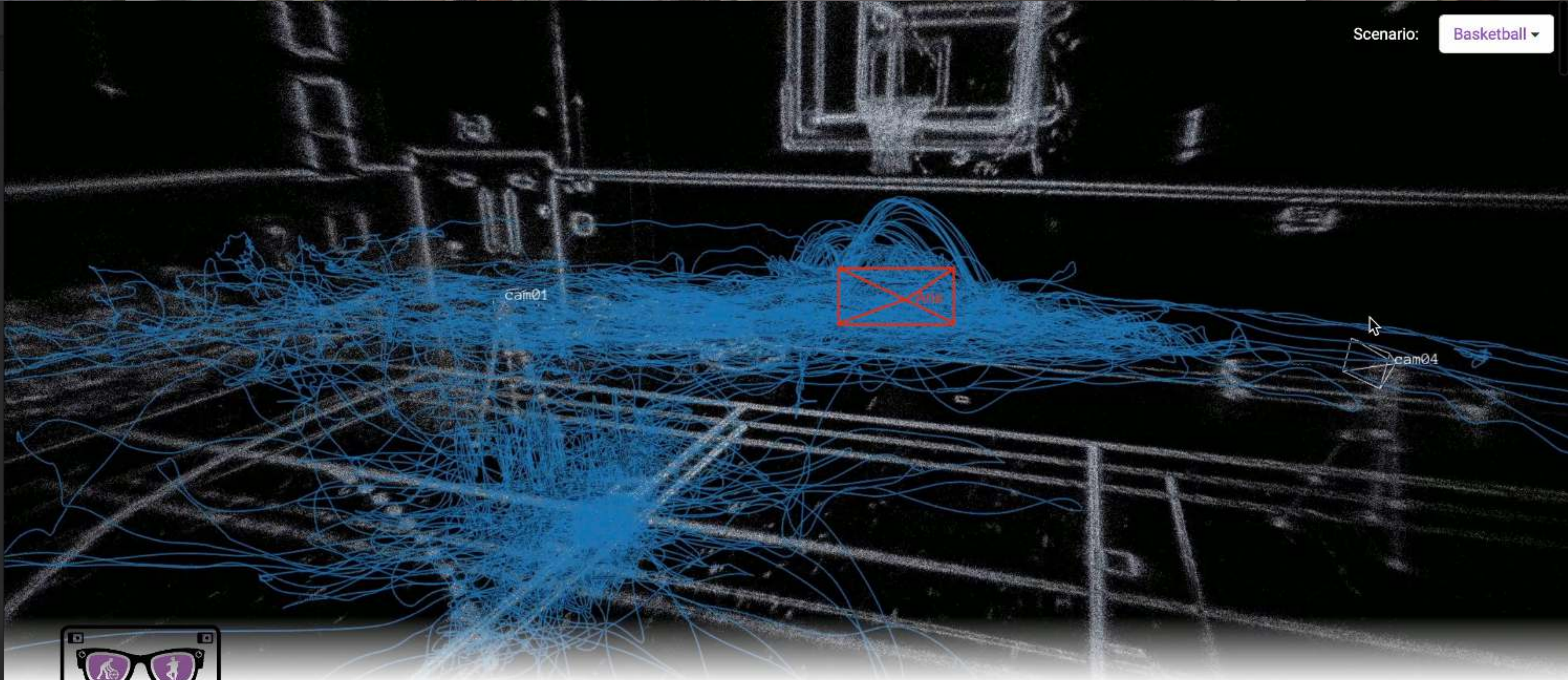
Ego-Exo4D

with: Kristen Grauman
+102 authors



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

Scenario: Basketball



EGO-EXO4D

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

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

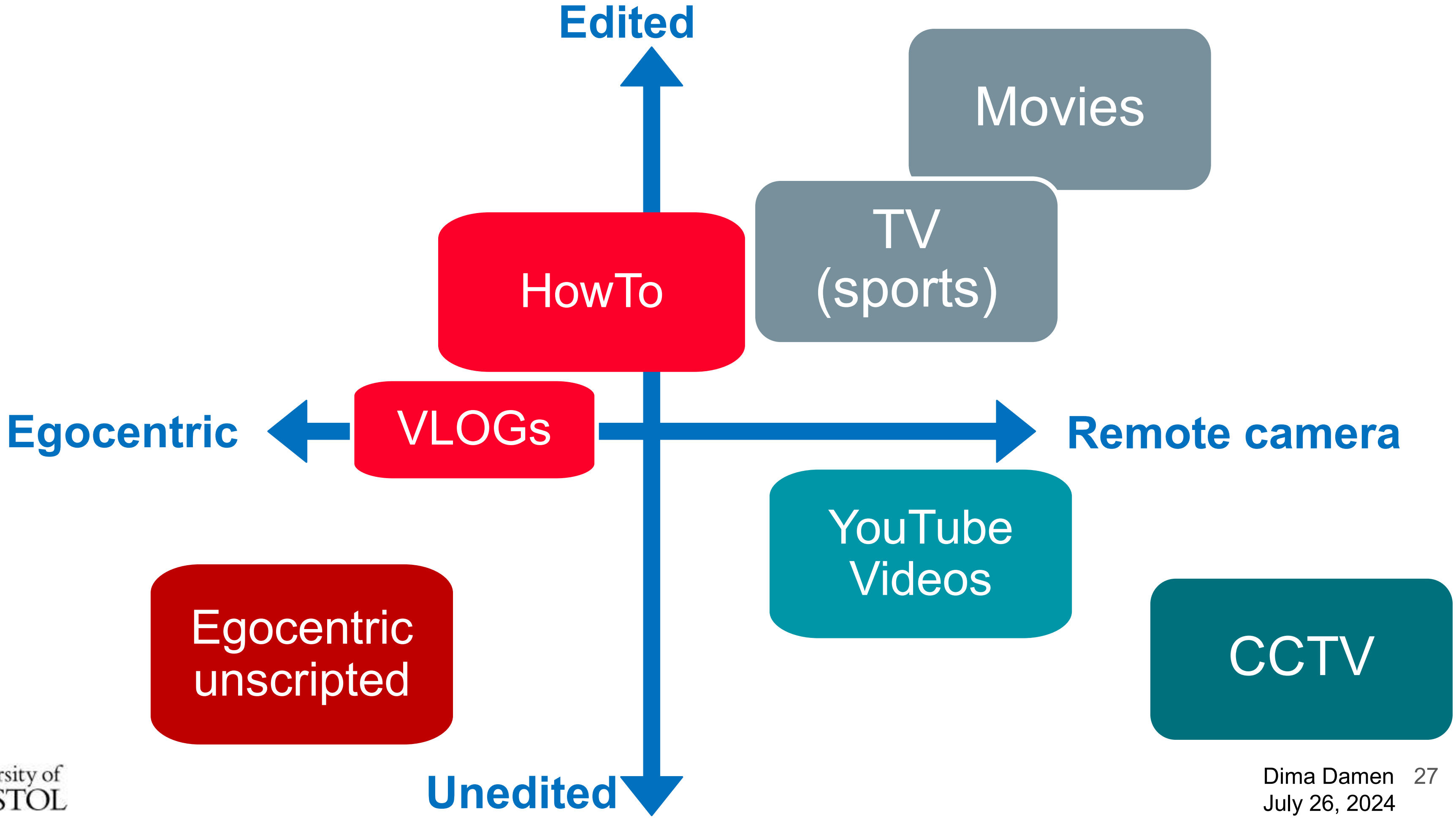
Learn More ↓ | Watch Video ↗ | Start Here ↗


WHAT
DO YOU
THINK?

If you were to do research in video understanding, which video type(s) would you explore? Why?

University of BRISTOL

The history of VIDEO understanding



sli.do

Joining as a participant?

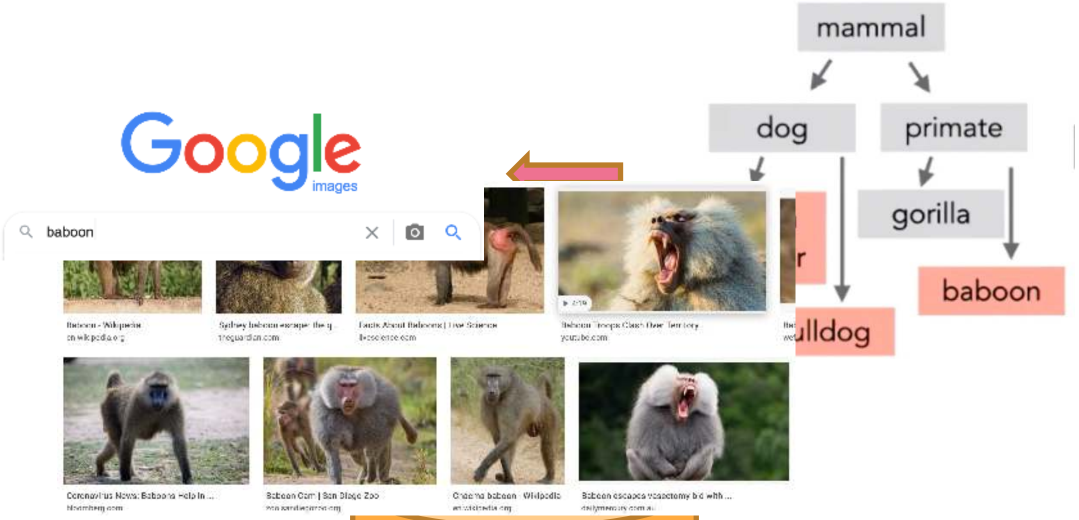
#3639 120

How to collect a dataset of videos?

Dima Damen
July 26, 2024

ImageNet Dataset

Object Recognition



Dataset

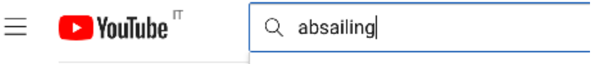
Kinetics Dataset

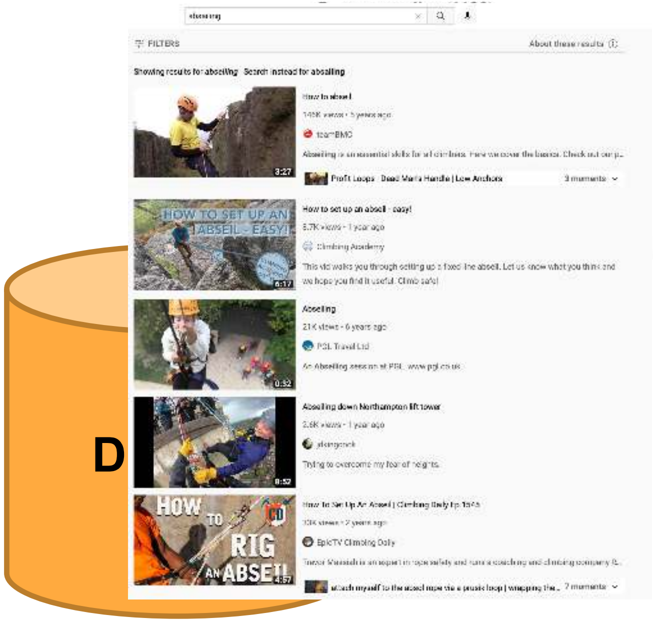
A. List of Kinetics Human Action Classes

This is the list of classes included in the human action video dataset. The number of clips for each action class is given by the number in brackets following each class name.

1. abseiling (1146)
2. air drumming (1132)
3. answering questions (478)
4. applauding (411)
5. applying cream (478)
6. archery (1147)

Action Recognition





Kinetics Dataset…

A. List of Kinetics Human Action Classes

This is the list of classes included in the human action video dataset. The number of clips for each action class is given by the number in brackets following each class name.

1. abseiling (1146)
2. air drumming (1132)
3. answering questions (478)
4. applauding (411)
5. applying cream (478)
6. archery (1147)
7. arm wrestling (1123)
8. arranging flowers (583)
9. assembling computer ([illegible])
10. auctioning (478)
11. baby waking up (611)
12. baking cookies (927)
13. balloon blowing (826)
14. bandaging (569)
15. barbequing (1070)

Statistics: The dataset has 400 human action classes, with 400–1150 clips for each action, each from a unique video. Each clip lasts around 10s. The current version has 306,245 videos, and is divided into three splits, one for training having 250–1000 videos per class, one for validation with 50 videos per class and one for testing with 100 videos per class. The statistics are given in table 2.

One Exception

Machine Learning in Practice

- Autonomous Driving…

Welcome to the KITTI Vision Benchmark Suite!

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

Share

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

stereo flow sceneflow depth odometry object tracking road semantics raw data

Machine Learning in Practice

~~Object Recognition~~

Let’s collect Data!

EPIC KITCHENS-100

Scaling and Rescaling Egocentric Vision:
The EPIC-KITCHENS Dataset


Dima Damen


Hazel Doughty


Giovanni M. Farinella


Sanja Fidler


Antonino Furnari


Evangelos Kazakos


Jian Ma


Davide Moltisanti


Jonathan Munro


Toby Perrett


Will Price


Michael Wray

EPIC-KITCHENS

Scaling and Rescaling Egocentric Vision

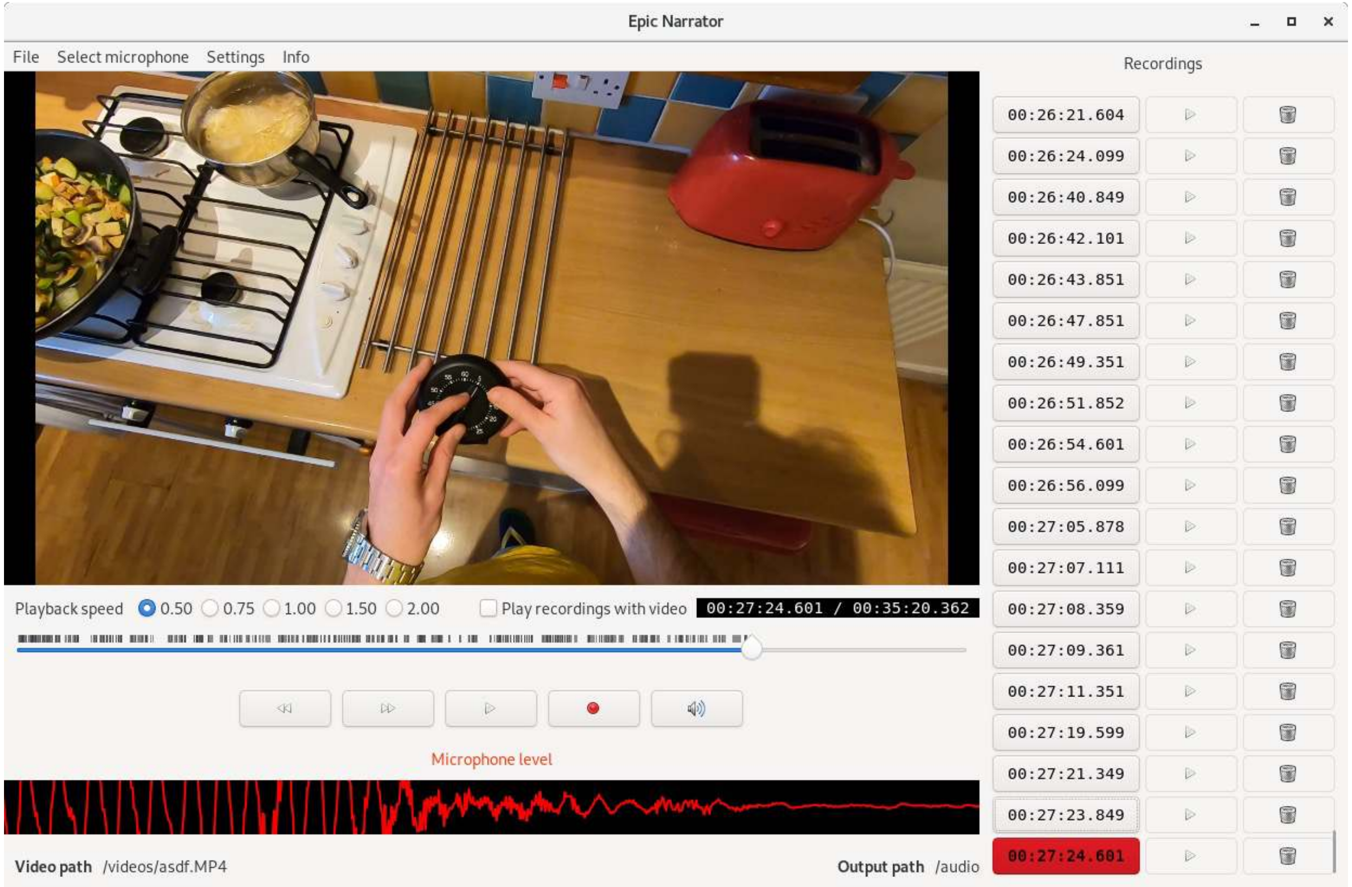
- Head-Mounted Go-Pro, adjustable mounting
- Recording starts immediately before entering the kitchen
- Only stopped before leaving the kitchen

EPIC-KITCHENS


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


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

EPIC-KITCHENS



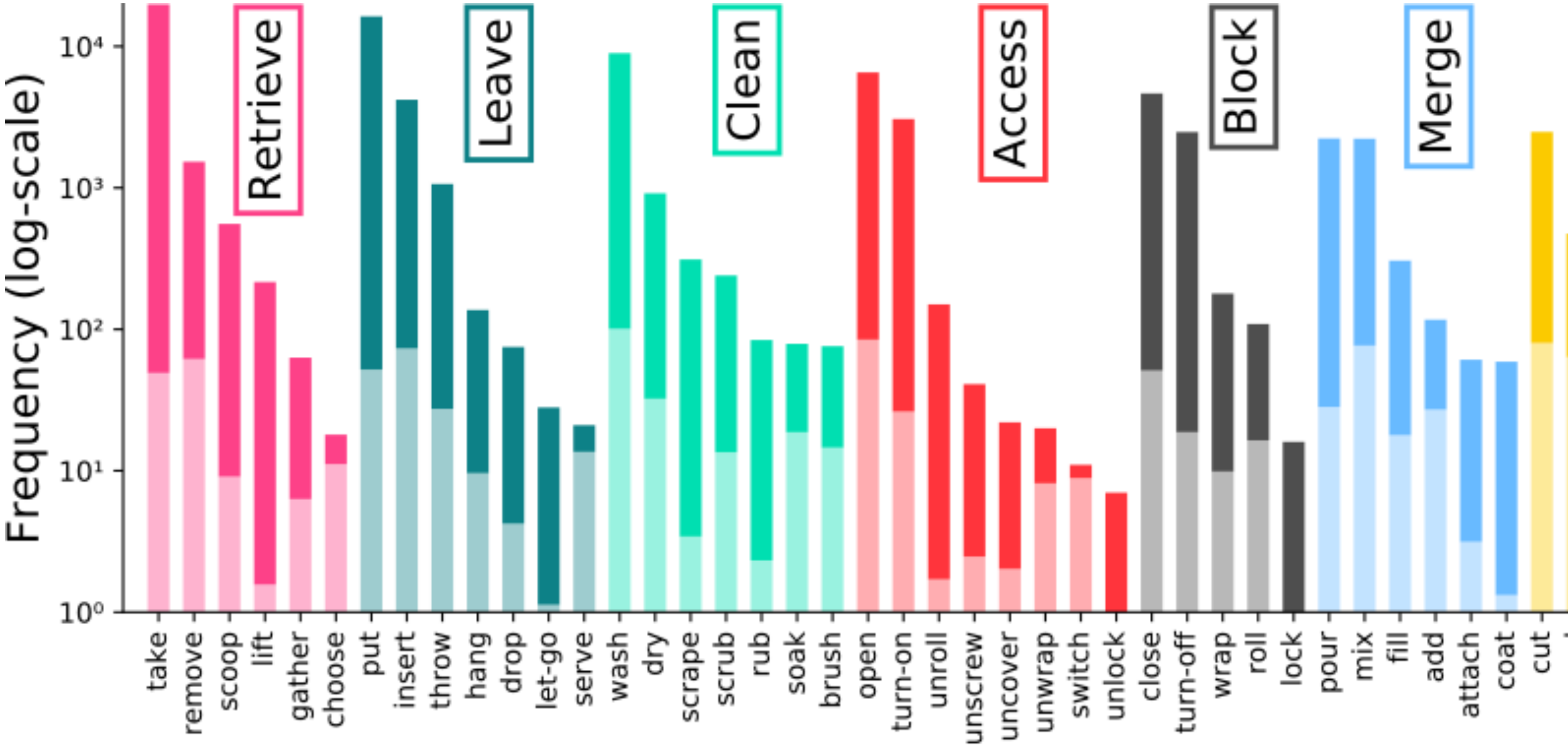
EPIC-KITCHENS and Ego4D

Store
Lay-down **PUT**
Place Rest-on

stove burner
hob
stovetop gas

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

EPIC-KITCHENS-100 Statistics



with: Kristen Grauman
+83 authors

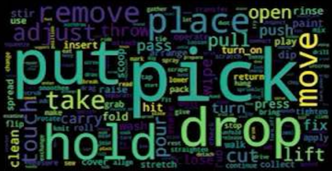
Narration

C: camera wearer

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

13.2 sentences/min
3.8 M sentences

1,772 verbs



4,336 nouns

Data Collection Exercise

Pascal VOC
ImageNet
Kinetics
Something-Something

Labels

Data

EPIC-KITCHENS
Ego4D
…
KITTI

The chicken or the egg...

Data

Naturally unbalanced

Harder to label (exposes ambiguity)

Closer to application

Multiple tasks

Labels

Unnaturally balanced (or nearly)

Easier to label (hides ambiguity)

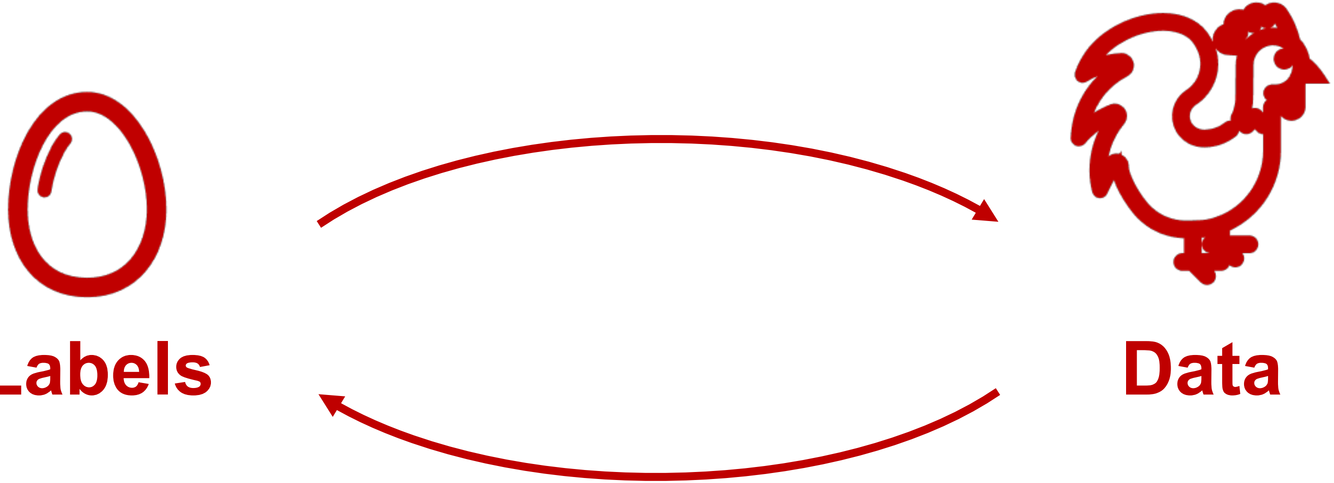
Can be expanded

Single task

University of BRISTOL

Data Collection Exercise

What should come first? Labels or Data



sli.do

Joining as a participant?

#3639 120

Video source and data collection approach heavily influences video understanding tasks

Dataset

+

Labels

Tasks and Methods

Part II: Labelling a Dataset

What type of labels can we provide?

- Temporal labels – Strong vs. Weak labels
- Semantic labels – Open-vocab. vs Closed-vocabulary
- Ranking labels – video-to-video comparisons
- Pixel-level labels – segmentation labels

What type of labels can we provide?

- Temporal labels – Strong vs. Weak labels
- Semantic labels – Open-vocab. vs Closed-vocabulary
- Ranking labels – video-to-video comparisons
- Pixel-level labels – segmentation labels

Action Recognition Challenge

Given a trimmed action segment:
$(t_{\text{start}}, t_{\text{stop}})$
classify the action within.

$\hat{y}_{\text{verb}} = $ open

$\hat{y}_{\text{noun}} = $ oven

$\hat{y}_{\text{action}} = $ (open, oven)

University of BRISTOL


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


280
309
322
348
280
309
322
348


280
322
309
348
280
309
322
348

with: Davide Moltisanti

Inconsistencies of temporal bounds across datasets for the same action

BEOID: take cup

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

GTEA Gaze+

ground truth

predicted class: take knife

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

with: Davide Moltisanti

How to achieve consistent temporal labels?

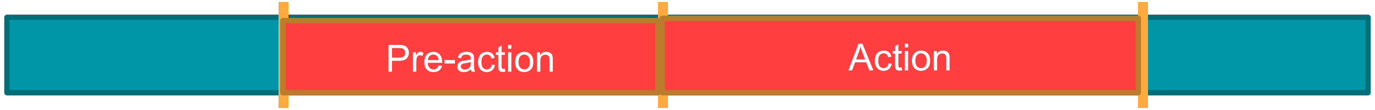
D Moltisanti et al (2017). Trespassing the Boundaries: Labeling Temporal Bounds for Object Interactions in Egocentric Video. *International Conference on Computer Vision (ICCV).*

Dima Damen
July 26, 2024

The Rubicon Boundaries

with: Davide Moltisanti

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



initiate motion

Point of no return

Goal completed

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

The Rubicon Boundaries

with: Davide Moltisanti

Cut pepper (GTEA Gaze+)

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

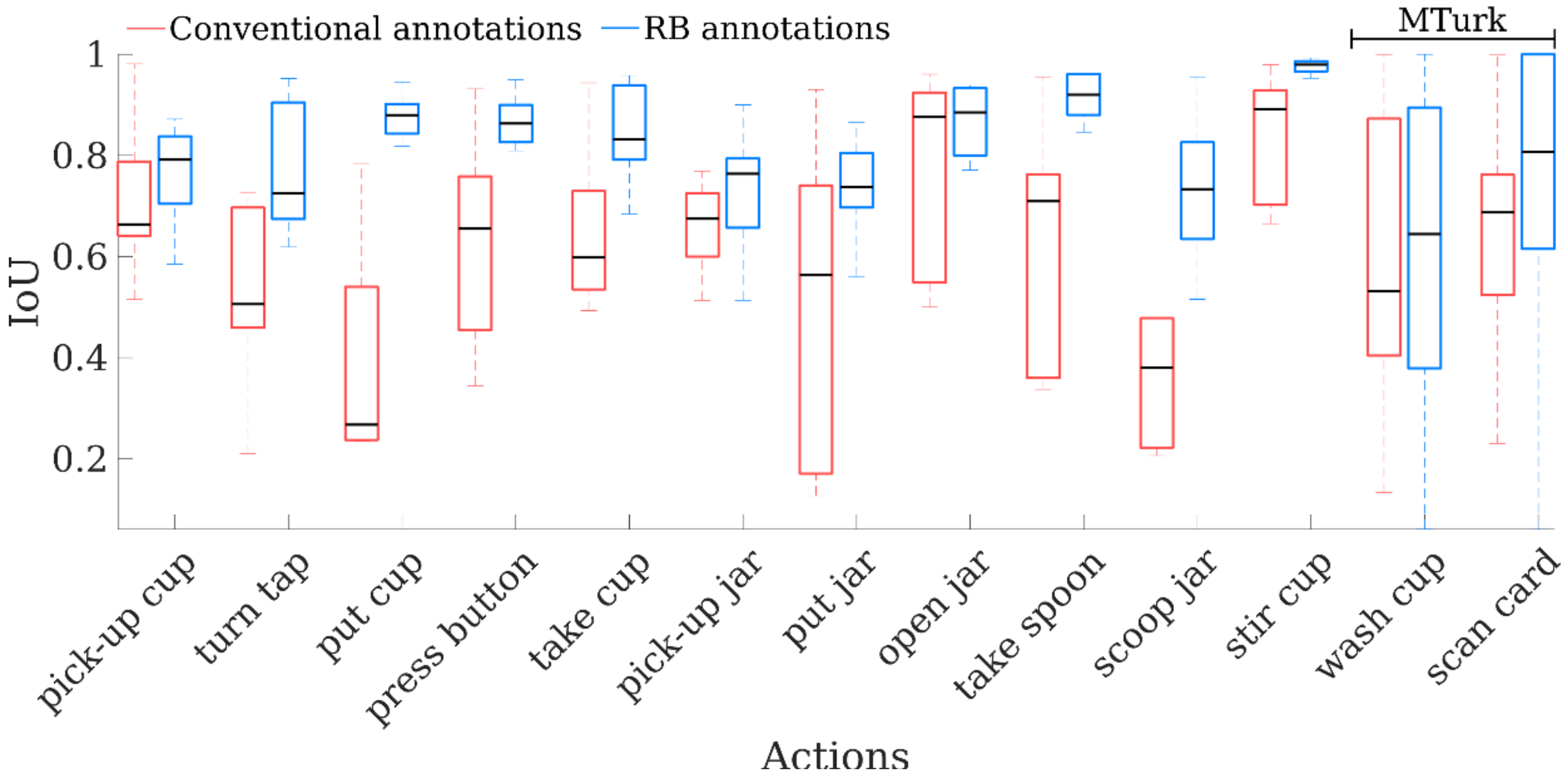
Rubicon Boundaries

Now we show some object interactions segmented
by multiple annotators using conventional labeling,
along with the same actions labeled by different annotators
following the Rubicon Boundaries (ref. Figure 3).

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



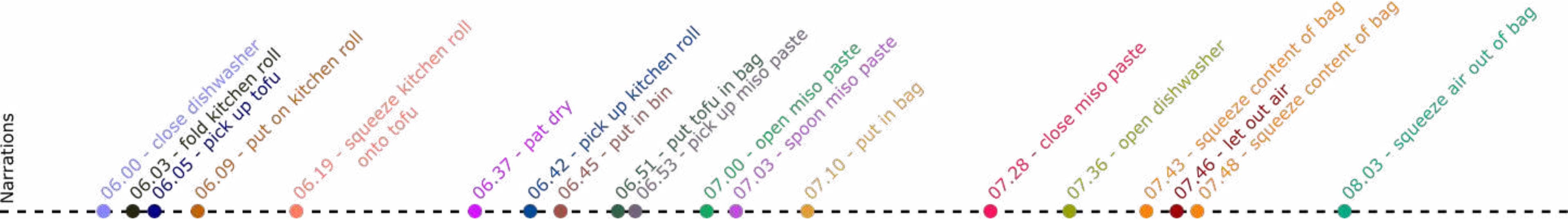
D Moltisanti et al (2017). Trespassing the Boundaries: Labeling Temporal Bounds for Object Interactions in Egocentric Video. *International Conference on Computer Vision (ICCV).*

The power of temporal labels

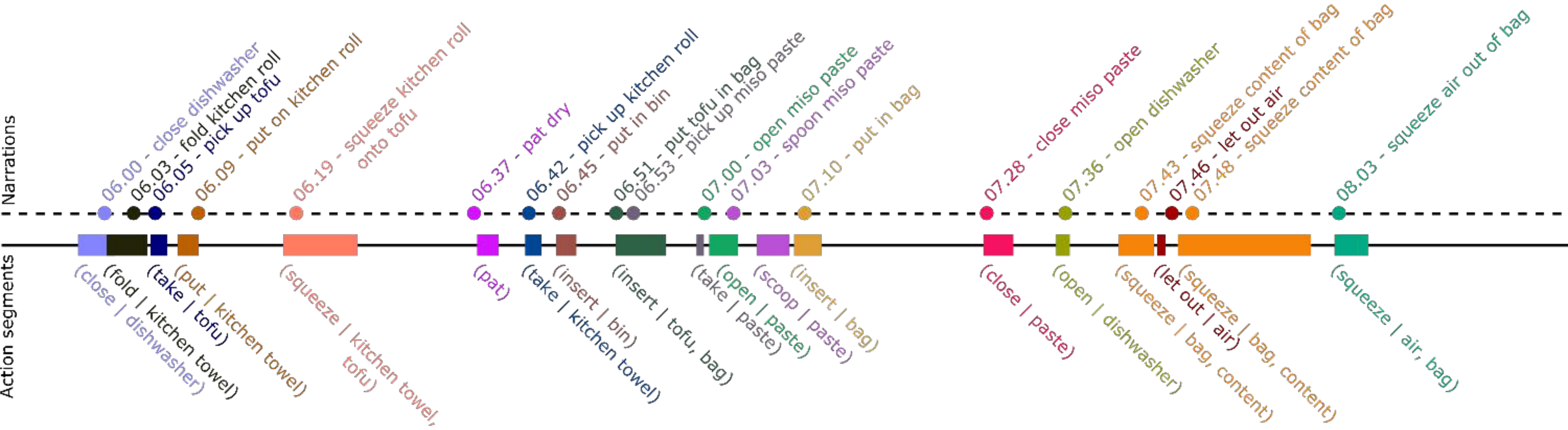
Other approaches to temporal boundaries

- Start-and-end times
  - Inconsistent
  - Consistent
- Fixed segment lengths
  - Kinetics Dataset -- 10 seconds videos
  - Moments in Time Dataset – 3 seconds videos
- No temporal annotations
  - Charades Dataset – Video-Level supervision (3-4 actions per video)
- Single-timestamp supervision

Scaling and Rescaling Egocentric Vision

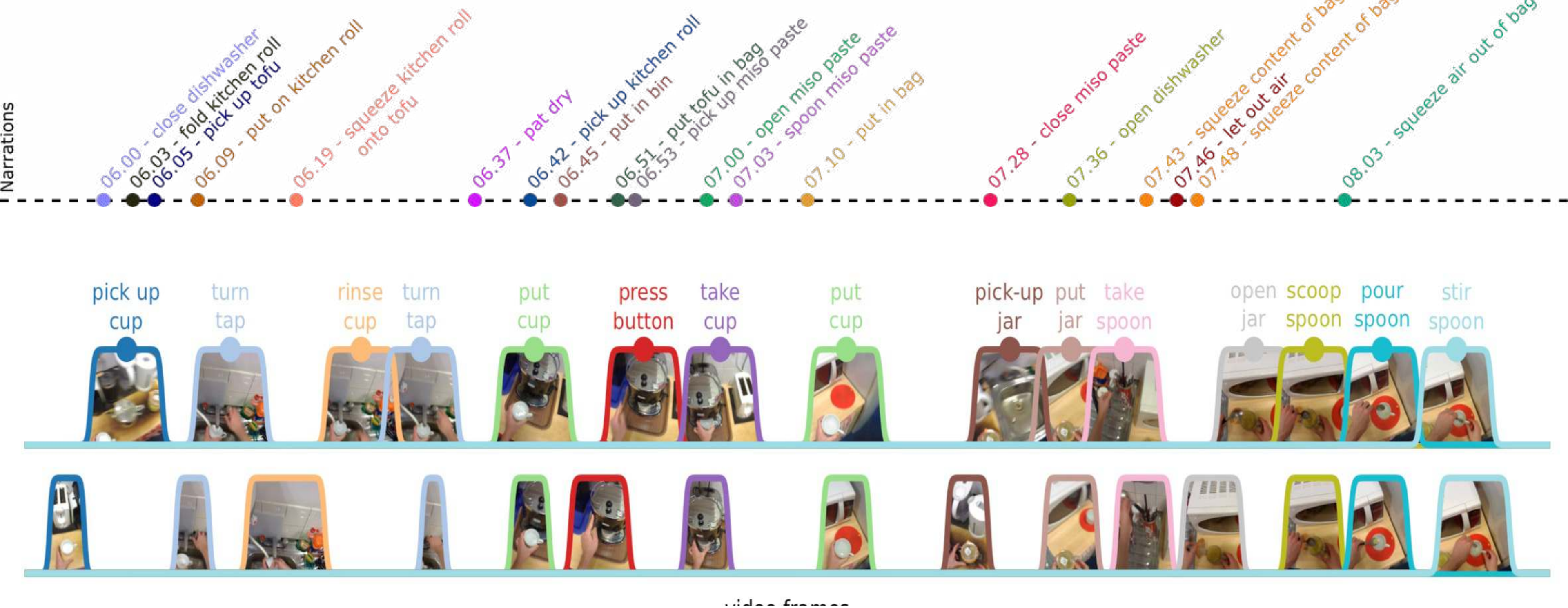


Scaling and Rescaling Egocentric Vision



Learning from a Single Timestamp

with: Davide Moltisanti
Sanja Fidler



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

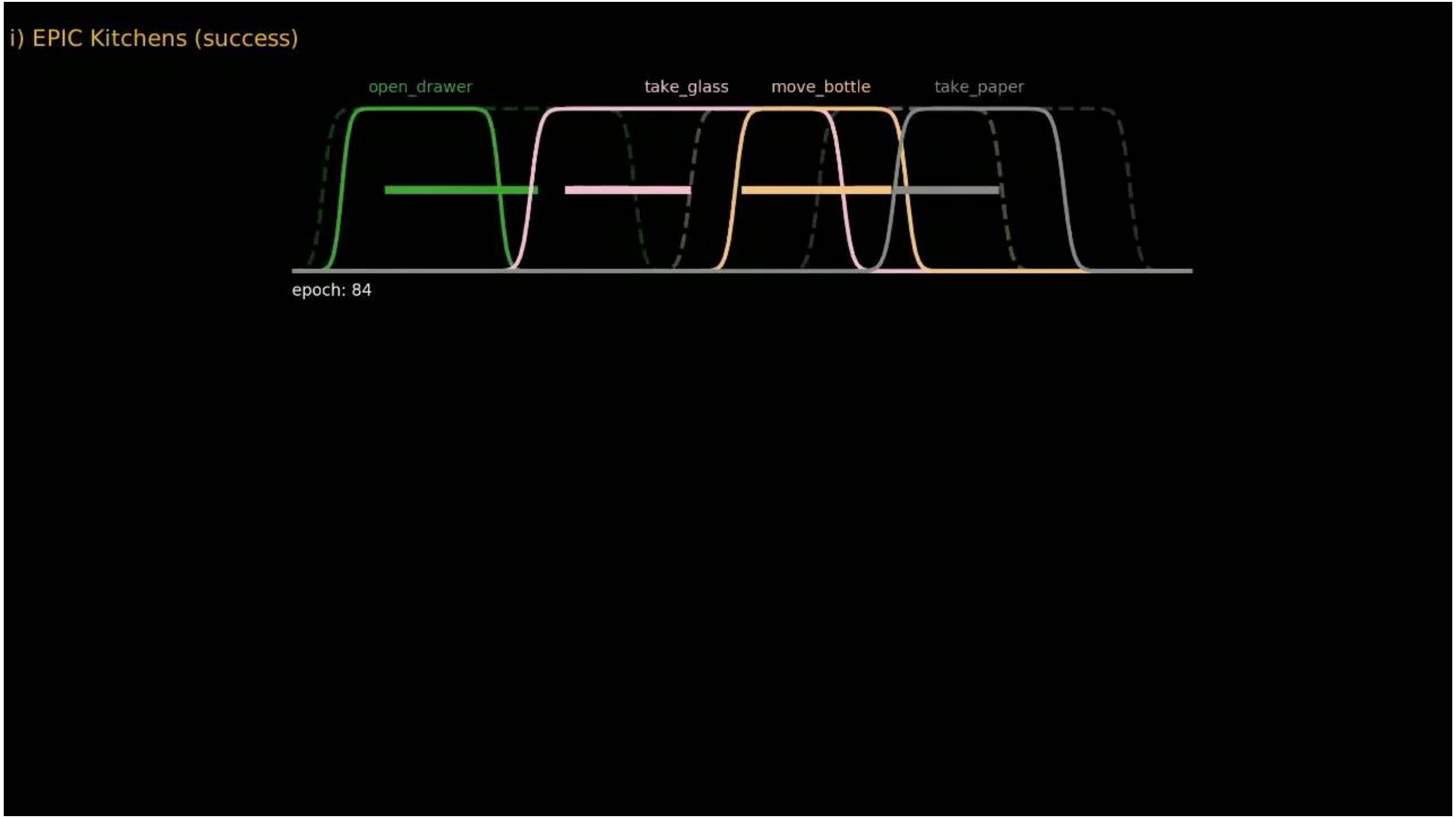
Learning from a Single Timestamp

with: Davide Moltisanti
Sanja Fidler

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

Learning from a Single Timestamp

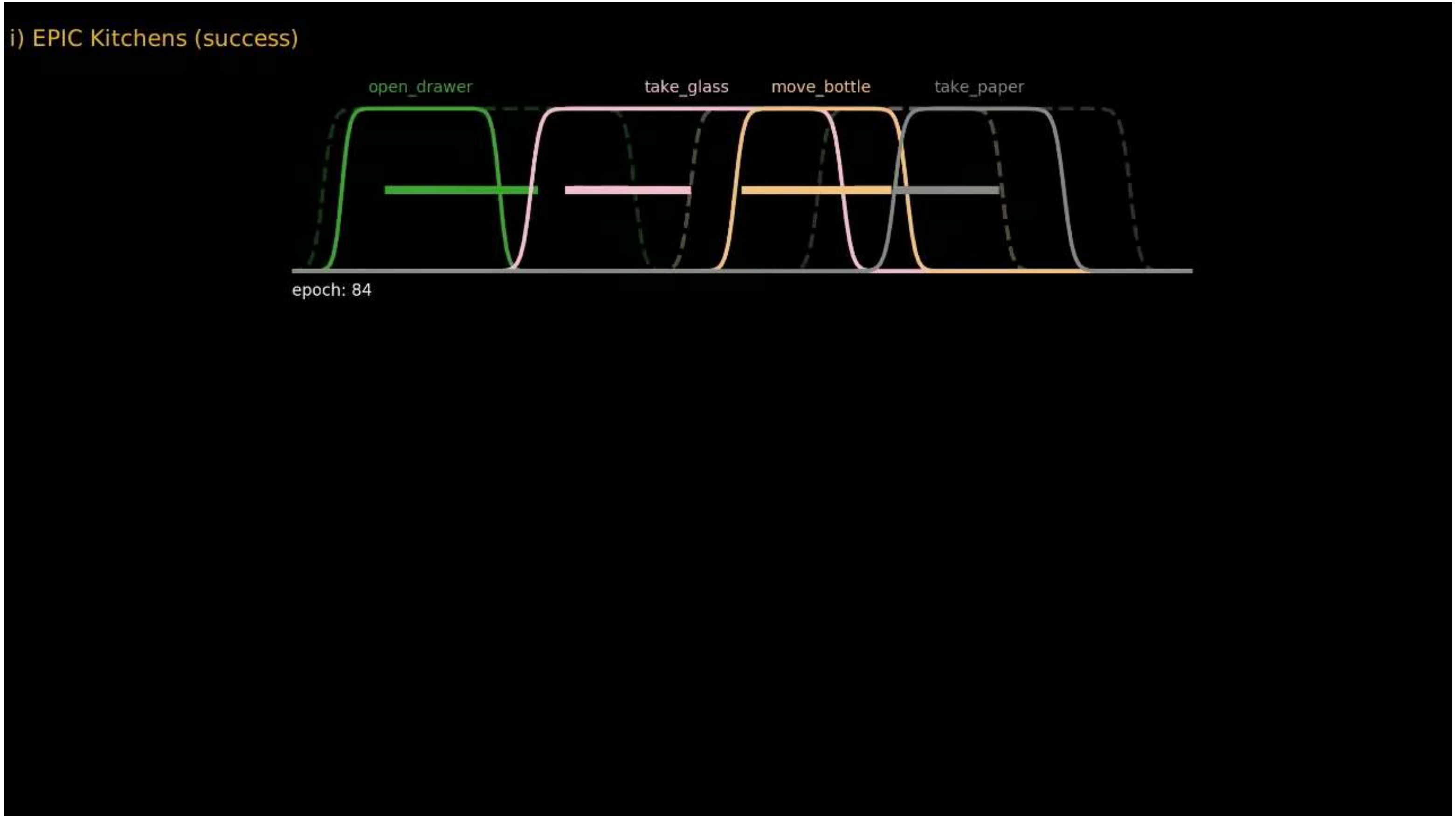
with: Davide Moltisanti
Sanja Fidler



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

Learning from a Single Timestamp

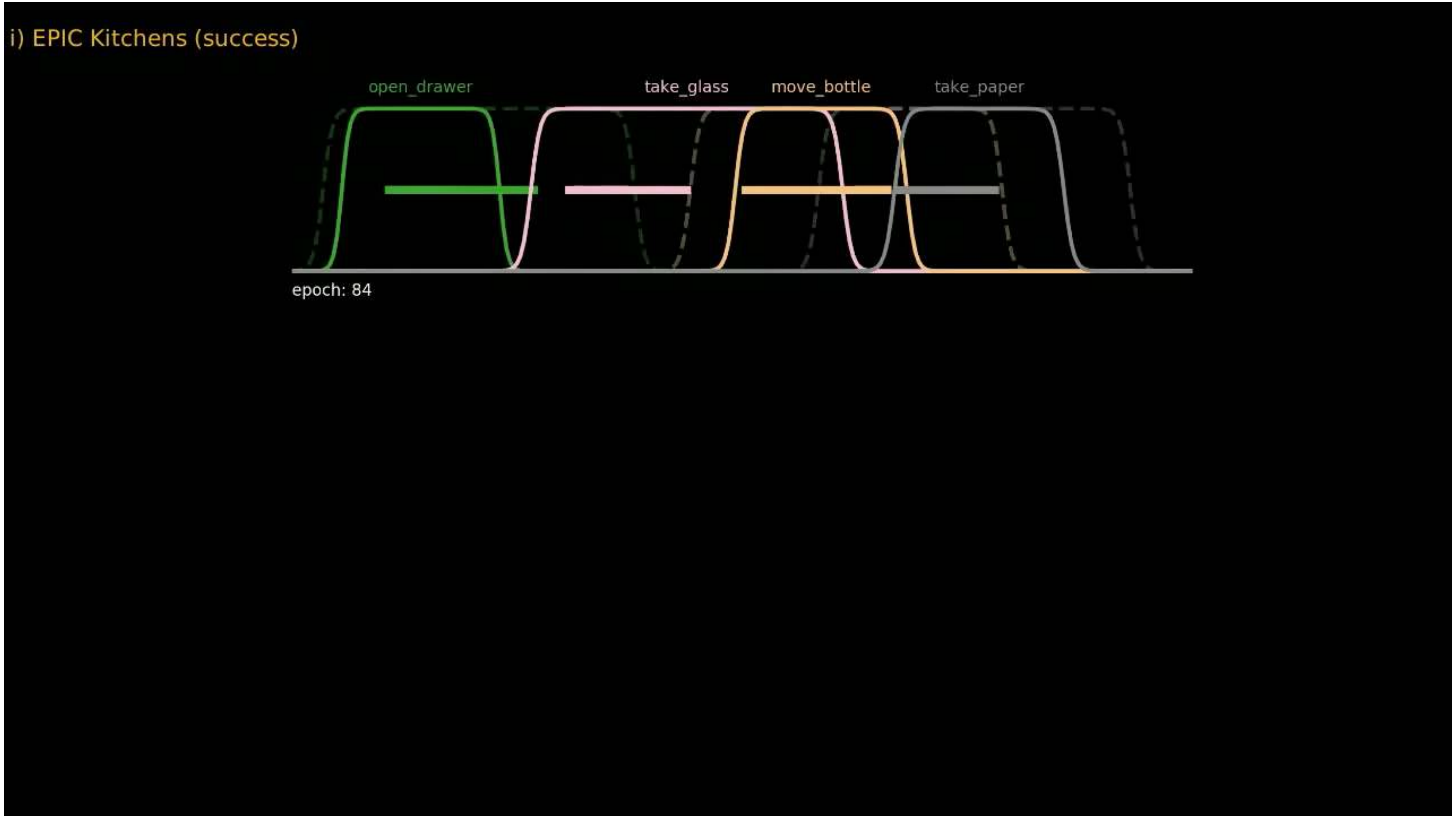
with: Davide Moltisanti
Sanja Fidler



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

Learning from a Single Timestamp

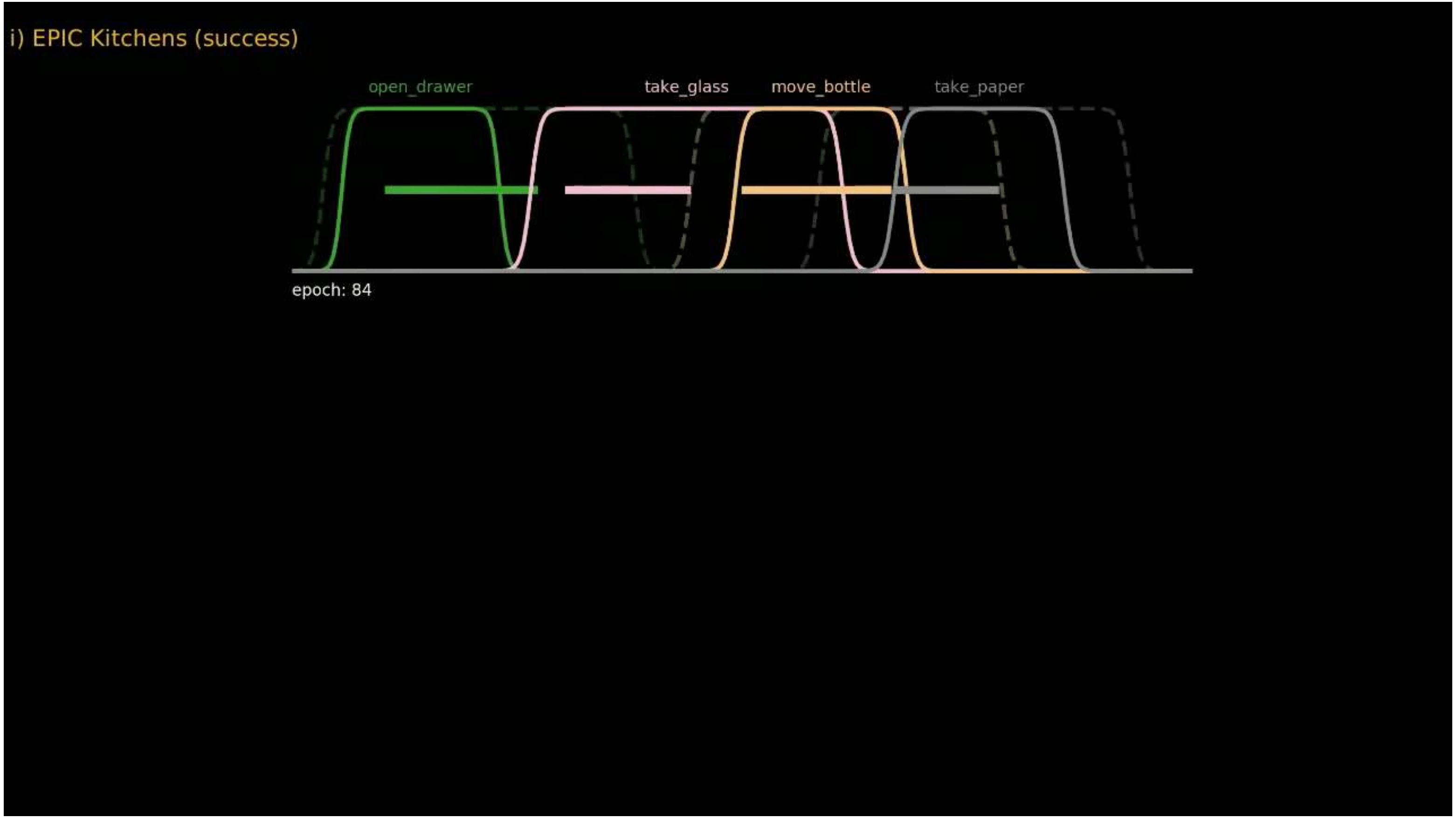
with: Davide Moltisanti
Sanja Fidler



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

Learning from a Single Timestamp

with: Davide Moltisanti
Sanja Fidler



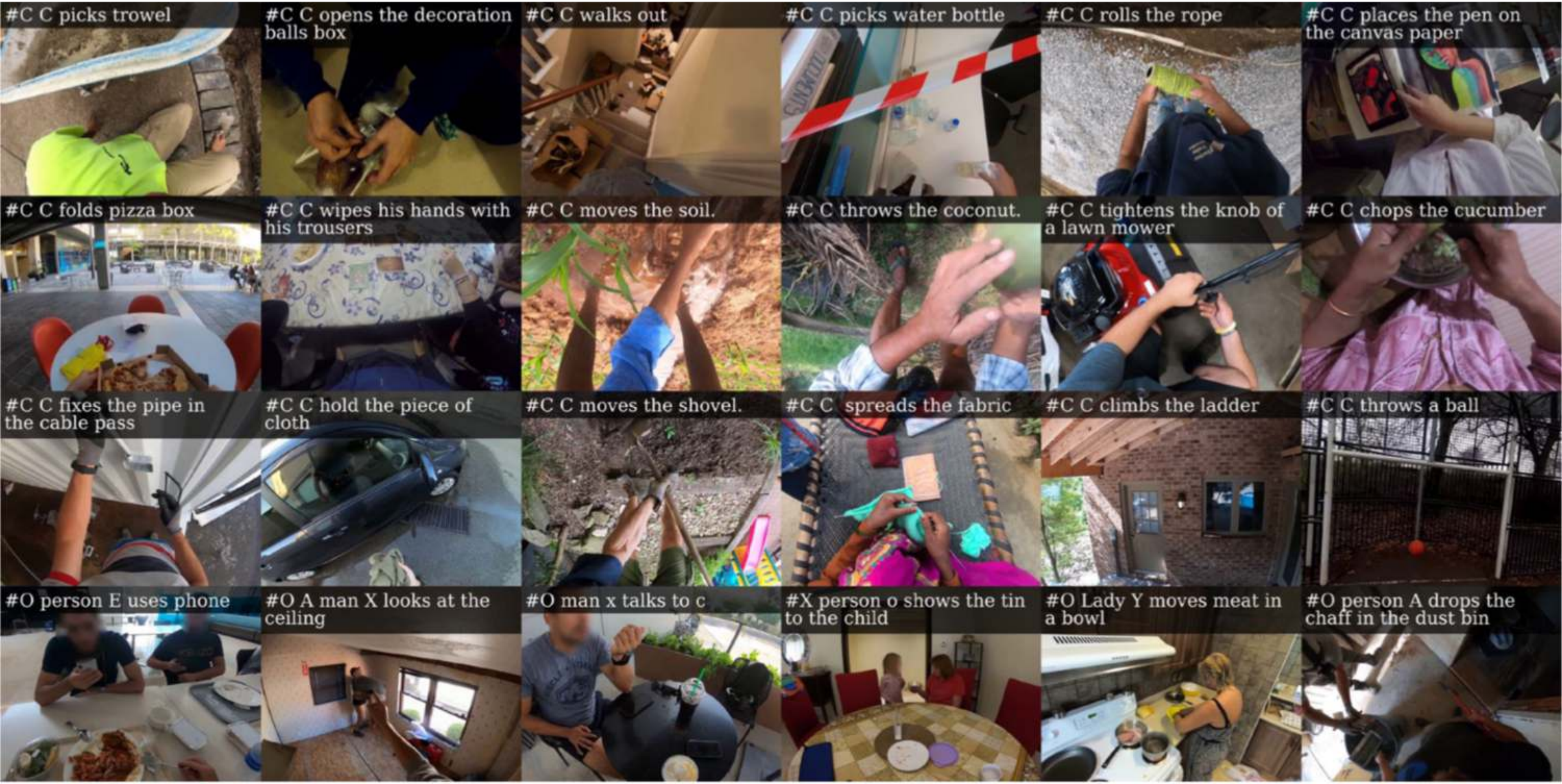
D Moltisanti et al (2019). Action Recognition from Single Timestamp Supervision in Untrimmed Videos. *Computer Vision and Pattern Recognition (CVPR)*

Learning from a Narration Timestamps

- Miech et al (2019). HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips

Ego4D

with: Kristen Grauman
+83 authors



Temporal labels vary across datasets… and should be consistent

Reversing Time

with: Will Price

W Price, D Damen (2019). Retro-Actions: Learning 'Close' by Time-Reversing 'Open' Videos. ICCV MDALC Workshop

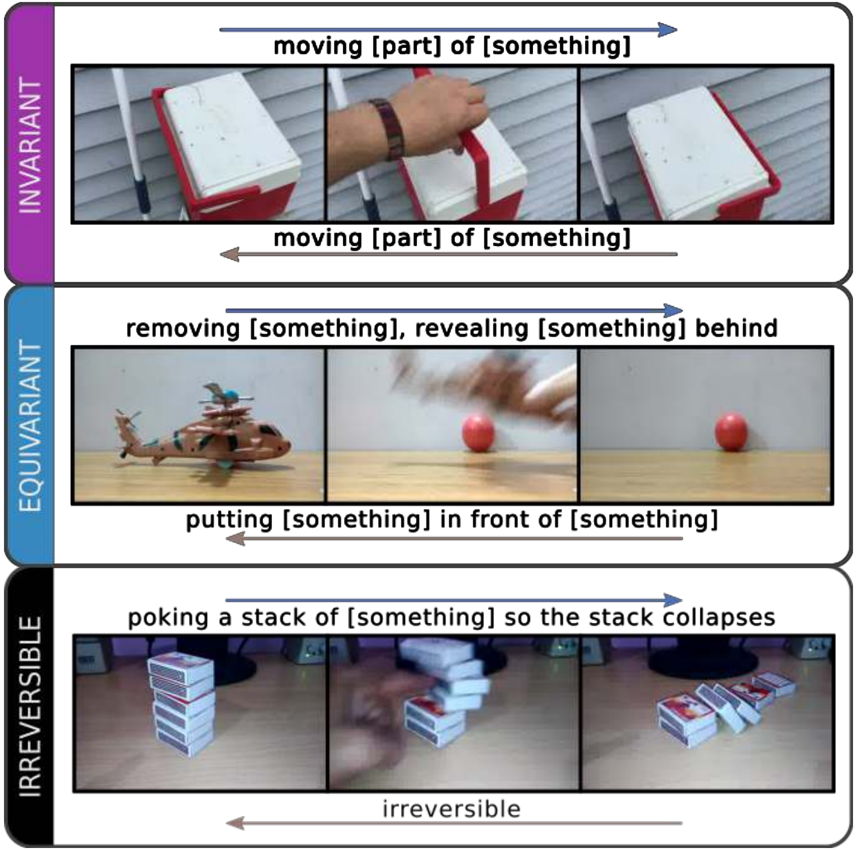
Reversing Time

This CVPR2014 paper is the Open Access version, provided by the Computer Vision Foundation.
The authoritative version of this paper is available in IEEE Xplore.

Seeing the Arrow of Time

Lyndsey C. Pickup[1] Zheng Pan[2] Donglai Wei[3] YiChang Shih[3] Changshui Zhang[2]
Andrew Zisserman[1] Bernhard Schölkopf[4] William T. Freeman[3]

Reversing Time



W Price, D Damen (2019). Retro-Actions: Learning 'Close' by Time-Reversing 'Open' Videos. ICCV MDALC Workshop

Reversing Time

with: Will Price

Open Close

Pull → Push ←

Camera ← Camera →

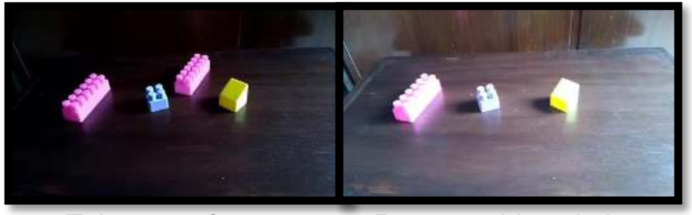
Moving _ and _ so they pass each other

Trying to bend _ unbendable

Removing _ revealing _ behind

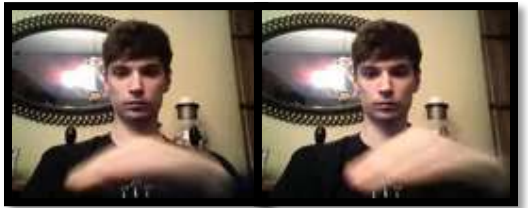
Putting _ in front of _

Swipe Swipe



Take one of many similar things on the table

Put something similar to other things already on the table



Roll hand forward

Roll hand backward

University of BRISTOL

Dima Damen
July 26, 2024

Reversing Time

with: Will Price

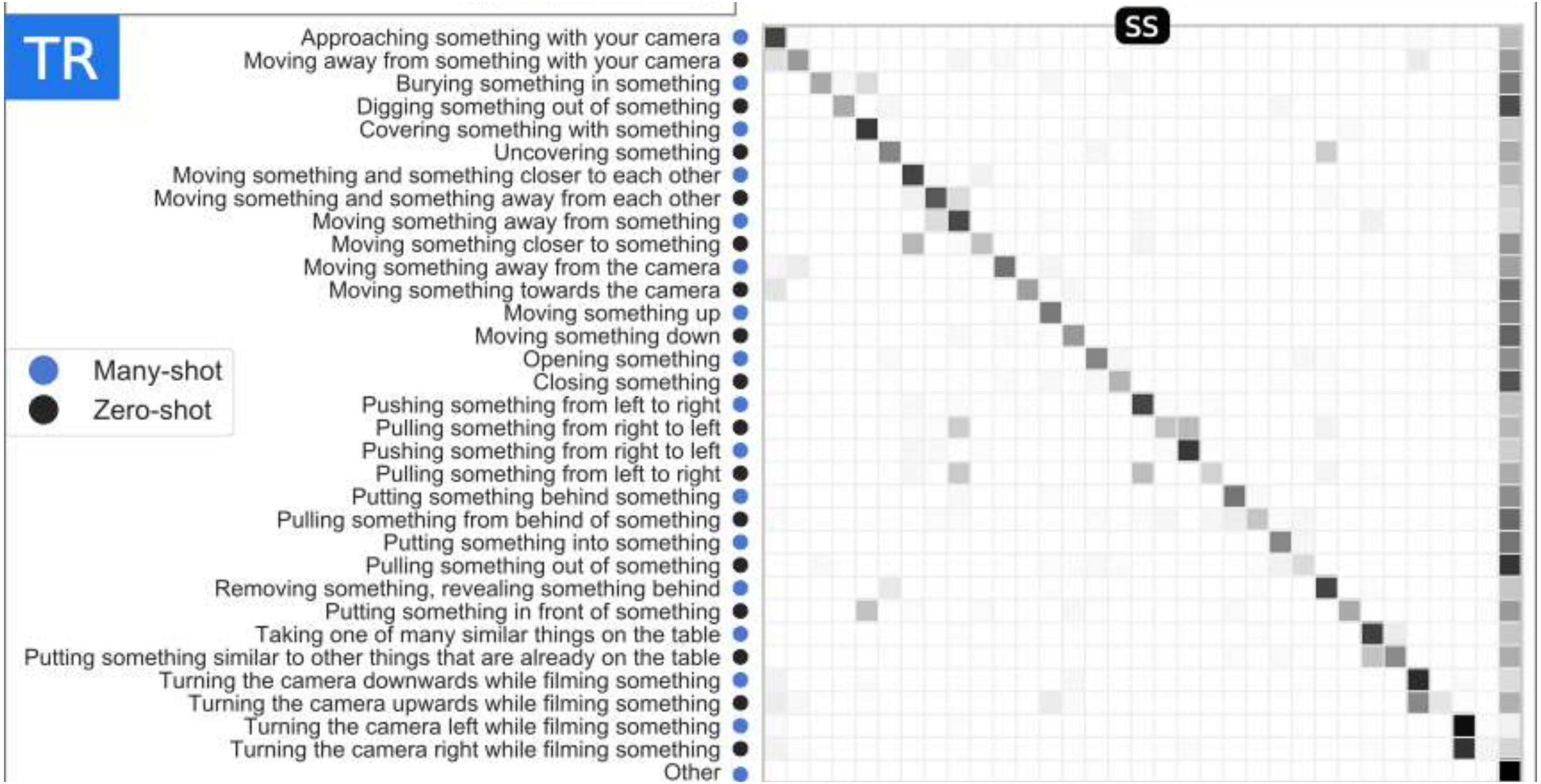
Can we trai

Learning ‘Close’ by time-reversing ‘Open’

W Price, D Damen (2019). Retro-Actions: Learning ‘Close’ by Time-Reversing ‘Open’ Videos. ICCV MDALC Workshop

Reversing Time

with: Will Price



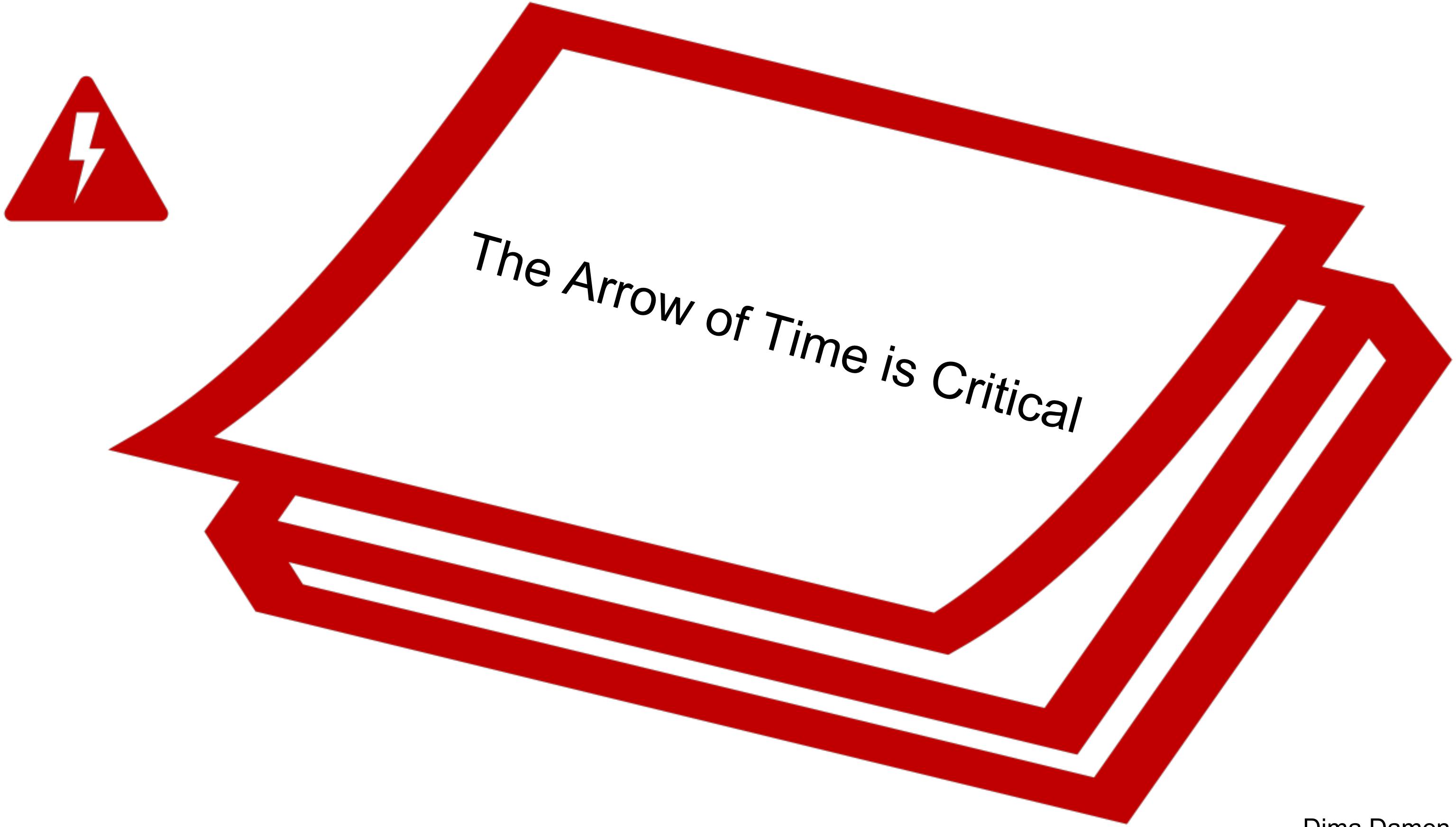
W Price, D Damen (2019). Retro-Actions: Learning ‘Close’ by Time-Reversing ‘Open’ Videos. ICCV MDALC Workshop

Reversing Time

with: Will Price

Now, results from a model supervised by
time-reversal example synthesis

W Price, D Damen (2019). Retro-Actions: Learning 'Close' by Time-Reversing 'Open' Videos. ICCV MDALC Workshop


The Arrow of Time is Critical


Videos are multi-modal

Multi-modal learning…

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

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

Multi-modal learning…

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

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

Multi-modal learning…

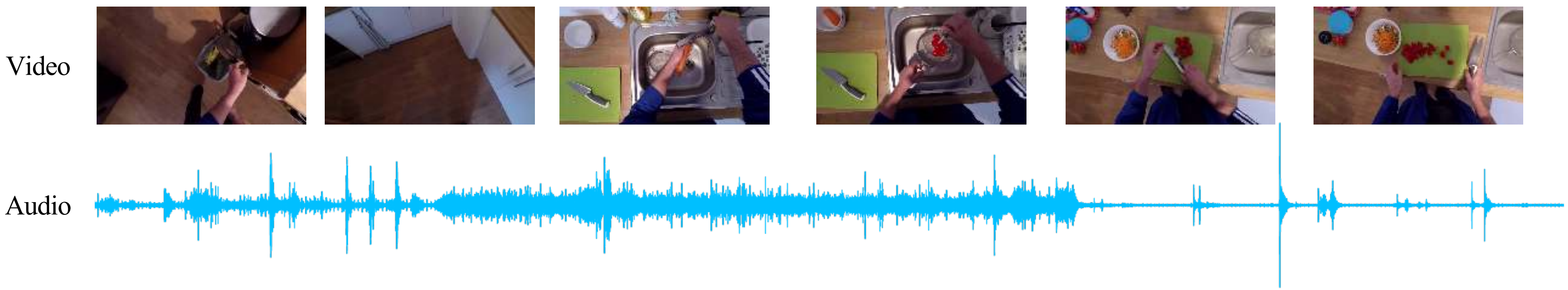
with: Vangelis Kazakos Jaesung Huh
Arsha Nagrani. Jacob Chalk
Andrew Zisserman

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

University of BRISTOL

Motivation

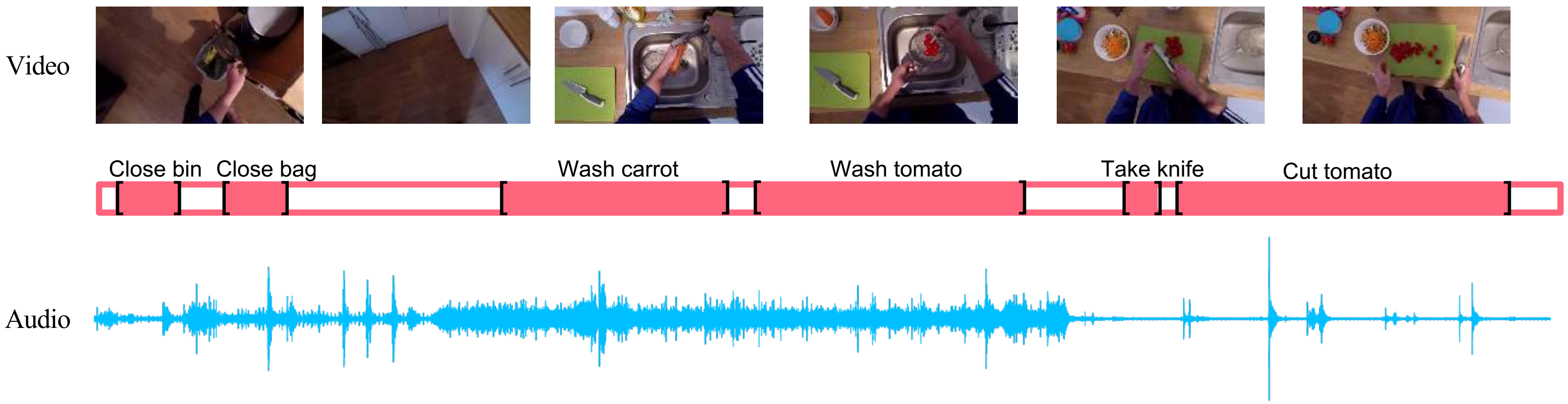
with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman



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

Motivation

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

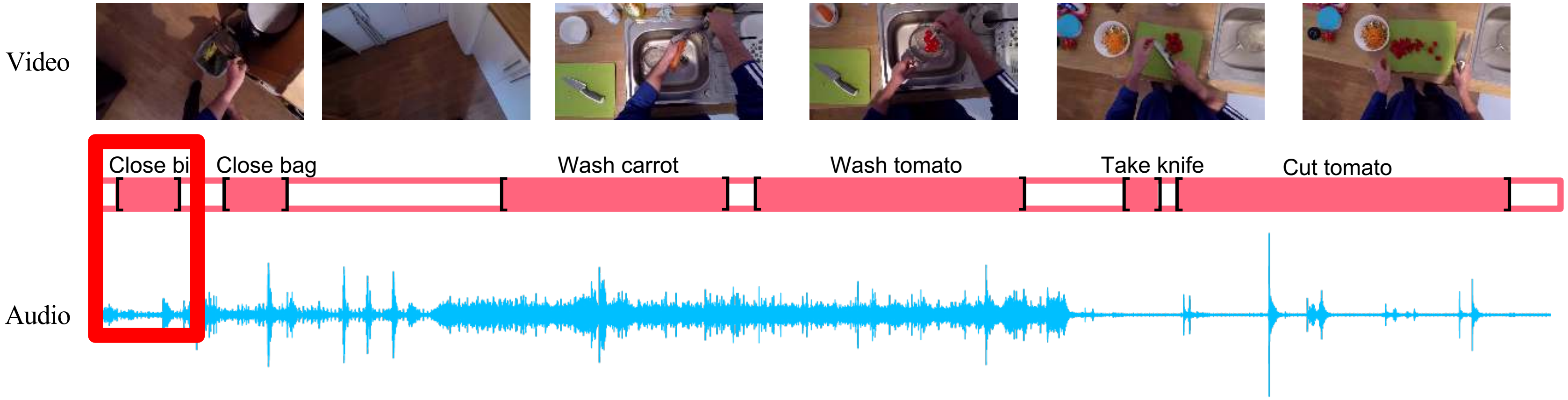


Motivation

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

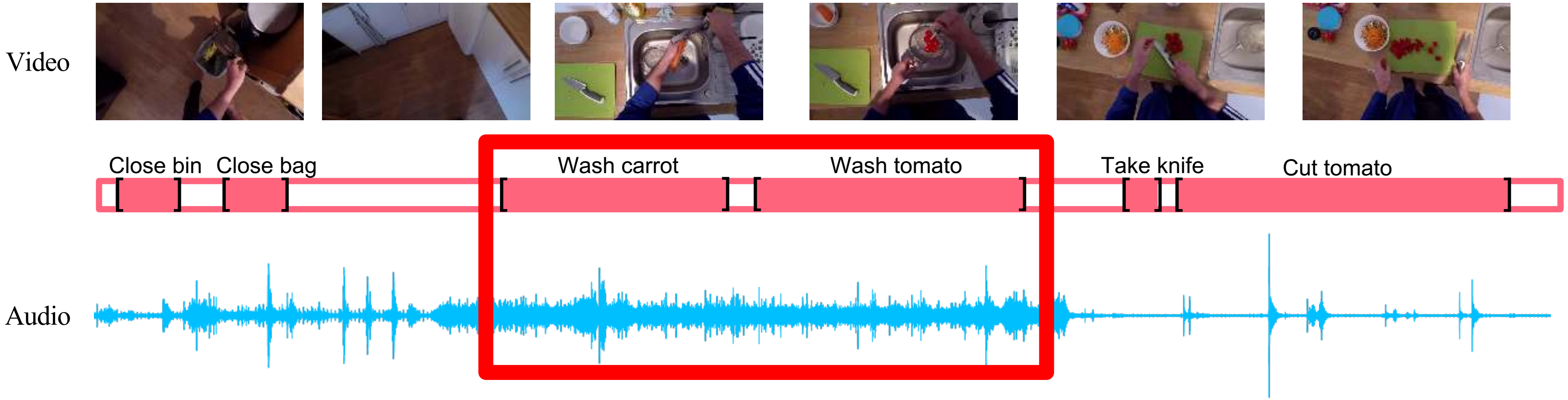
Motivation

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



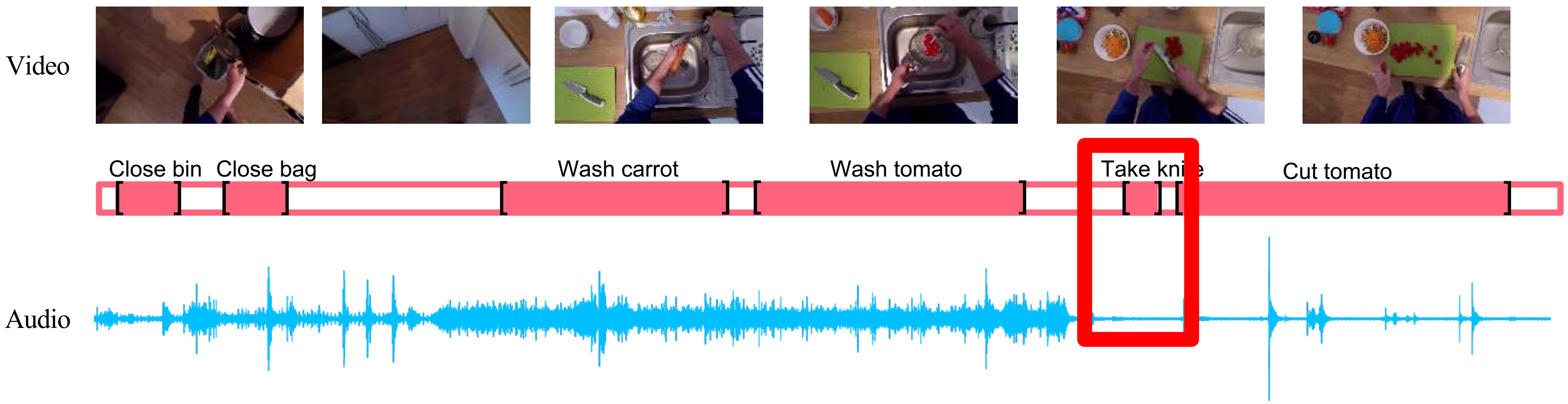
Motivation

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



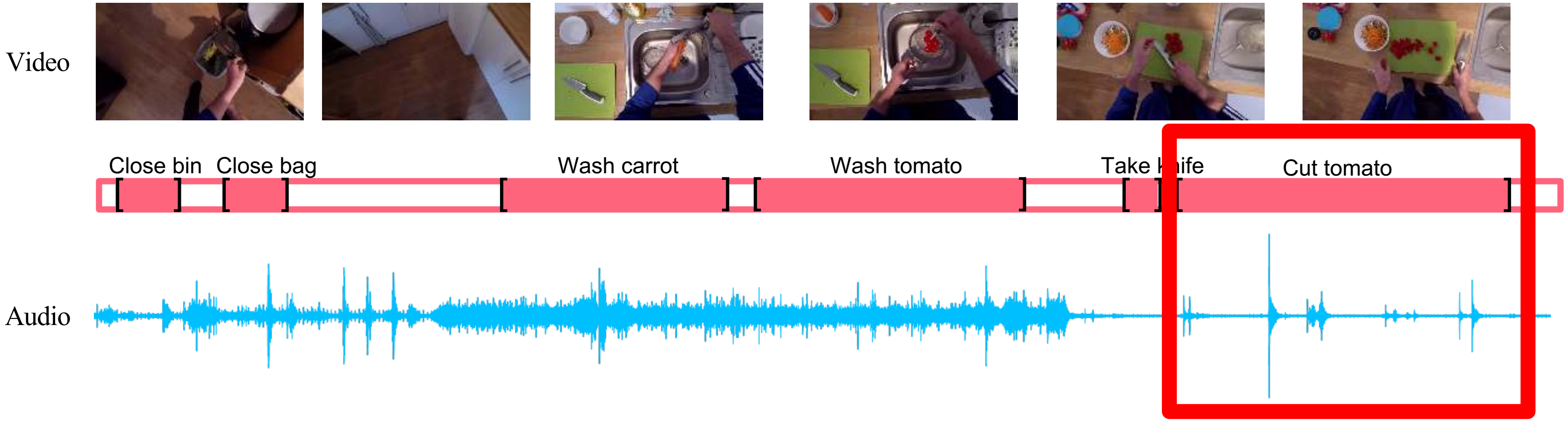
Motivation

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



Motivation

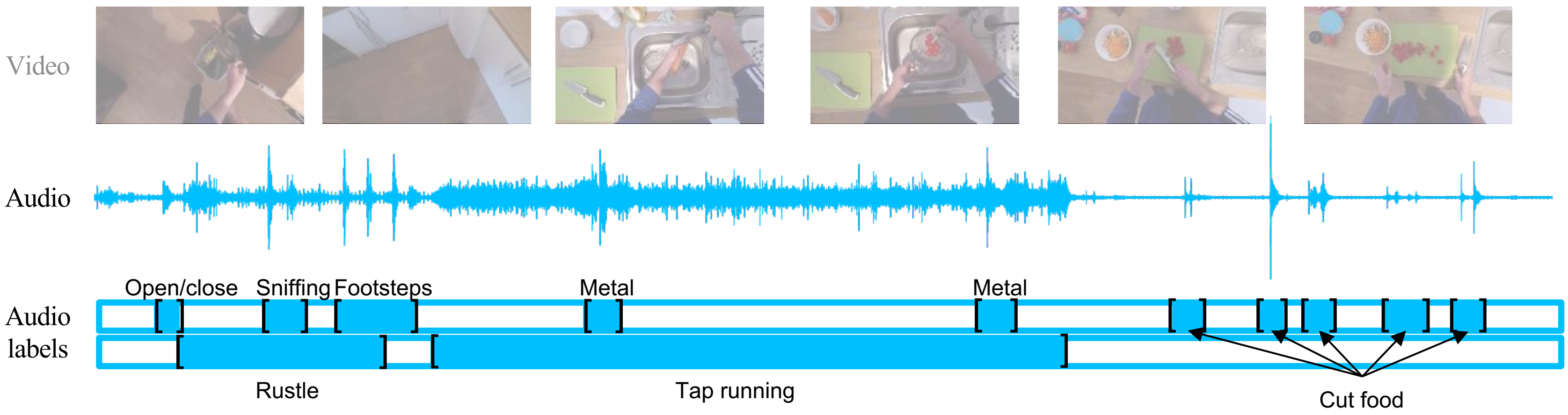
with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman



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

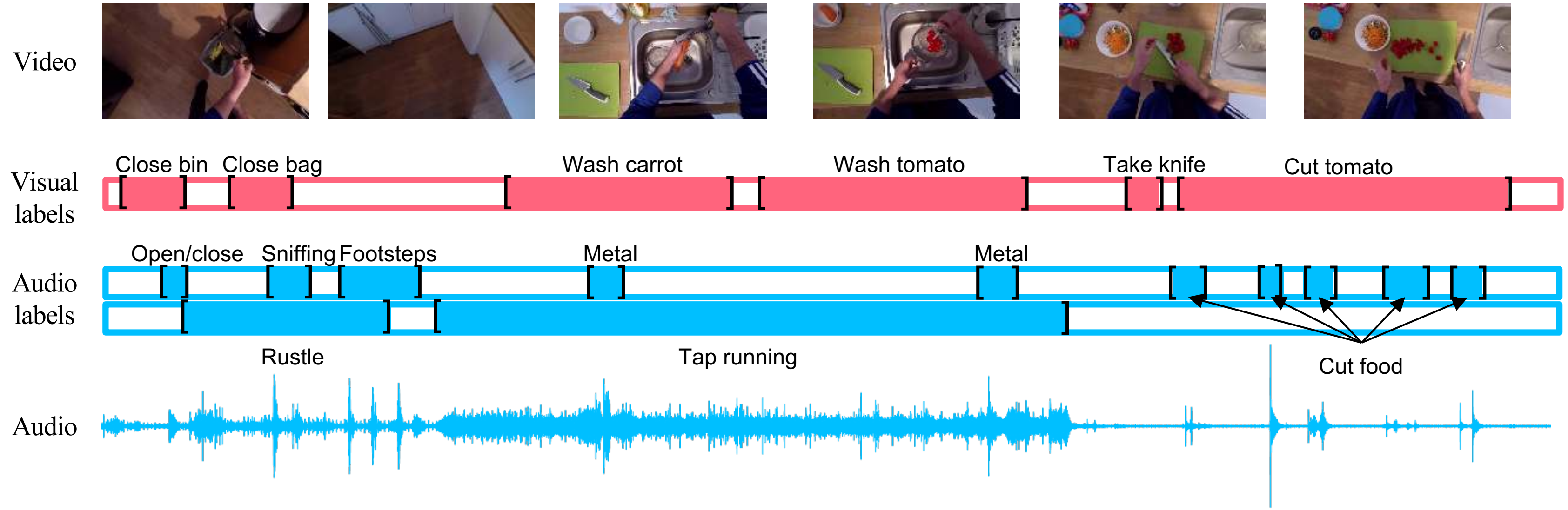
Motivation

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



Motivation

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



EPIC-SOUNDS

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

EPIC-KITCHENS VIDEOS
100 hours
45 kitchens

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

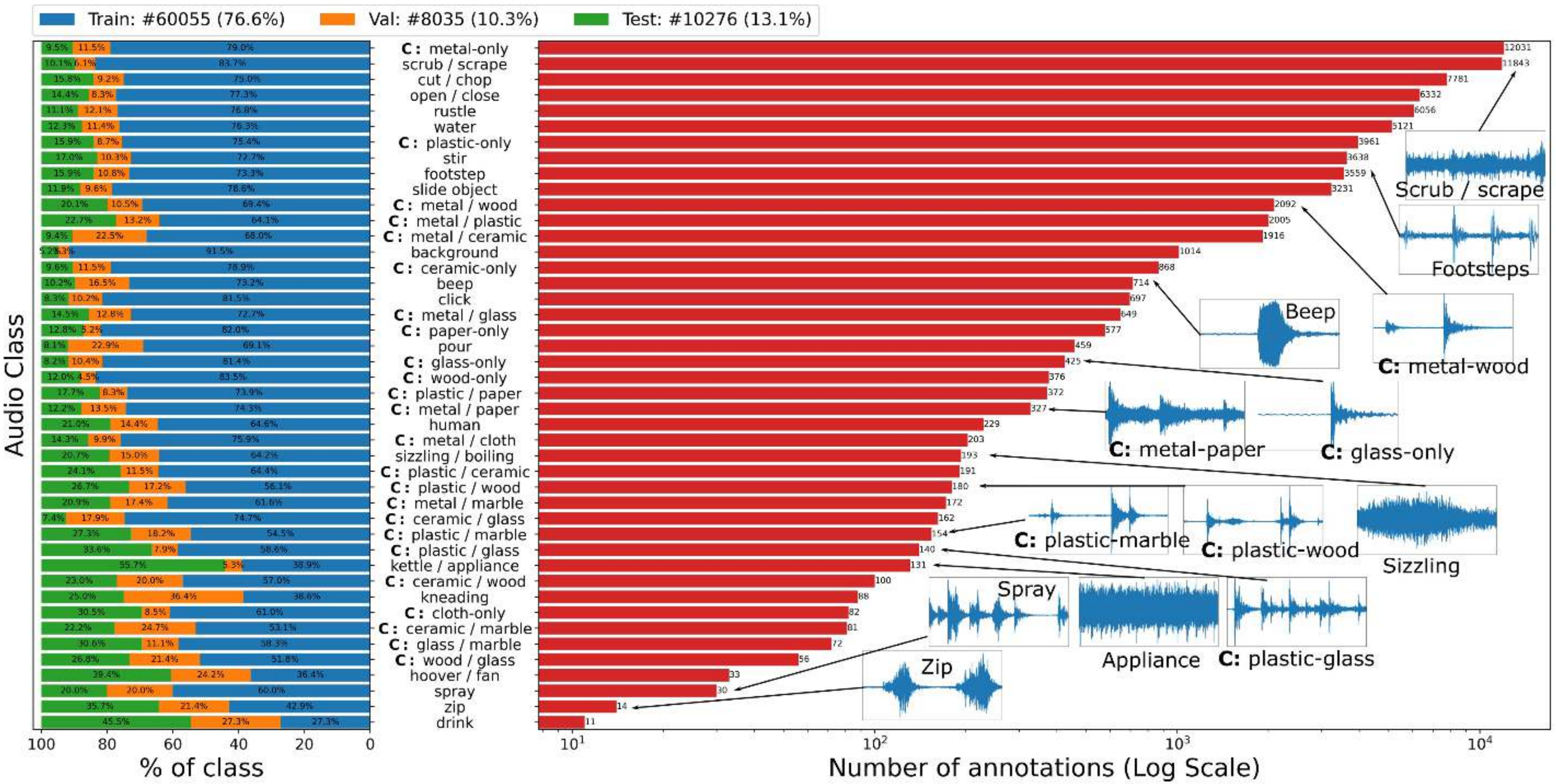
EPIC-Sounds
Audio-Based Annotations
79K categorised audio events
44 sound categories
39K uncategorised events

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


spray

EPIC-SOUNDS

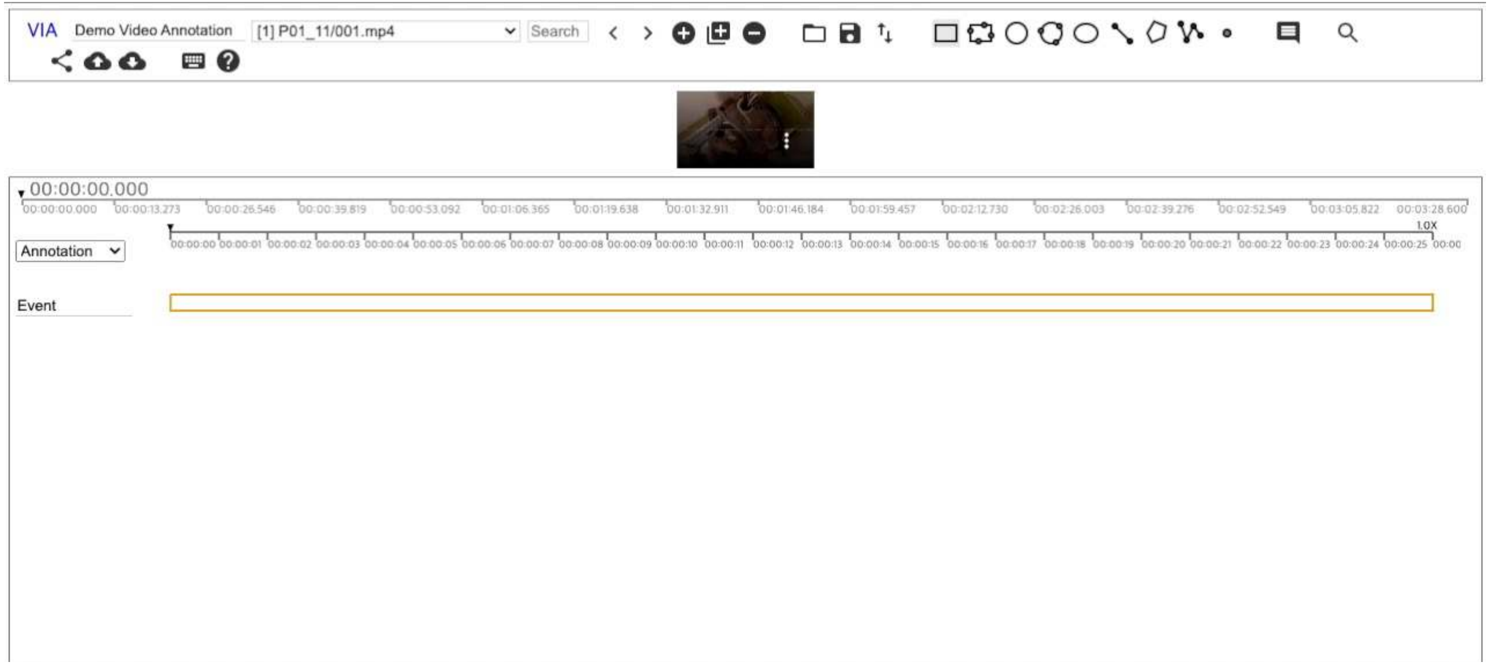
with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman



Annotations Pipeline

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

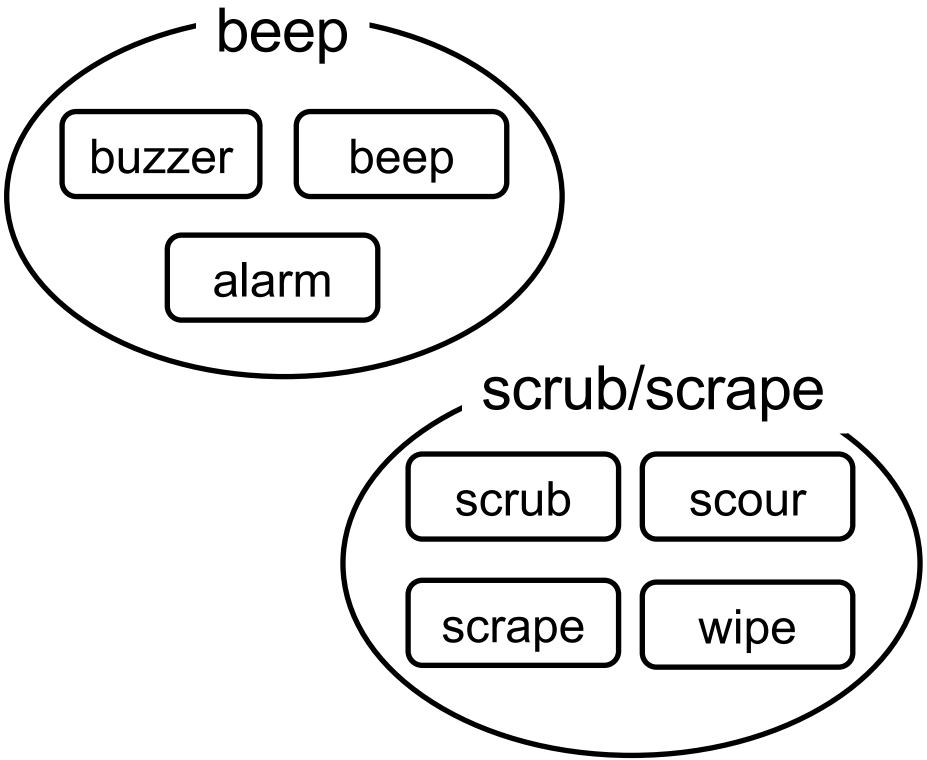
- We annotate all the distinctive sound events which consist of temporal intervals using free-form sound descriptions.
- Using VGG Image annotator tool



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

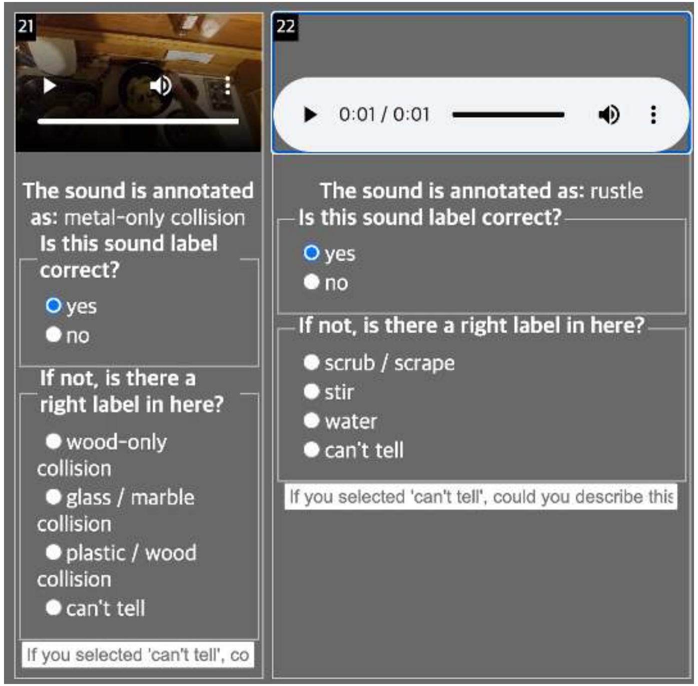

Plastic


Glass

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.

Temporal labels are modality-specific!

What type of labels can we provide?

- Temporal labels – Strong vs. Weak labels
- Semantic labels – Open-vocab. vs Closed-vocabulary
- Ranking labels – video-to-video comparisons
- Pixel-level labels – segmentation labels


Do
It
Yourself

Verb?

Noun?

sli.do

Joining as a participant?

#3639 120

Verbs:

add

pour

sprinkle

salt

season

Nouns:

salt

sea salt

seasoning

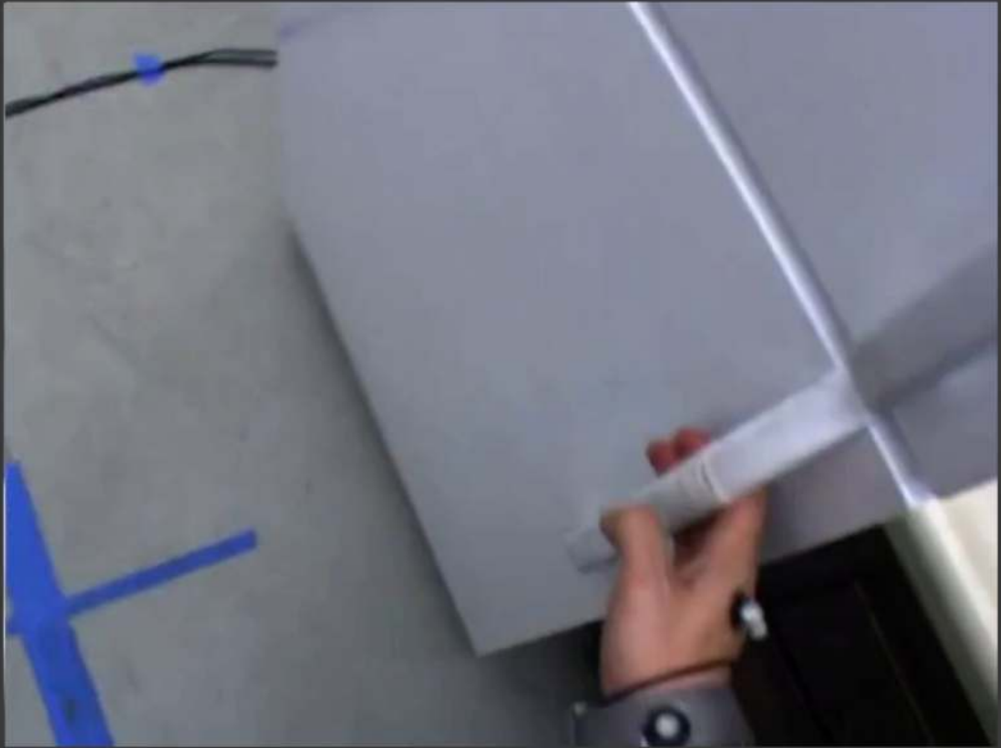
salt granules

sprinkle salt

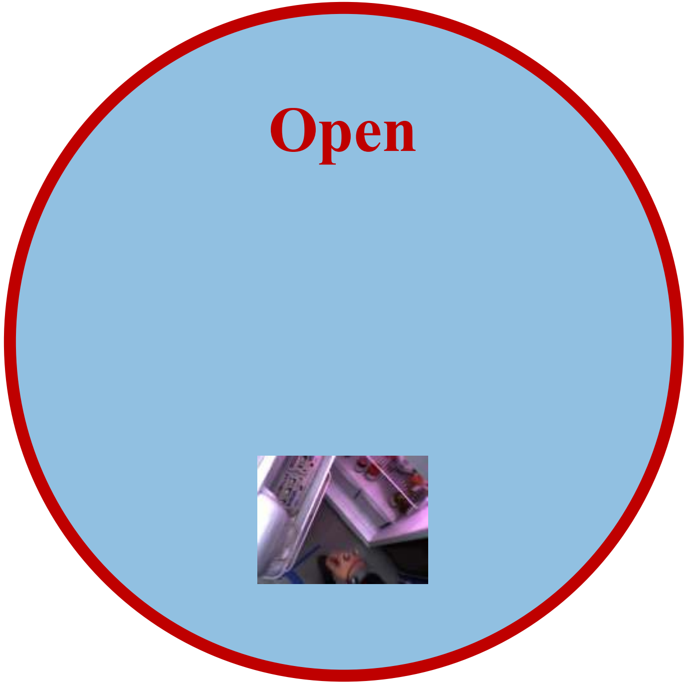
season meat

Think of an example of an *opening* action

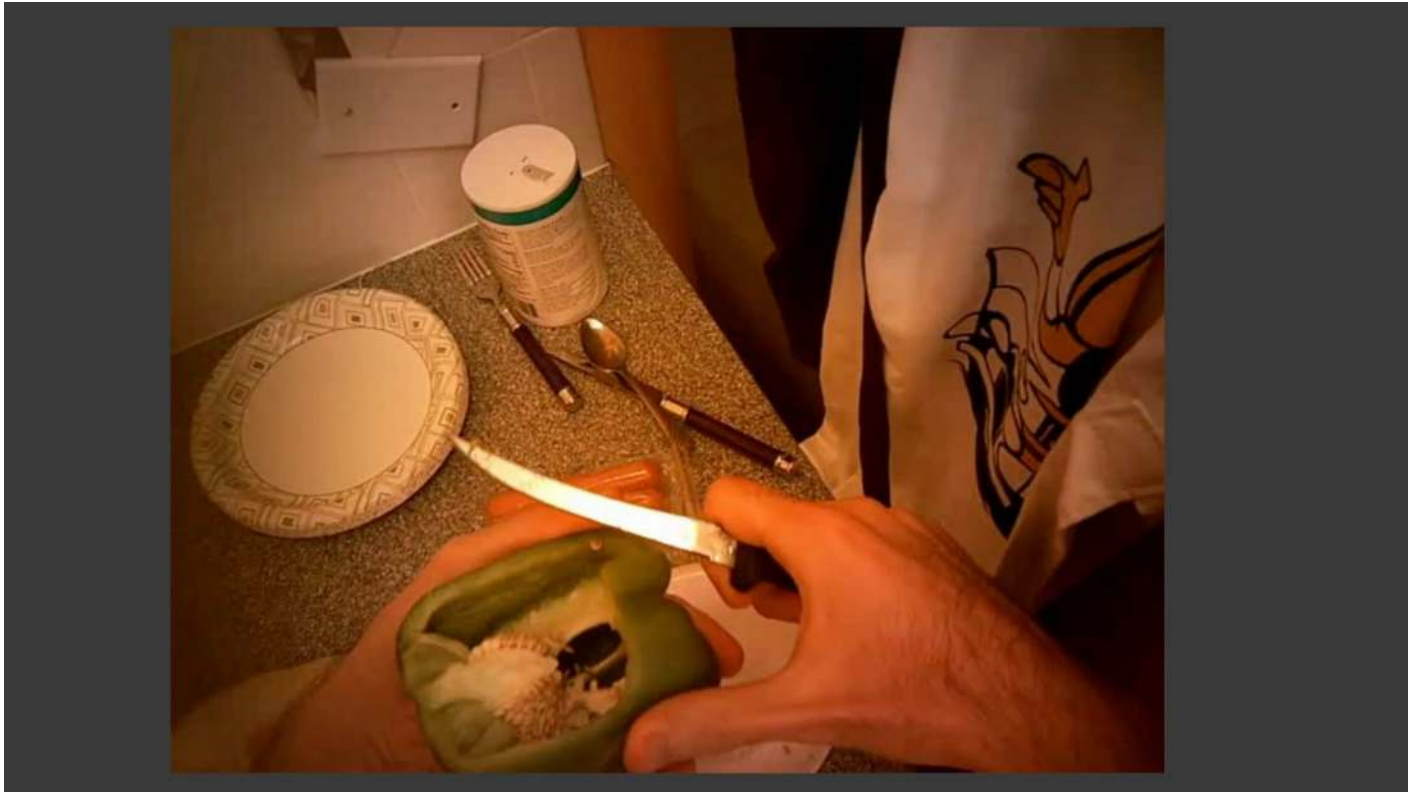
with: Michael Wray



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).

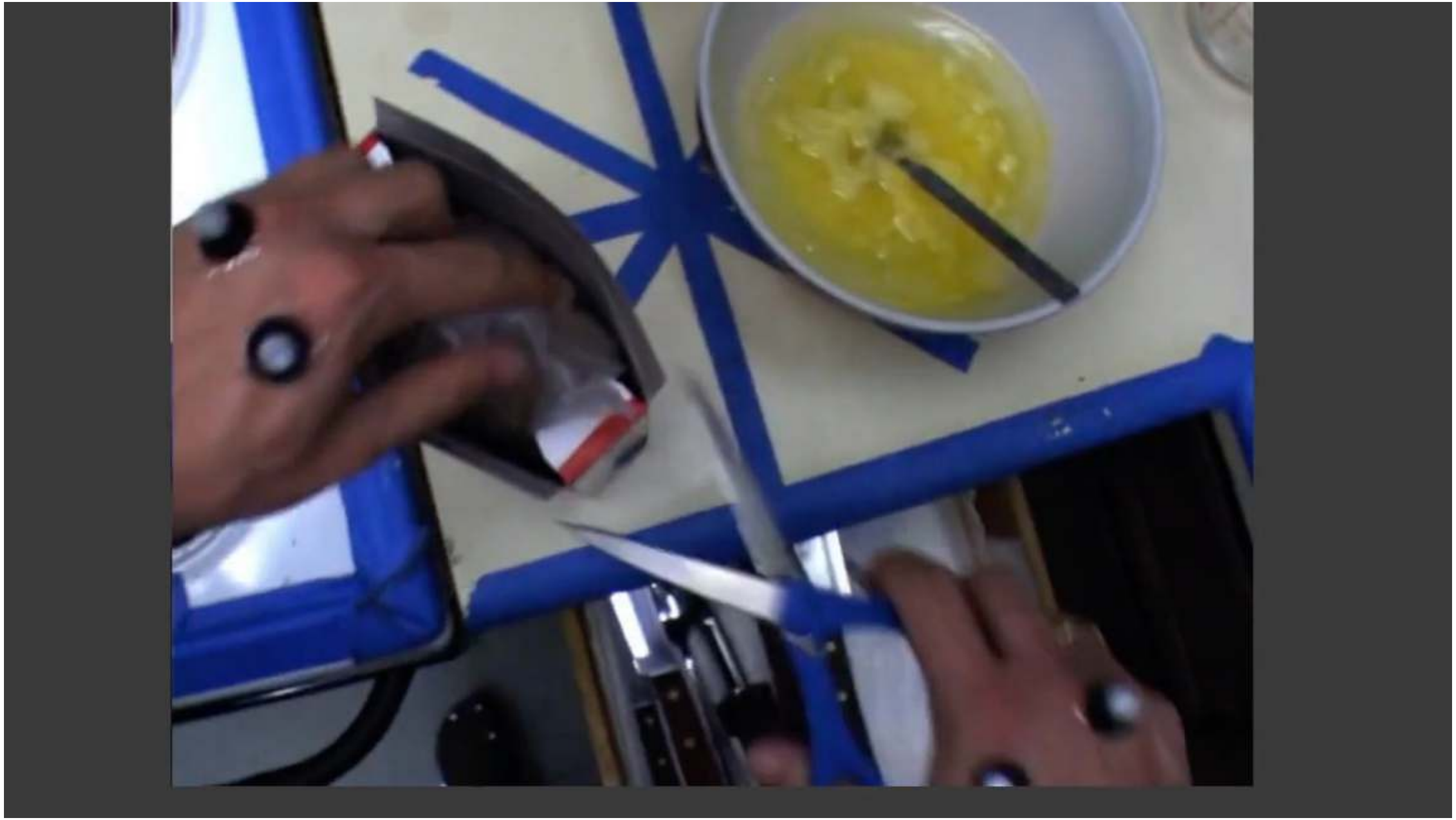
with: Michael Wray

Open

Cut

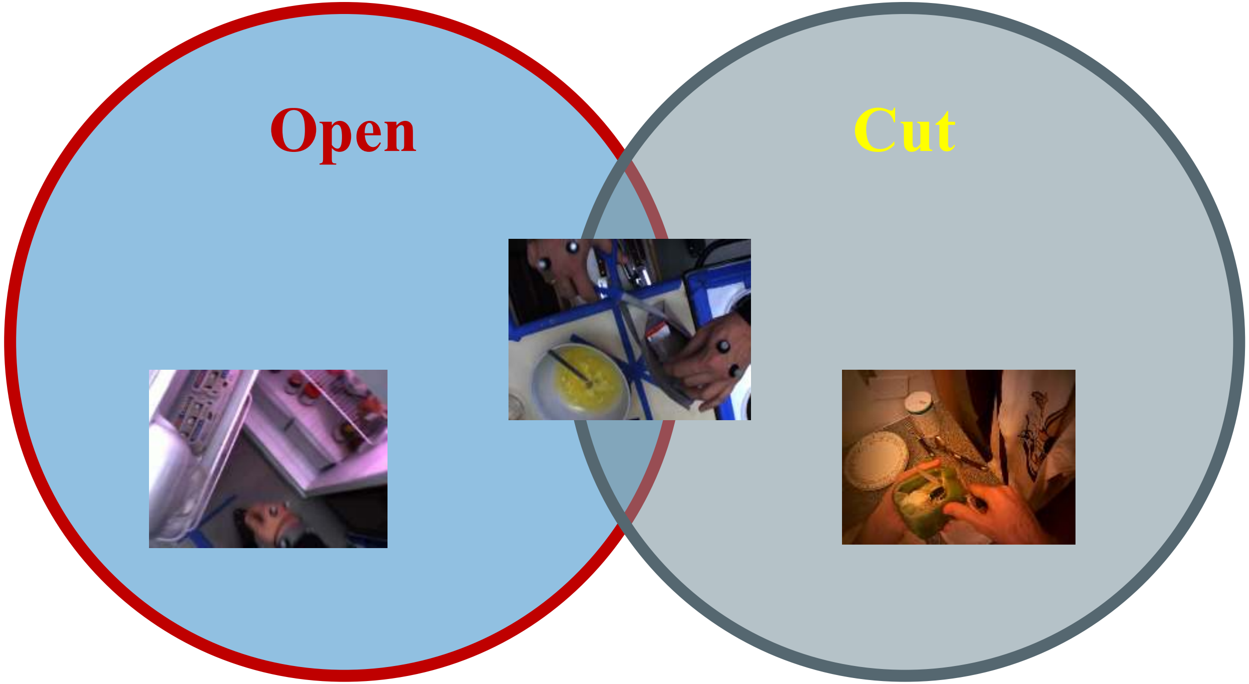
M Wray and D Damen (2019). Learning Visual Actions Using Multiple Verb-Only Labels. British Machine Vision Conference (BMVC).

with: Michael Wray



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

with: Michael Wray



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

Towards an unequivocal rep. of actions

with: Michael Wray

- Action representations using a single verb is highly-ambiguous
  - Solution1: pre-selected non-overlapping verbs (SL)
    - run, walk, open, close
  - Solution2: Using nouns to disambiguate actions (V-N)
    - open-drawer, open-bottle, open-fridge
    - actions constrained to known nouns
  - Solution3: Multi-verb labels (ML, SAML)
    - open, hold, pull

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

Towards an unequivocal rep. of actions

- Collected from AMT

SL

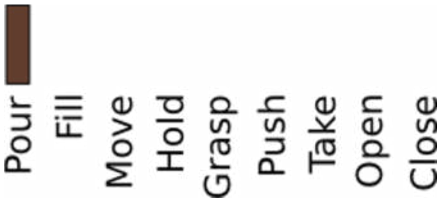
- Majority Vote.
- One-hot vector.

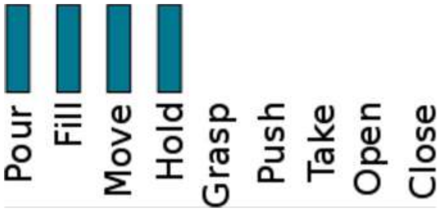
ML

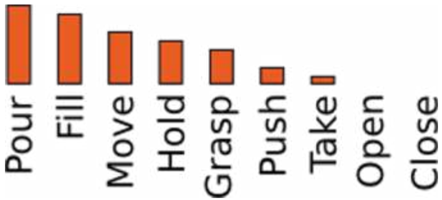
- Threshold of 0.5.
- Binary Vector

SAML

- Full Annotation.
- Continuous Vector.





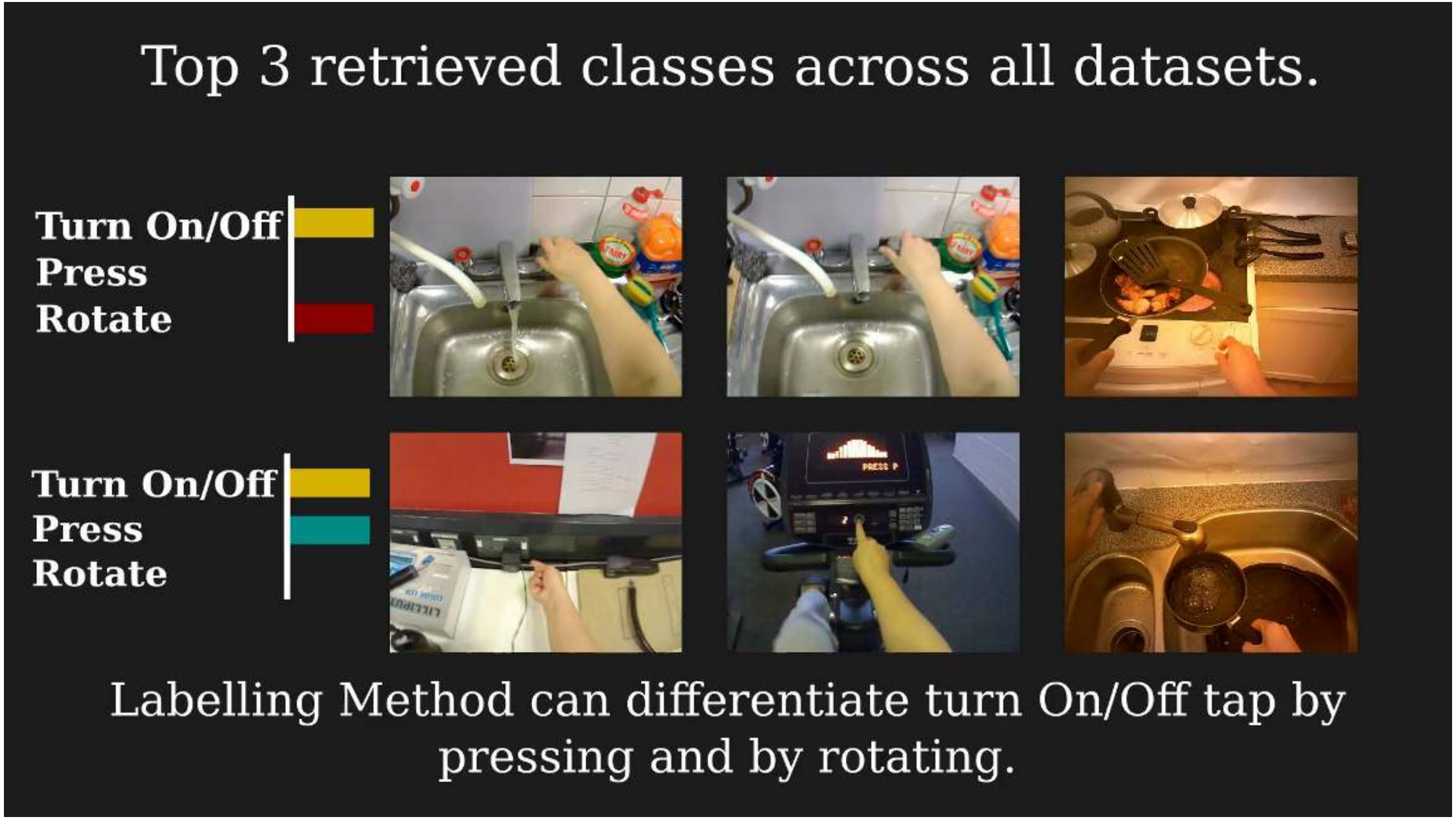


University of BRISTOL

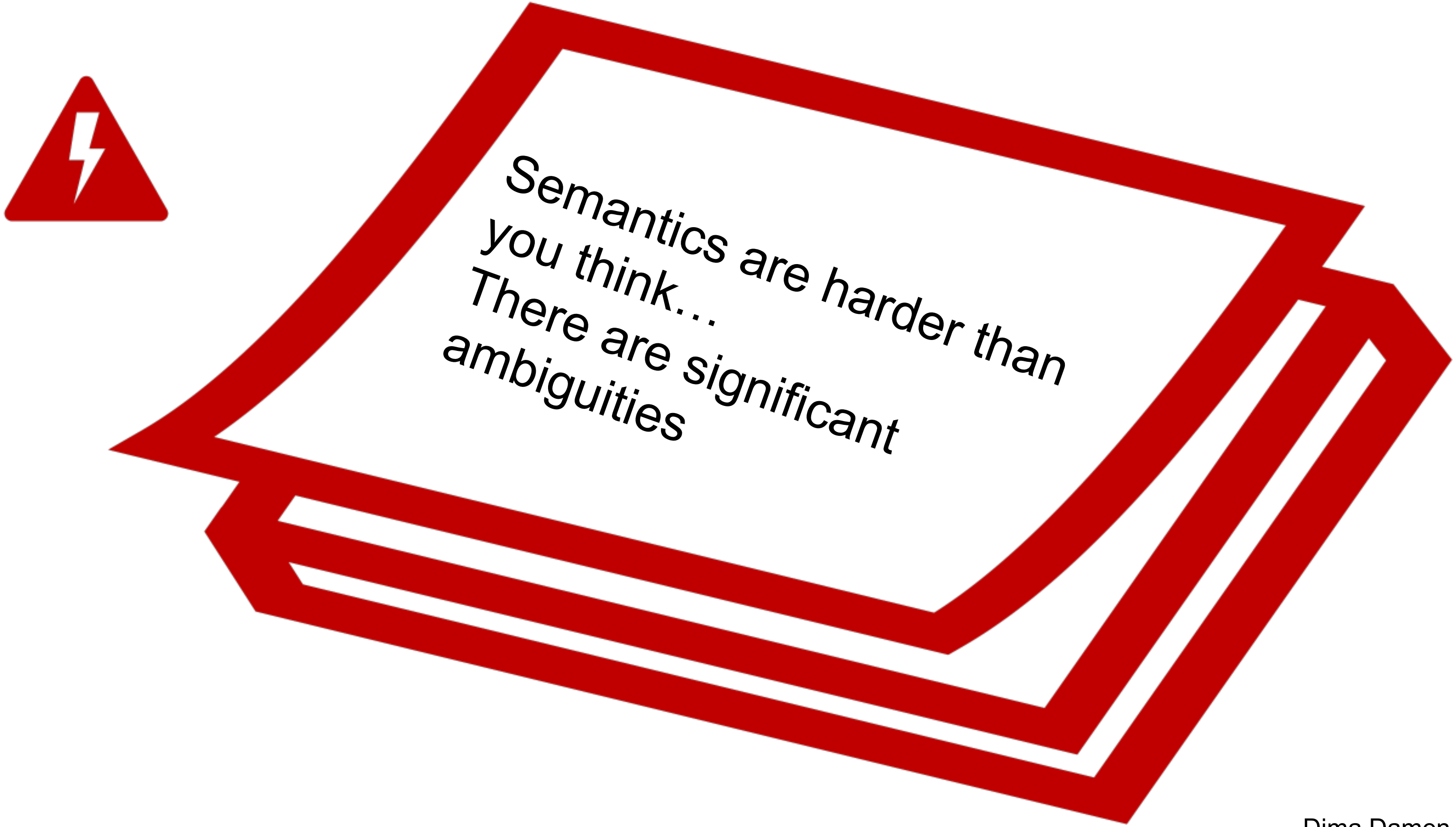
M Wray and D Damen (2019). Learning Visual Actions Using Multiple Verb-Only Labels. British Machine Vision Conference (BMVC).

Towards an unequivocal rep. of actions

with: Michael Wray



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

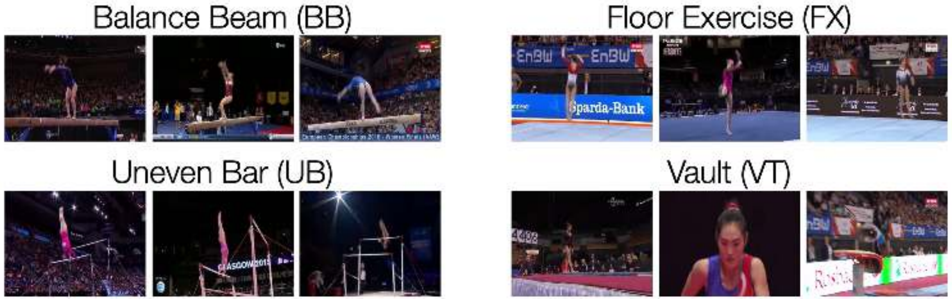

Semantics are harder than you think…
There are significant ambiguities

What type of labels can we provide?

- Temporal labels – Strong vs. Weak labels
- Semantic labels – Open-vocab. vs Closed-vocabulary
- Ranking labels – video-to-video comparisons
- Pixel-level labels – segmentation labels

Quality of Actions…

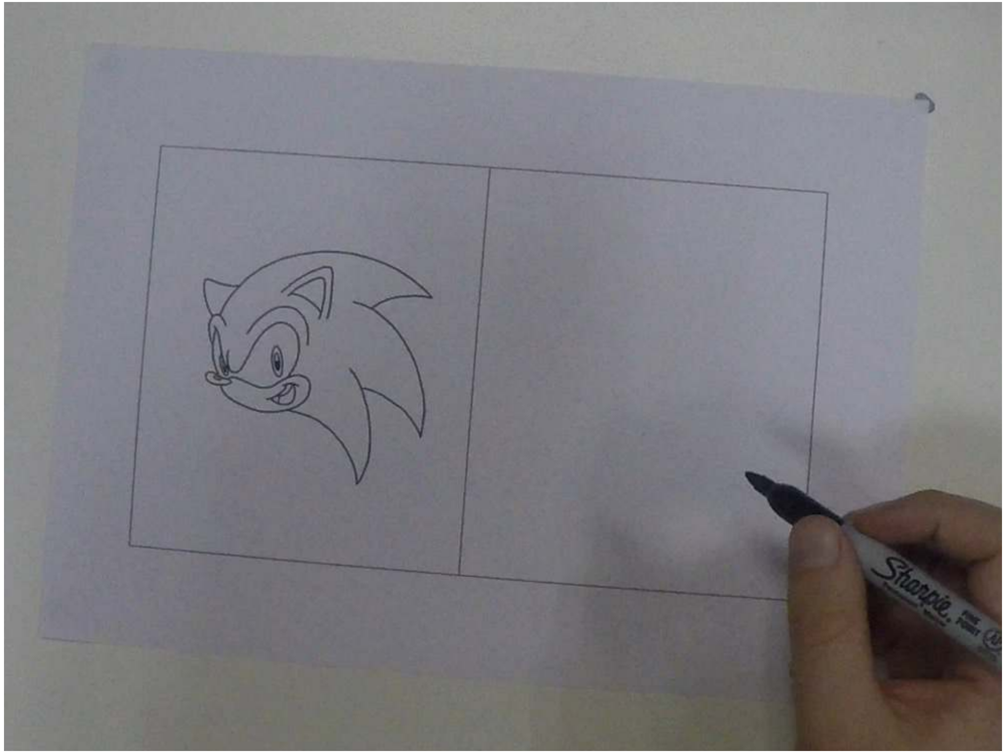
Pirsiavash et al, ECCV 2014



Shao et al, CVPR 2020

Quality of Actions…

with: Hazel Doughty
Walterio Mayol-Cuevas



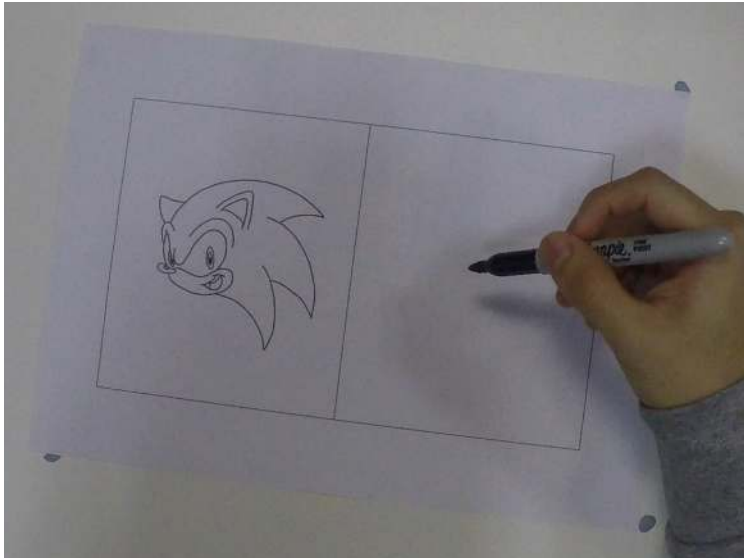
University of BRISTOL

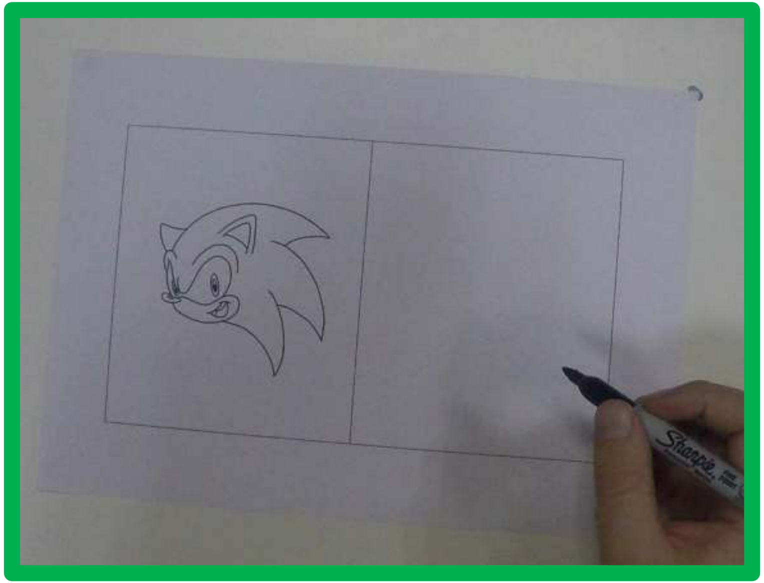
H Doughty, W Mayol-Cuevas, D Damen (2019). The Pros and Cons: Rank-aware Temporal Attention for Skill Determination in Long Videos. *Computer Vision and Pattern Recognition (CVPR)*

Skill determination in video

with: Hazel Doughty
Walterio Mayol-Cuevas

Pairwise annotations of videos, indicating higher skill or no skill preference





H Doughty, W Mayol-Cuevas, D Damen (2019). The Pros and Cons: Rank-aware Temporal Attention for Skill Determination in Long Videos. *Computer Vision and Pattern Recognition (CVPR)*

What type of labels can we provide?

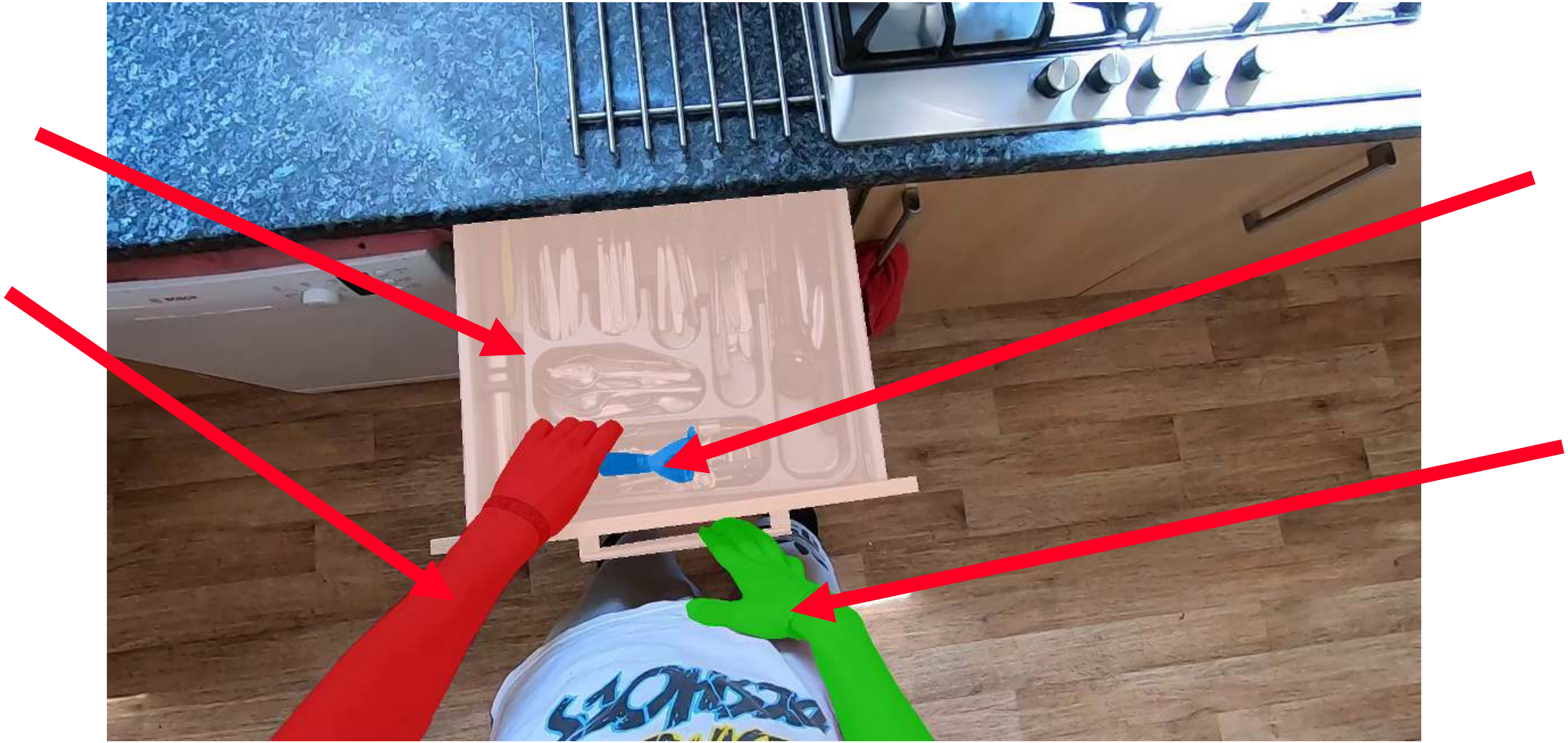
- Temporal labels – Strong vs. Weak labels
- Semantic labels – Open-vocab. vs Closed-vocabulary
- Ranking labels – video-to-video comparisons
- Pixel-level labels – segmentation labels

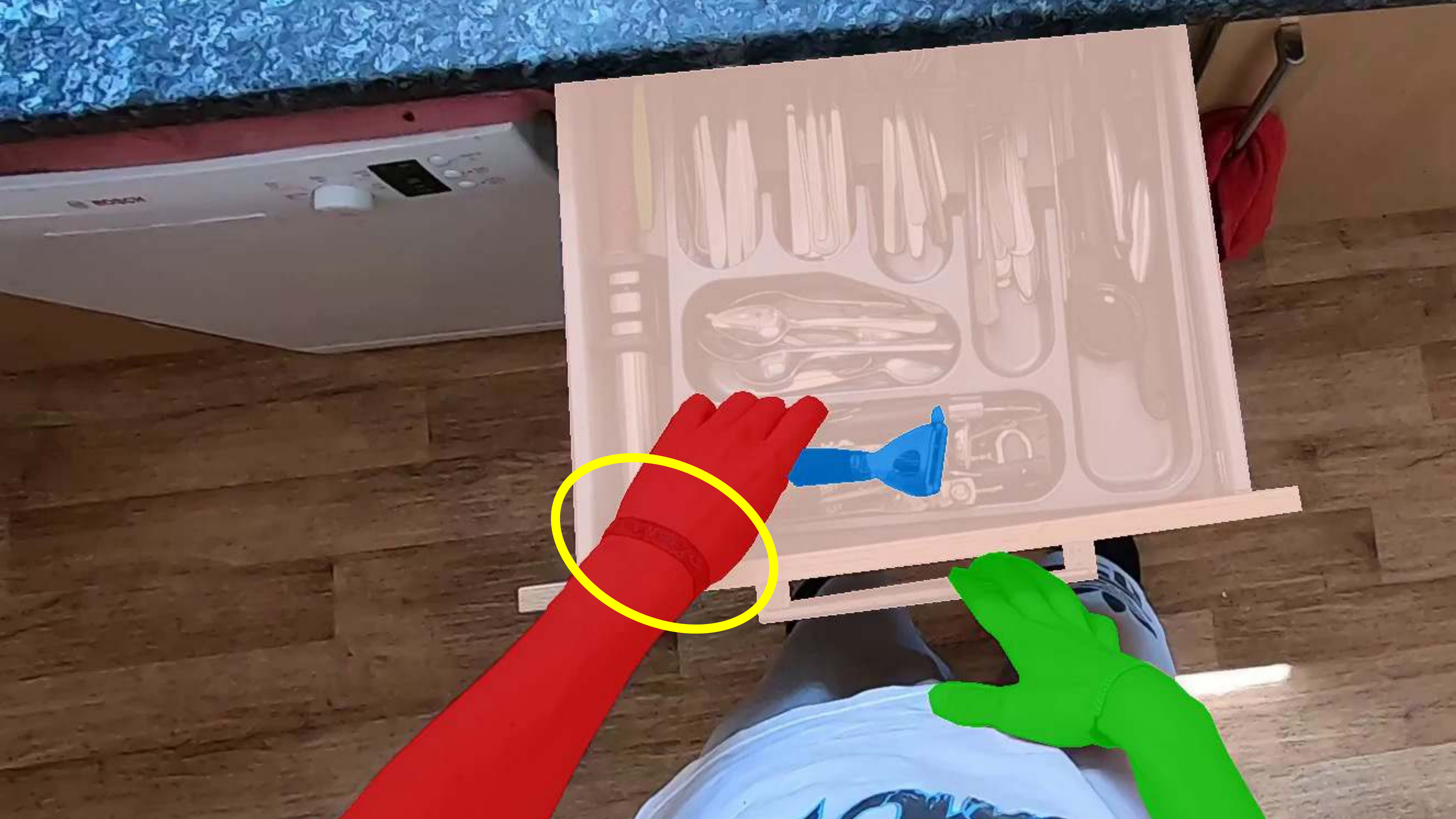
EPIC-KITCHENS VISOR

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

EPIC-KITCHENS VISOR

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

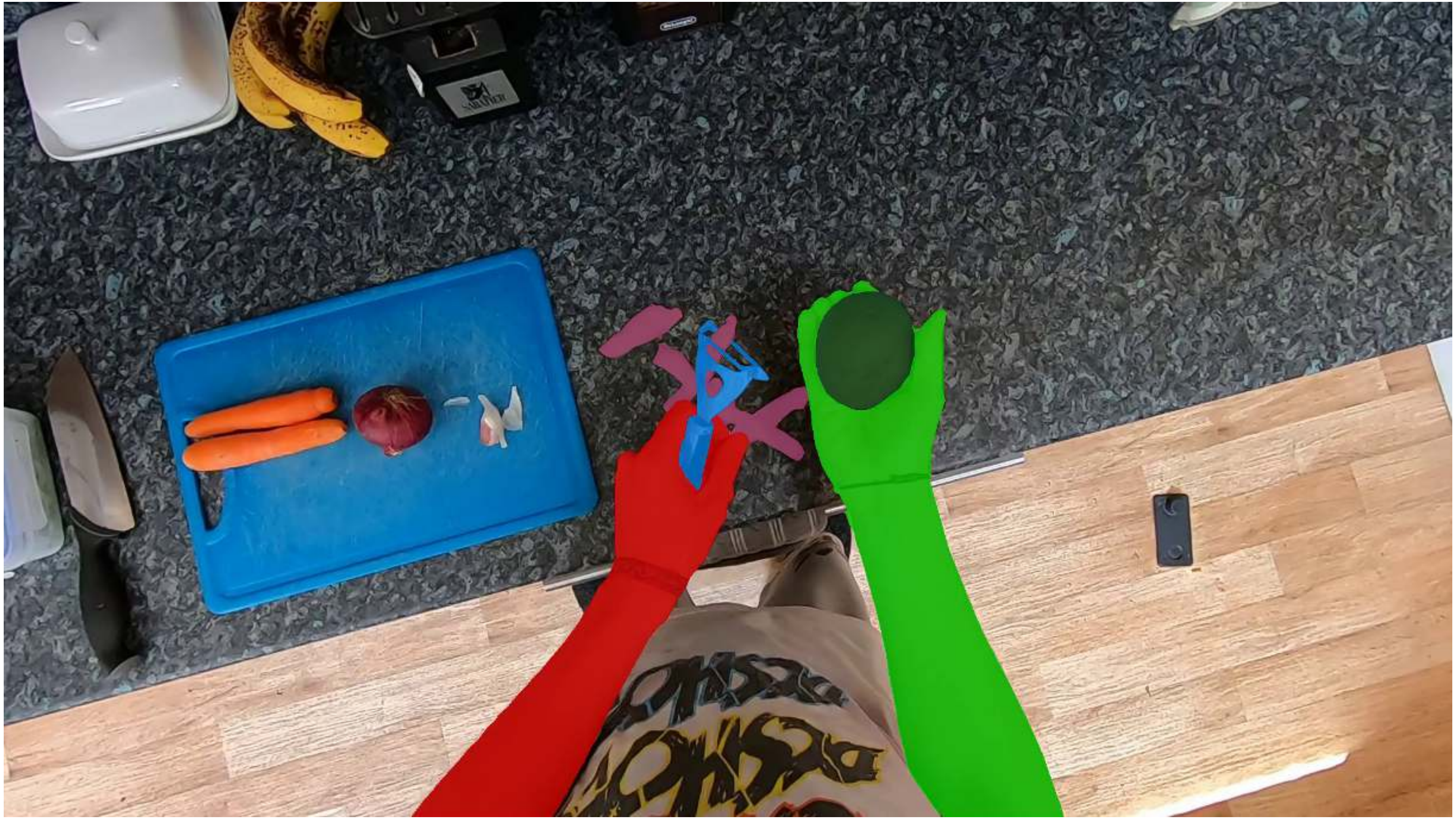


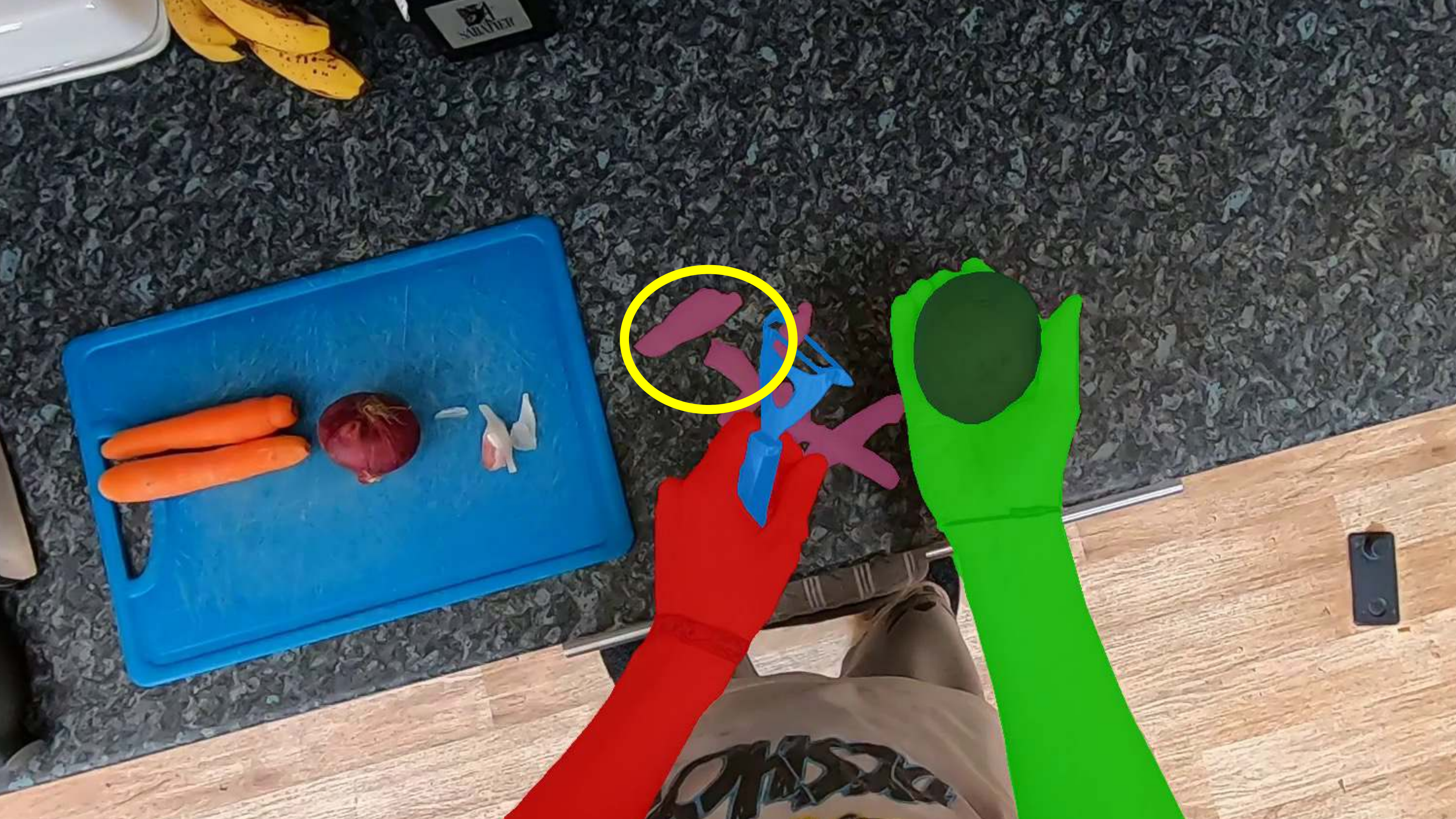



Depth 0 : peeler
Depth -1 : drawer

EPIC-KITCHENS VISOR

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





EPIC-KITCHENS VISOR

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

EPIC-KITCHENS VISOR

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

EPIC-KITCHENS VISOR

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

EPIC-KITCHENS VISOR

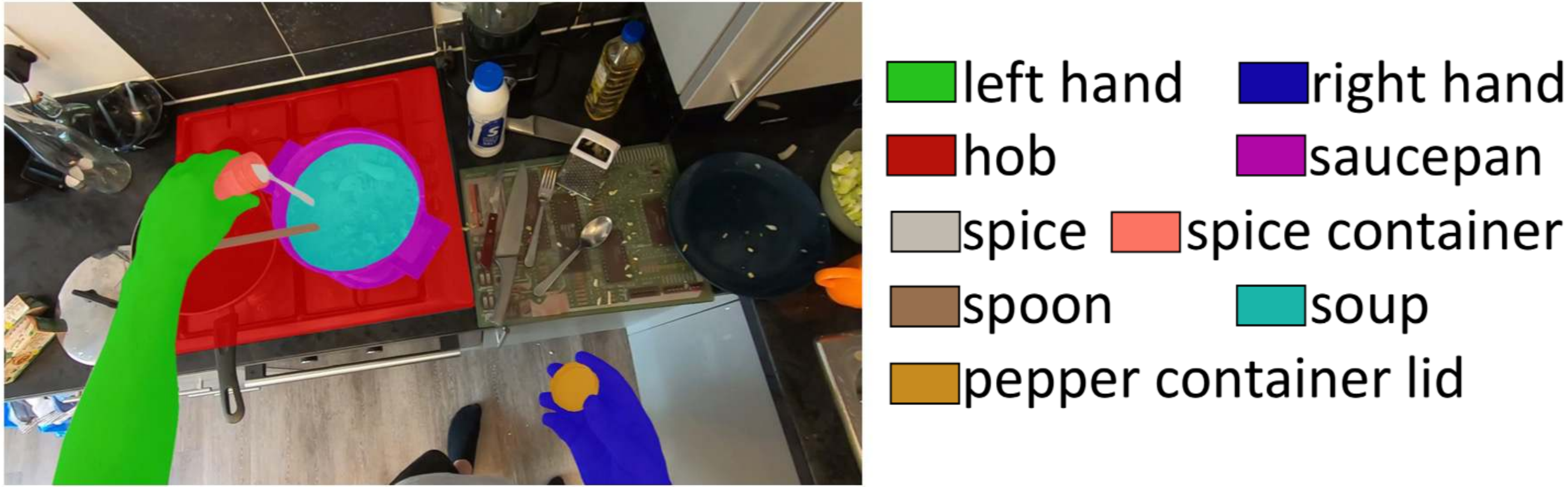
with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma,
Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler

EPIC-KITCHENS VISOR

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

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

pour spice

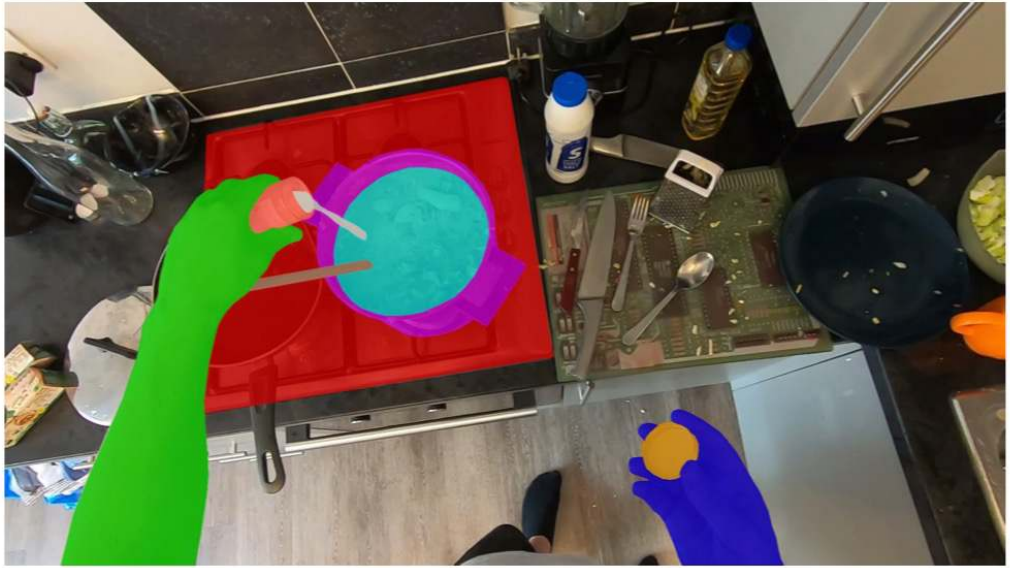


with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler, Dima Damen

pour spice

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

pour spice



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

saucepan → pan → cookware

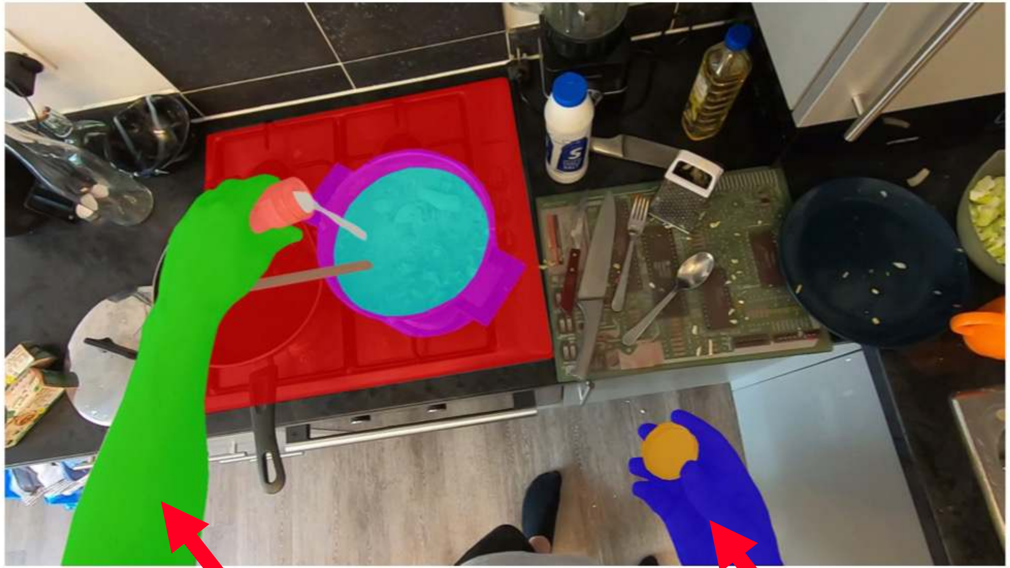
spoon → spoon → cutlery

VISOR is….

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

action

pour spice



in-contact (spice container)

in-contact (container lid)

University of BRISTOL

VISOR is….

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

pour spice

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

spoon (non-exhaustive)

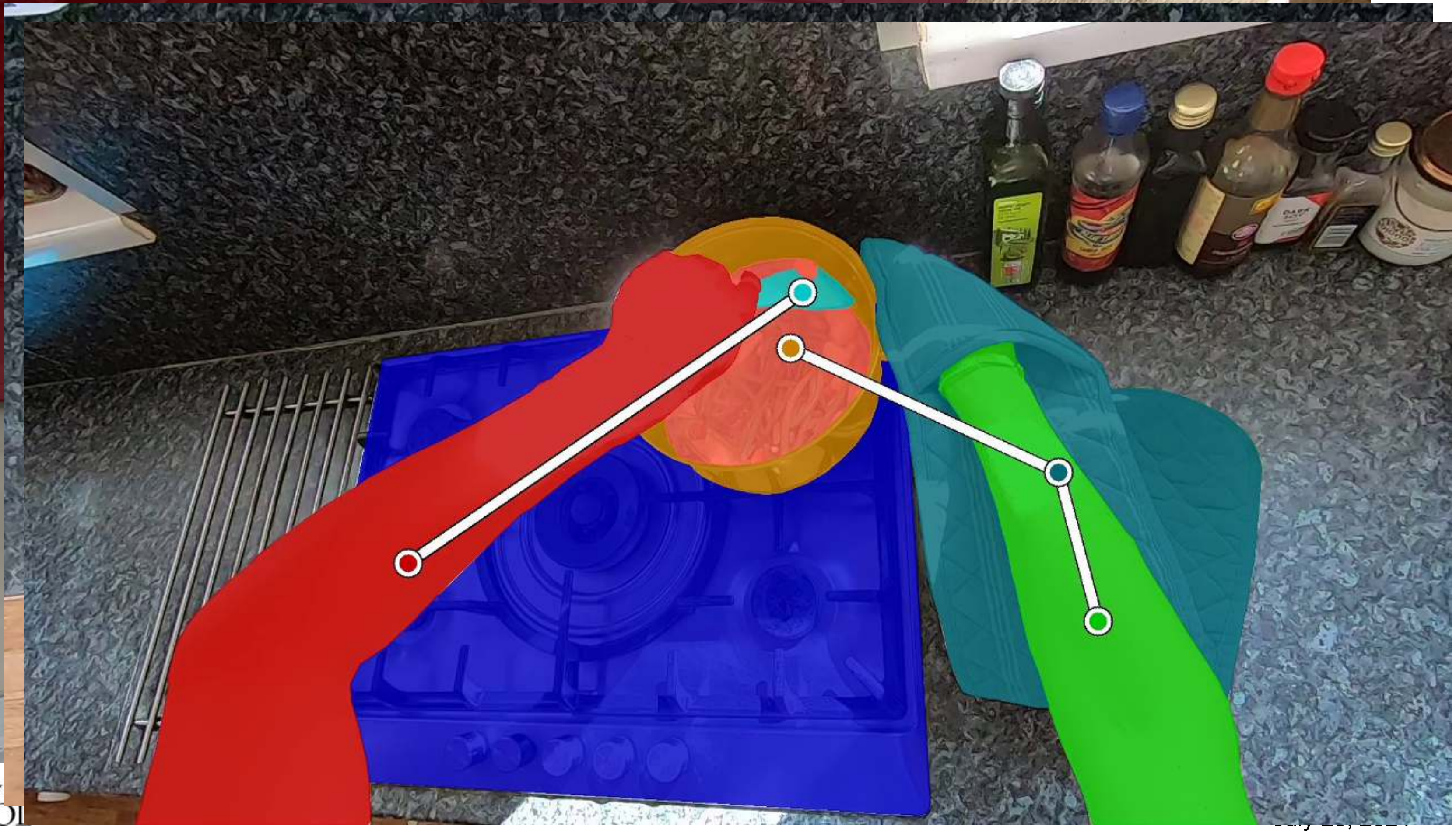
EPIC-KITCHENS VISOR

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

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

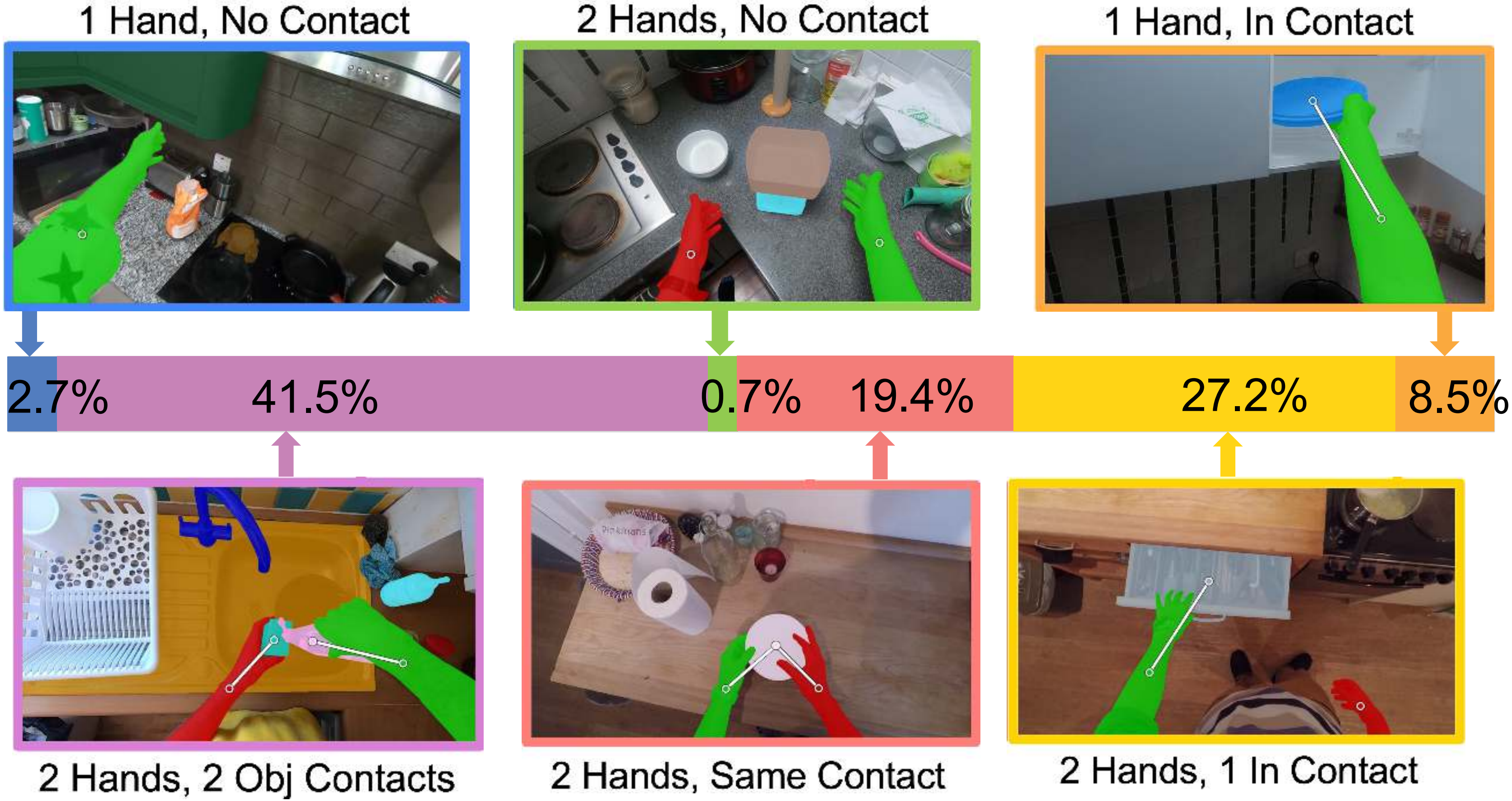
VISOR Relations

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



Object relation stats

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



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

EPIC-KITCHENS VISOR

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

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

Semantic segmentations during transformations is… very challenging

Data
bi
e always
researchers’

There is no Ground Truth – only biased labels

Dataset

Labels

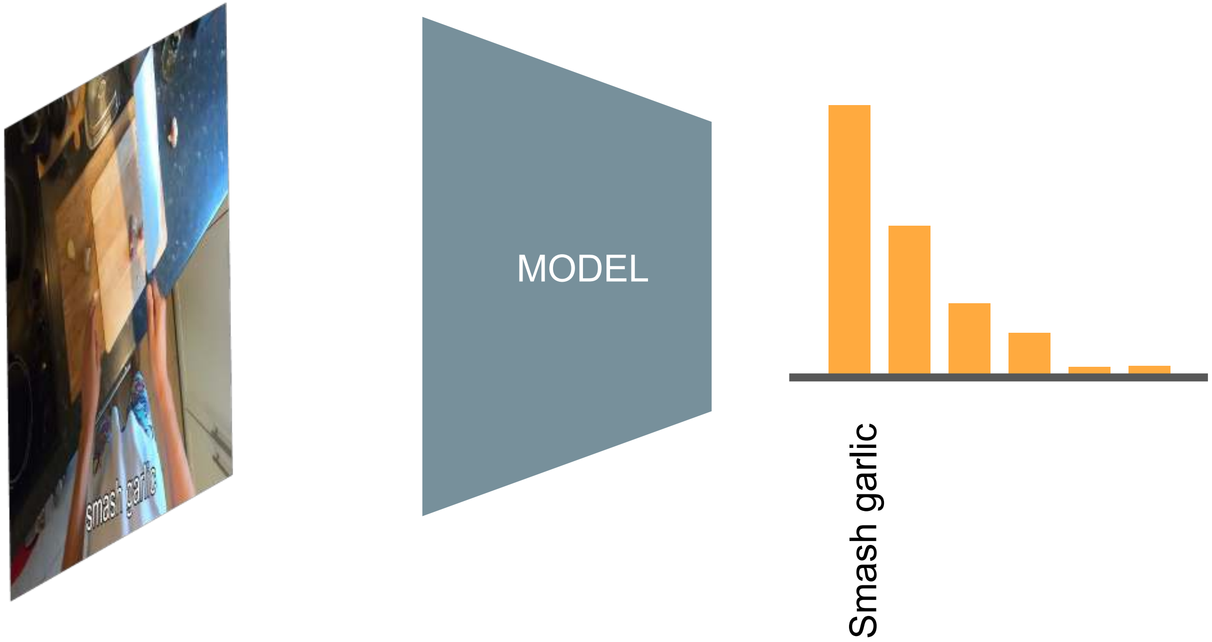
Tasks and Methods

Part III: Tasks and Methods

Sampling frames…

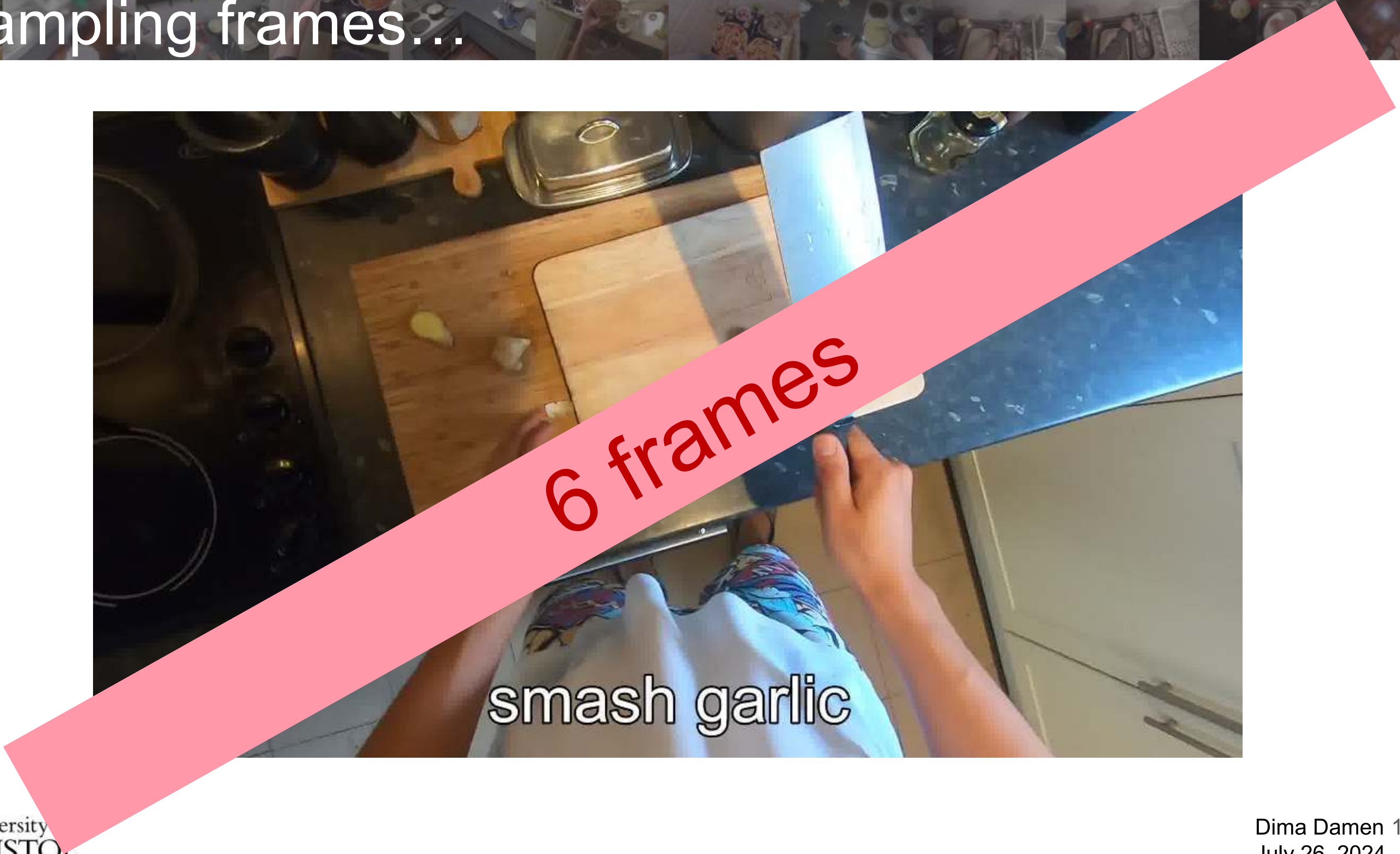
Sampling frames…

Sampling frames…



Sampling frames…

Sampling frames…

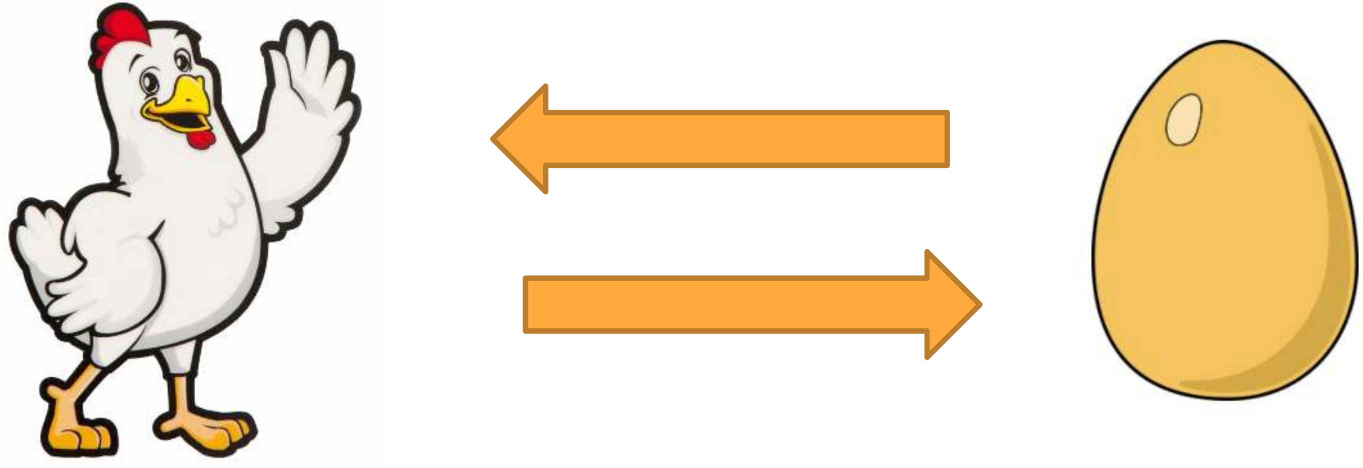


Sampling frames…

3 frames!!

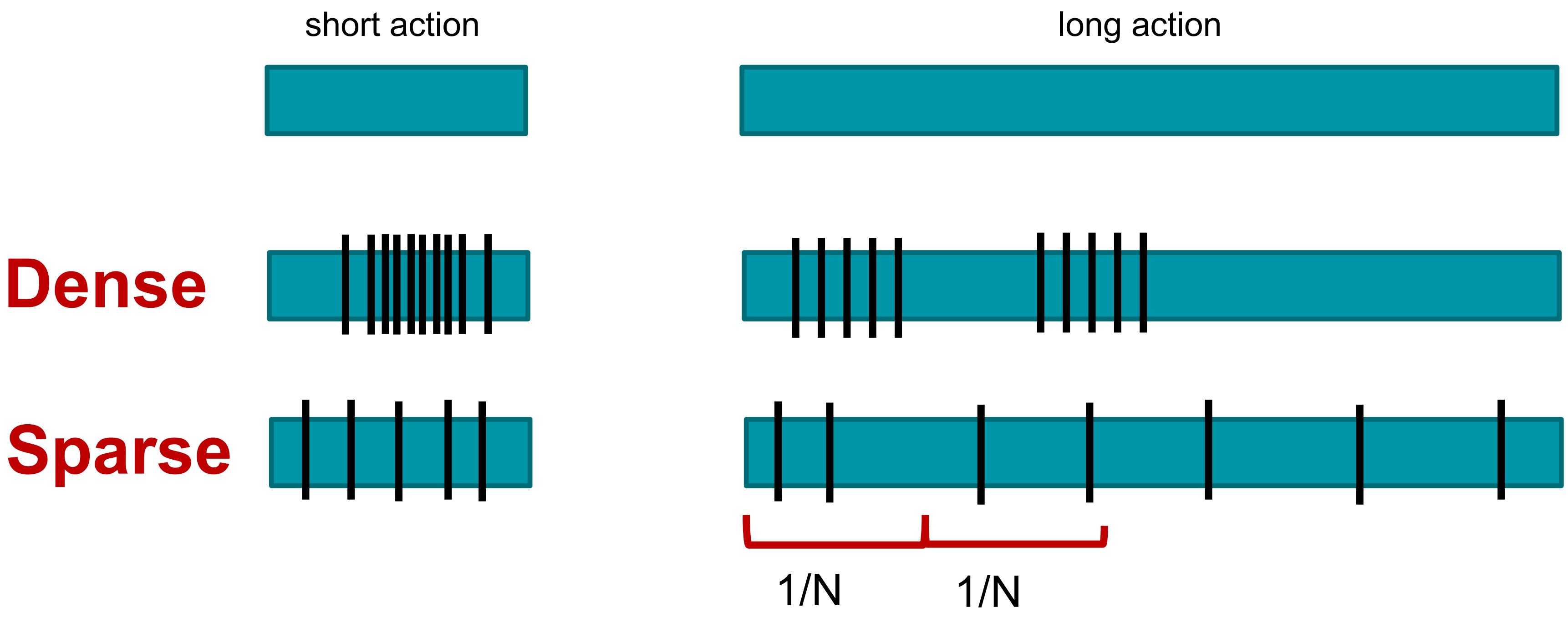
Sampling frames…

Frames to select | Action to recognise



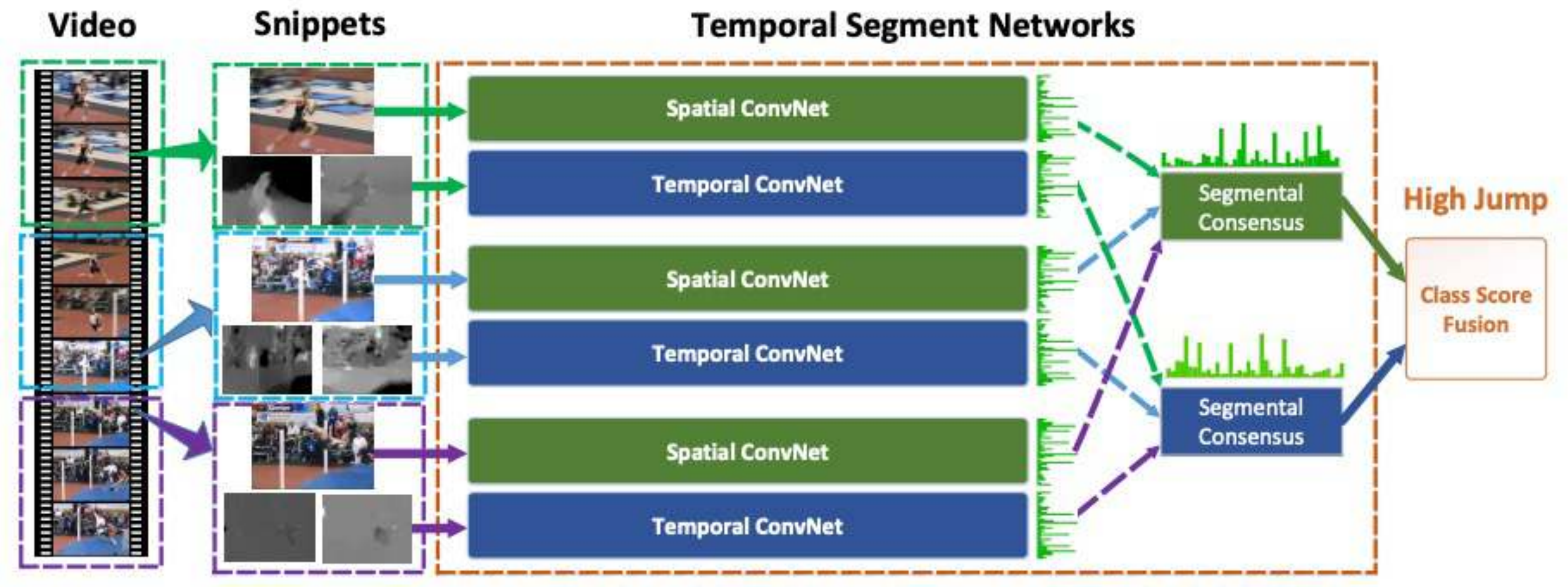
All models and methods sample frames…
Sampling is often hidden in implementation details…
It is ***critical*** …

Two sampling approaches



Sparse sampling

Temporal Segment Networks (TSN) – Wang et al, ECCV 2016



Two sampling approaches

Dense

Better motion features

Short motion signature

Easier to implement

Better for cross-dataset generalisation

Sparse

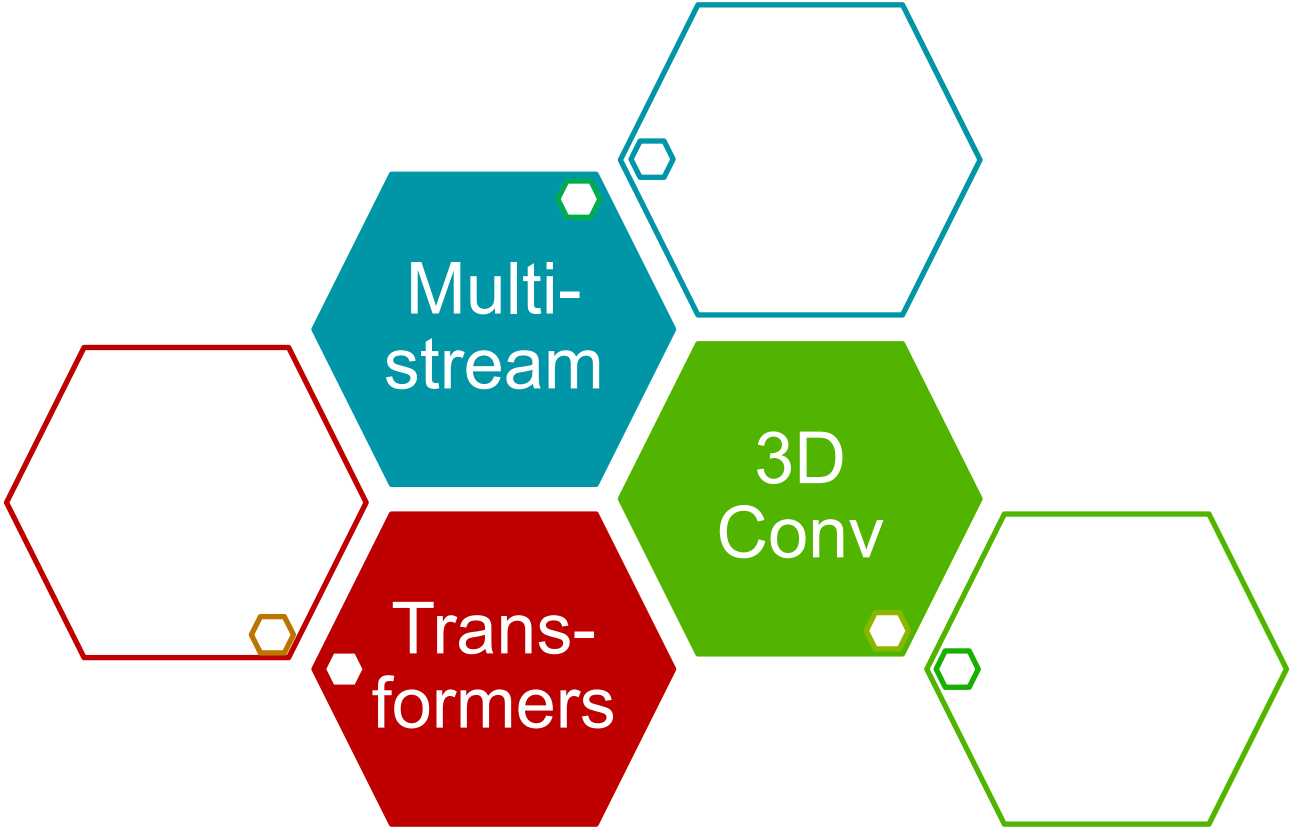
More complex features

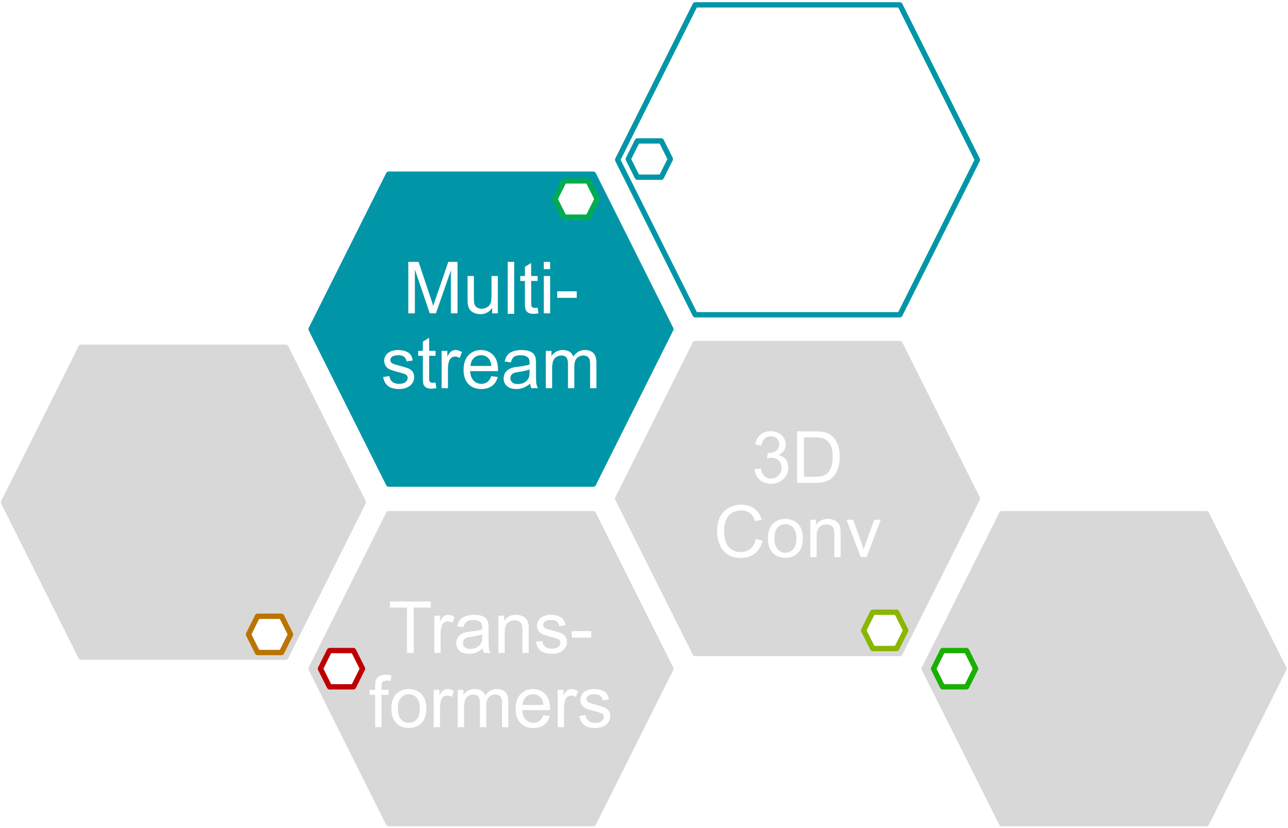
Complete action representation

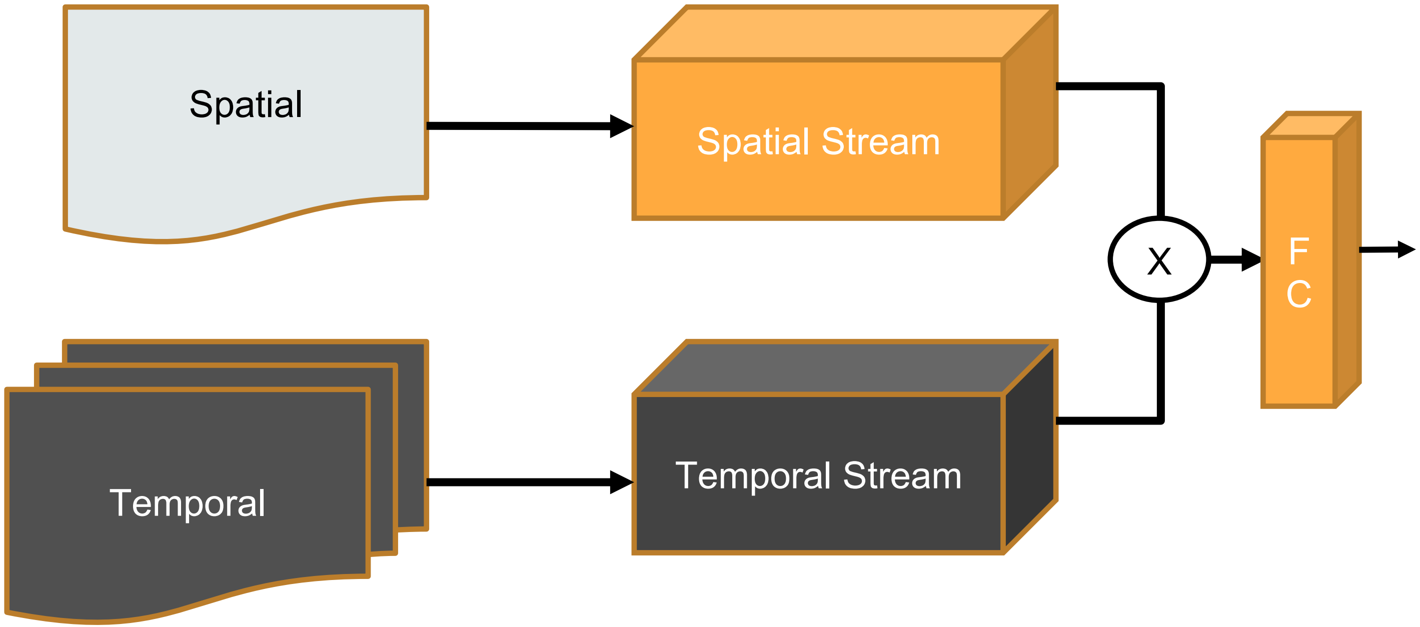
More augmentations

Better for *temporal* datasets

Models


Multi-
stream
3D
Conv
Trans-
formers


Multi-
stream
3D
Conv
Trans-
formers


Spatial
Spatial Stream
Temporal
Temporal Stream
X
F
C

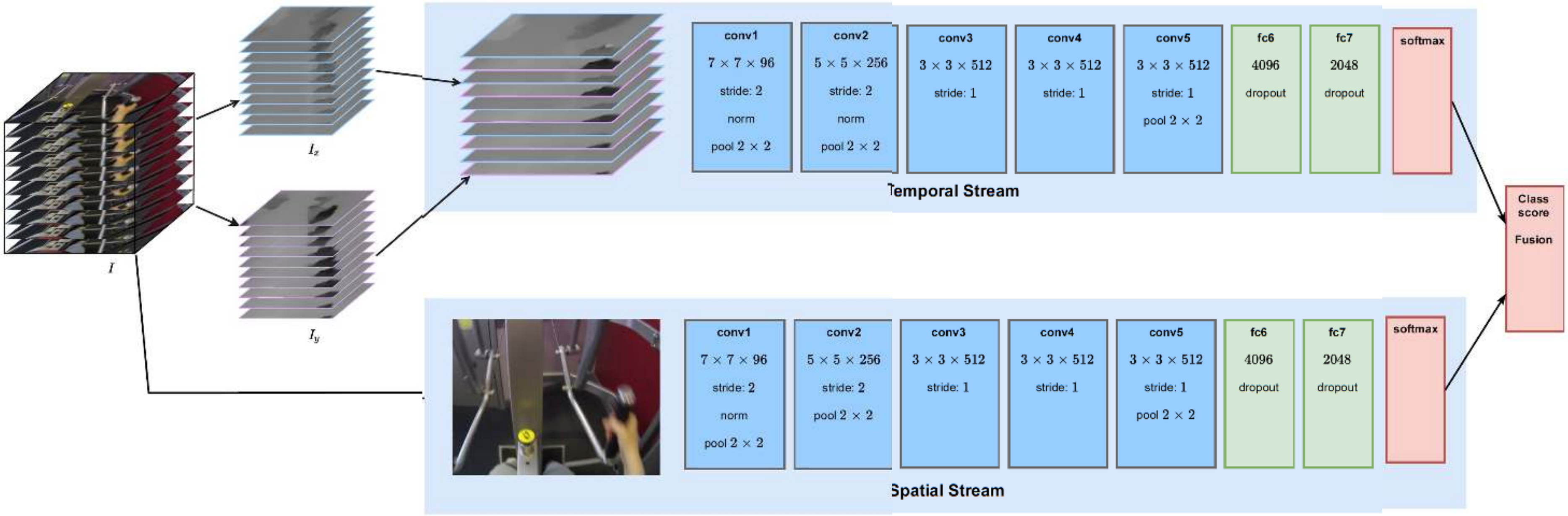


Figure by: Will Price, BSc Project, University of Bristol

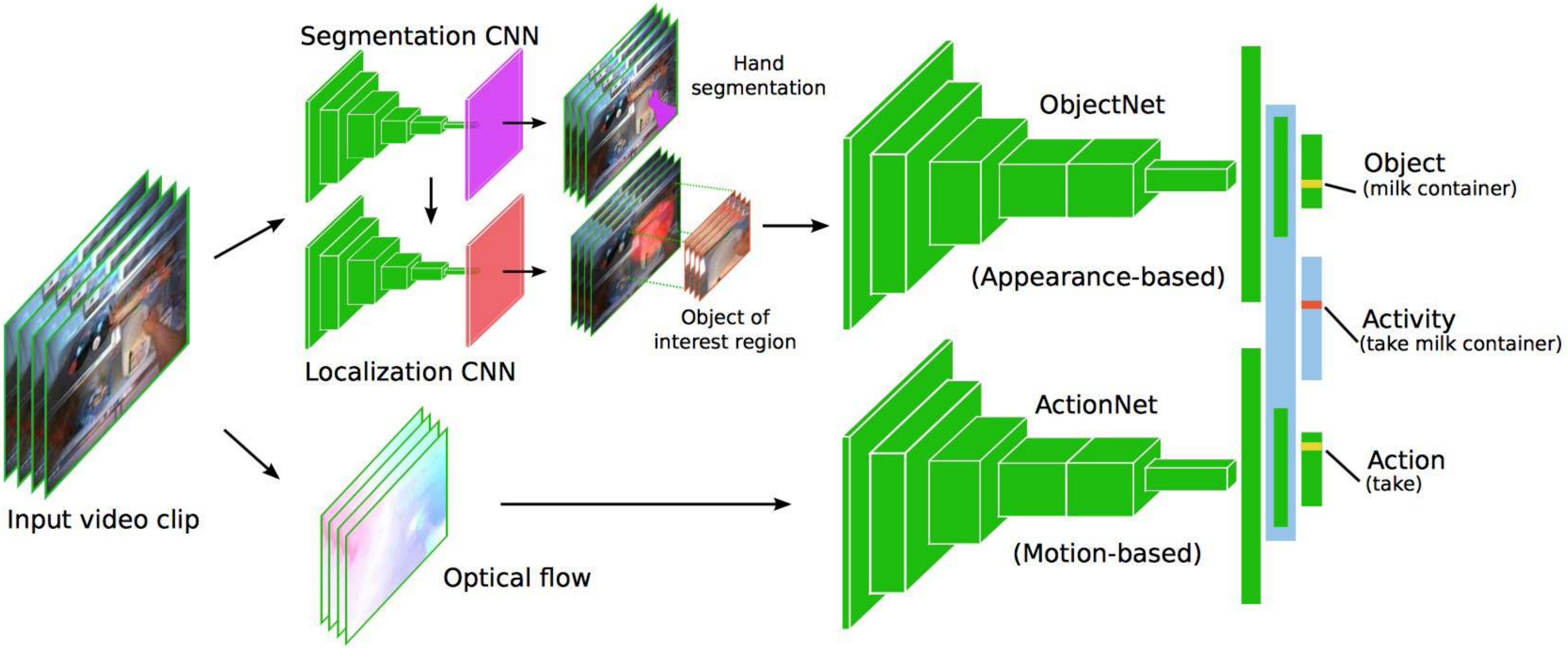
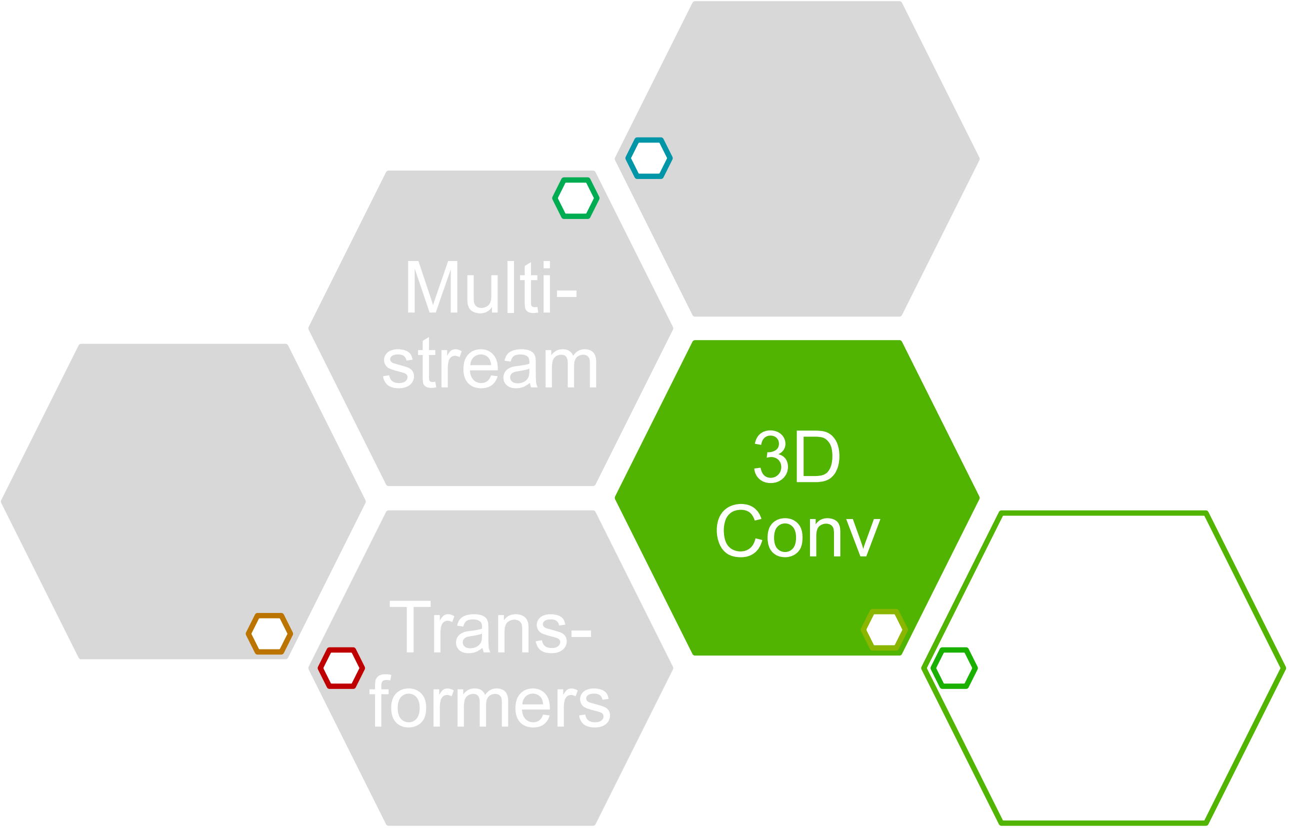
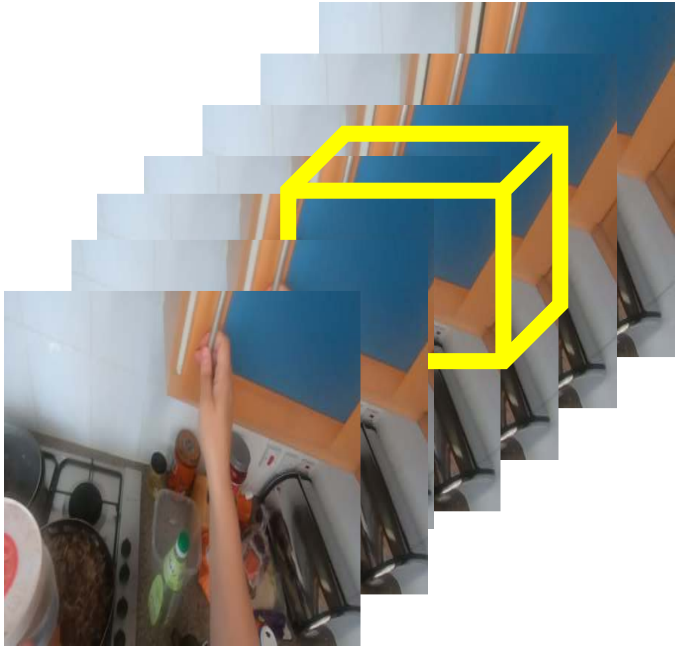
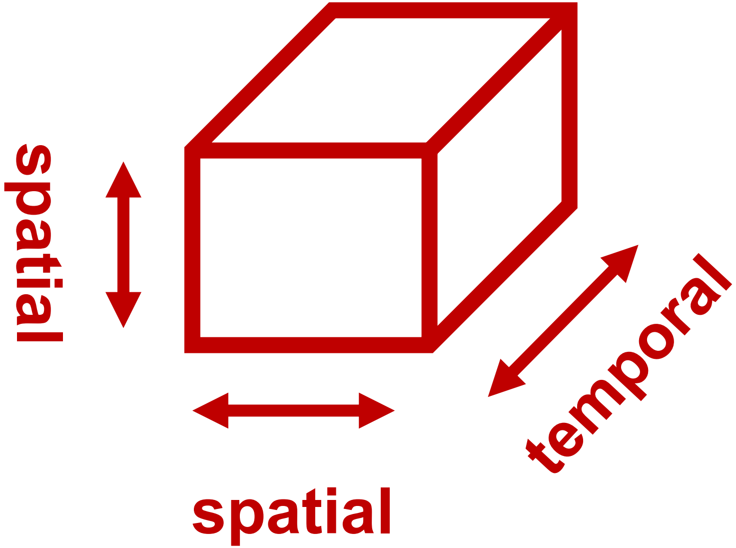


Figure from: Ma et al. Going Deeper into First-Person Activity Recognition. CVPR 2016


Multi-
stream
3D
Conv
Trans-
formers

3D Convolutional





3D Convolutional

- Initial attempts required initialisation from 2D Networks (i.e. ImageNet)
  - No presence of large scale video dataset
  - Inflated networks (I3D) Carriera et al

- The objective was to remove the need for optical flow…

3D Convolutional

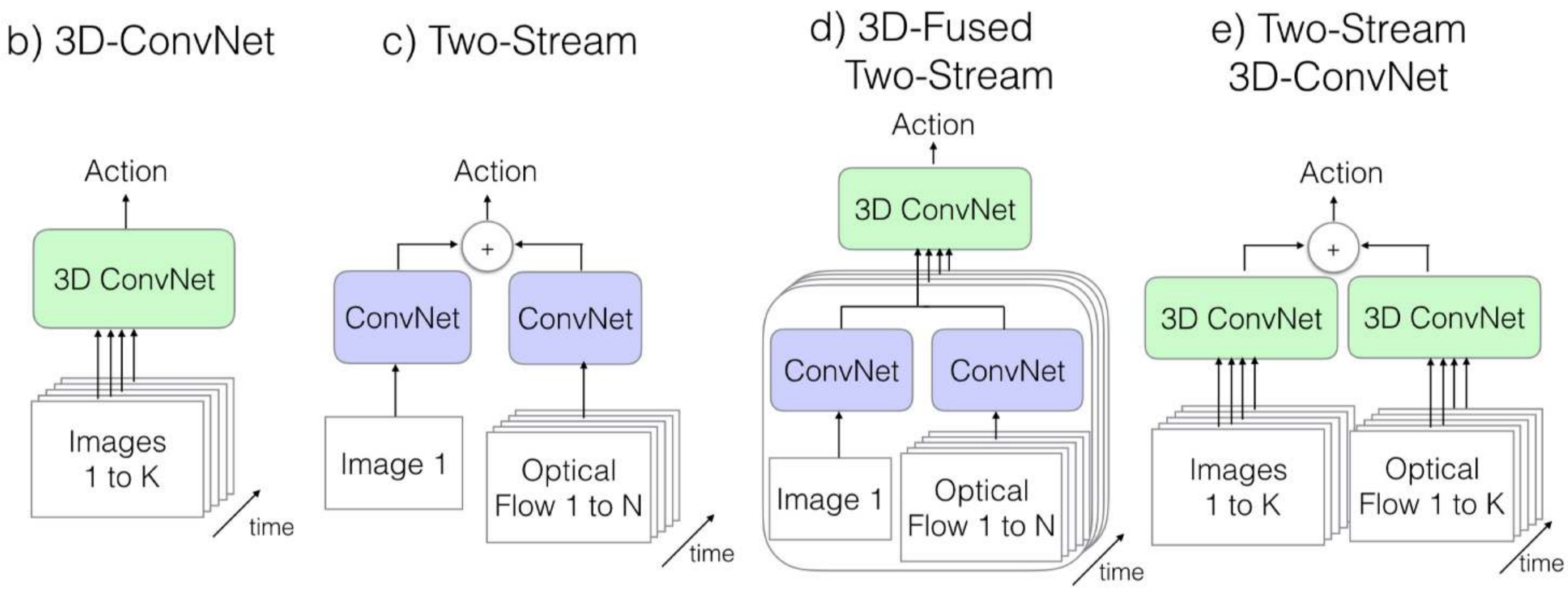


Figure from: Carriera et al. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. CVPR 2017

Via knowledge distillation

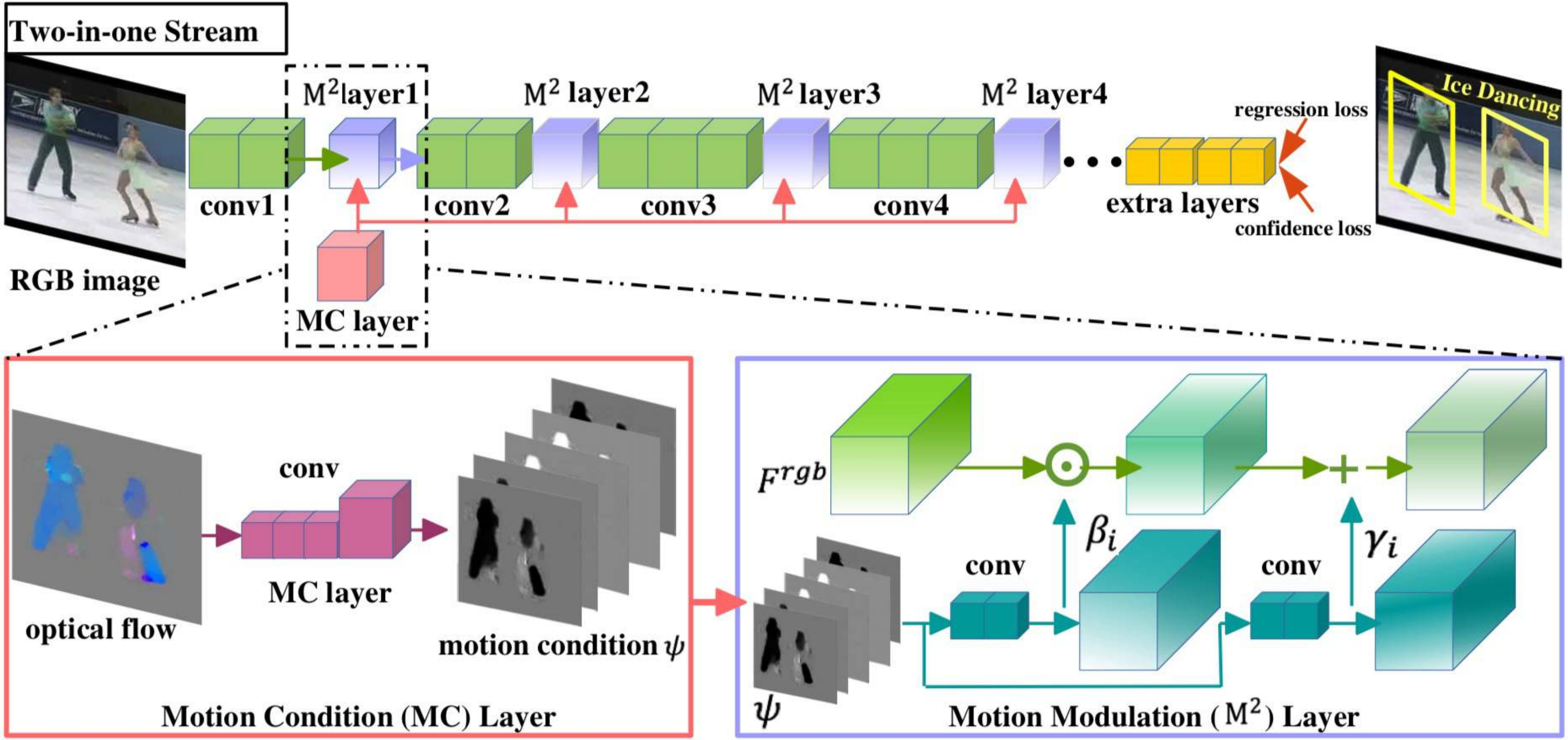
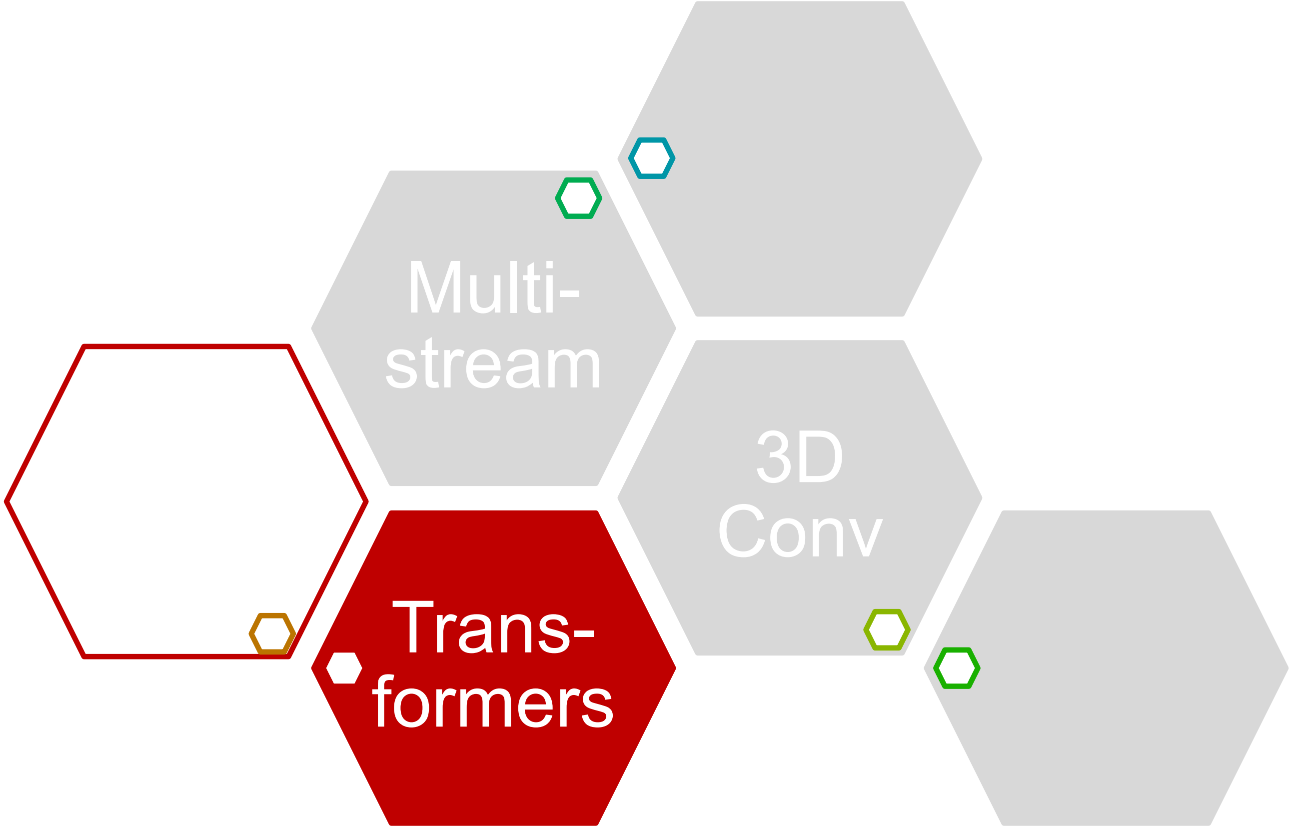
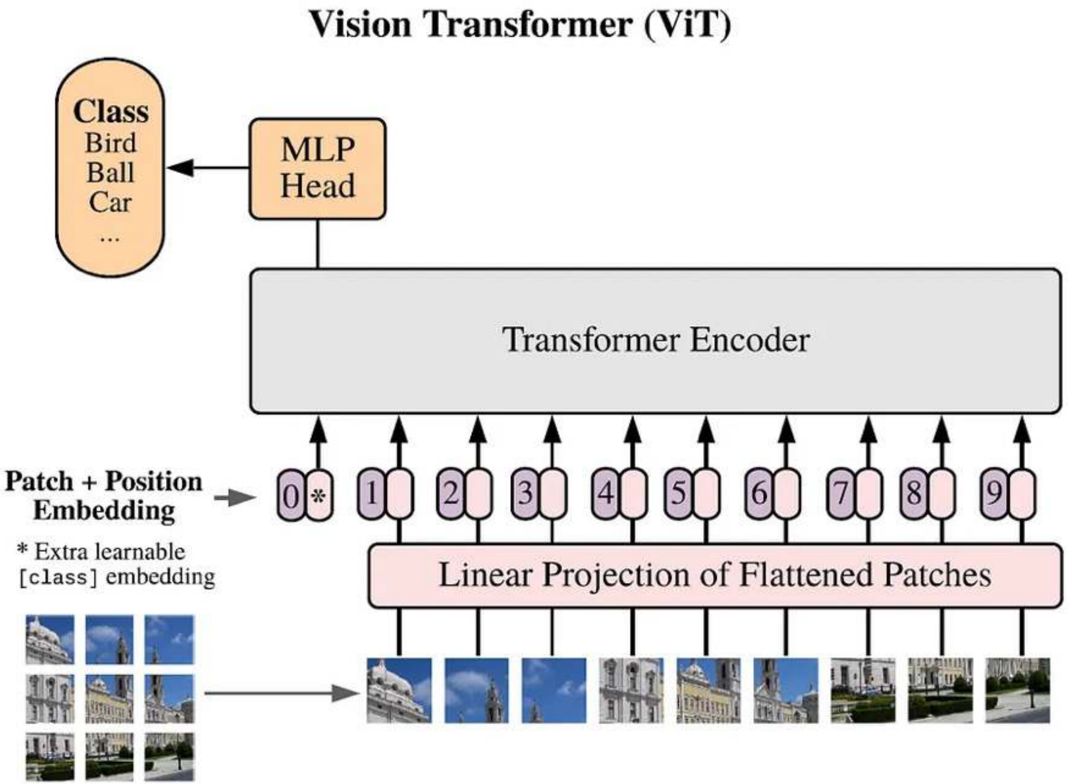
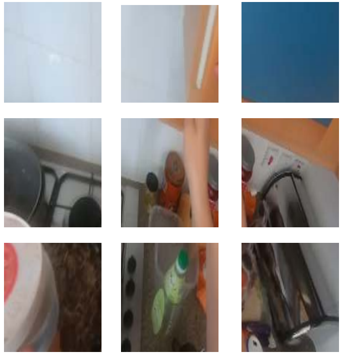


Figure from: Zhao and Snoek. Dance with Flow: Two-in-One Stream Action Detection. CVPR 2019

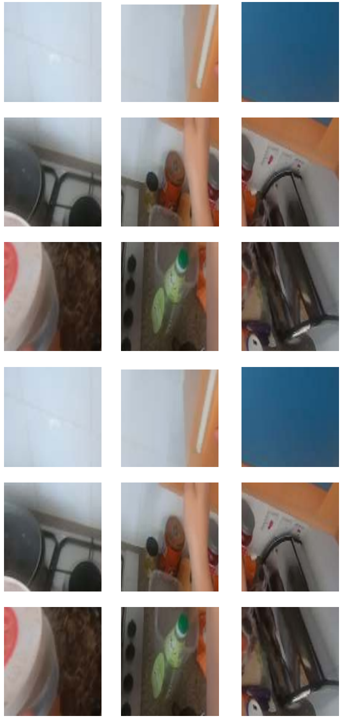

Multi-
stream
3D
Conv
Trans-
formers


Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP
Head
Transformer Encoder
Patch + Position
Embedding
* Extra learnable
[class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

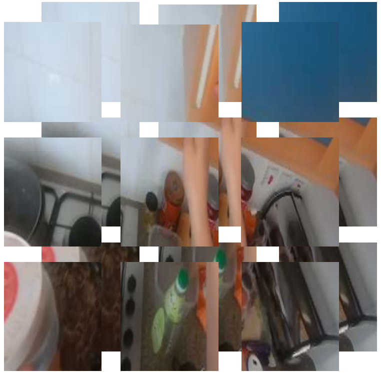
How to Patch-ify a Video?



H x W x C



[TH] x W x C

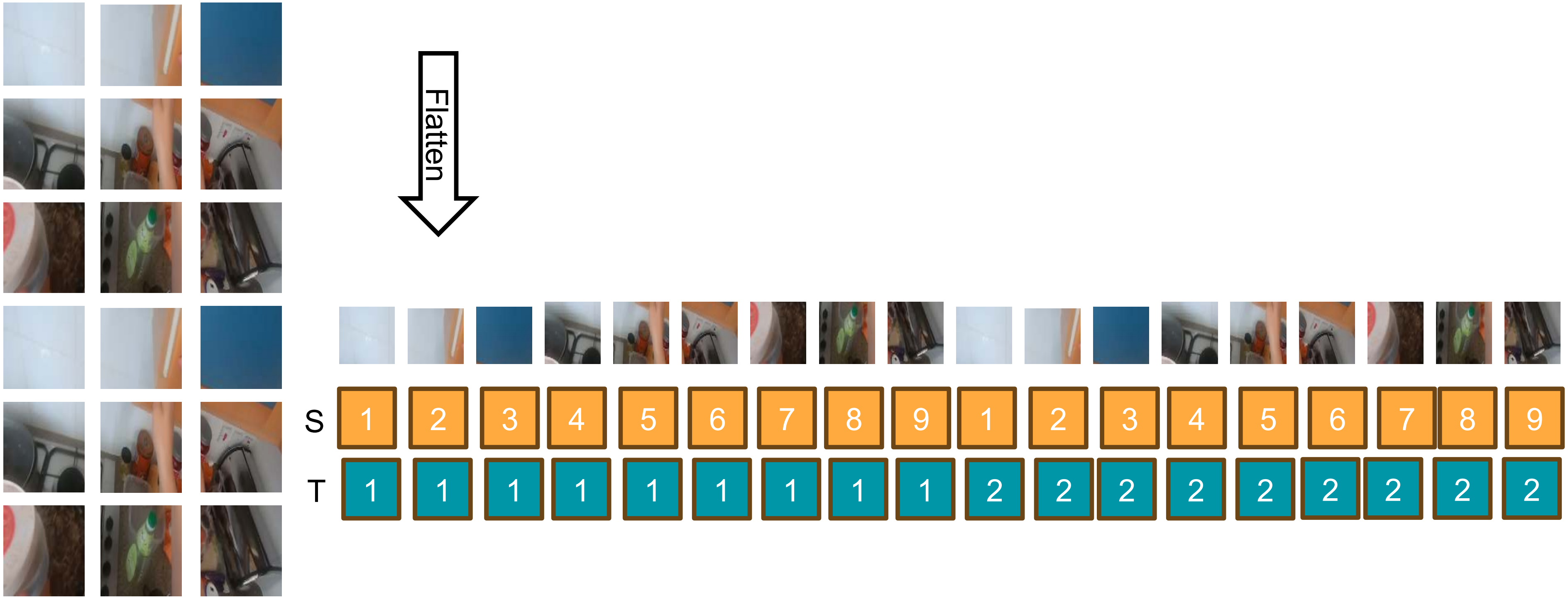


H x W x [CT]

How to Patch-ify a Video?

[TH] x W x C

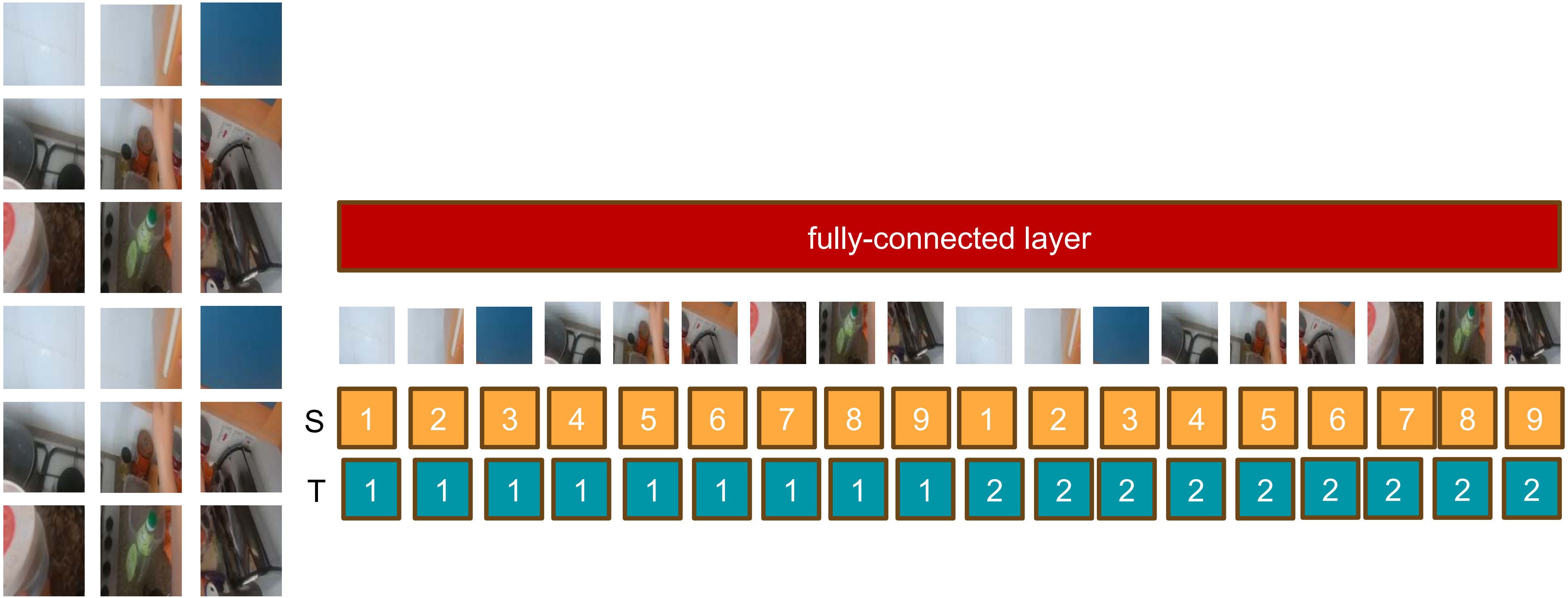
How to Patch-ify a Video?



[TH] x W x C

University of BRISTOL

How to Patch-ify a Video?



[TH] x W x C

TimeSFormer

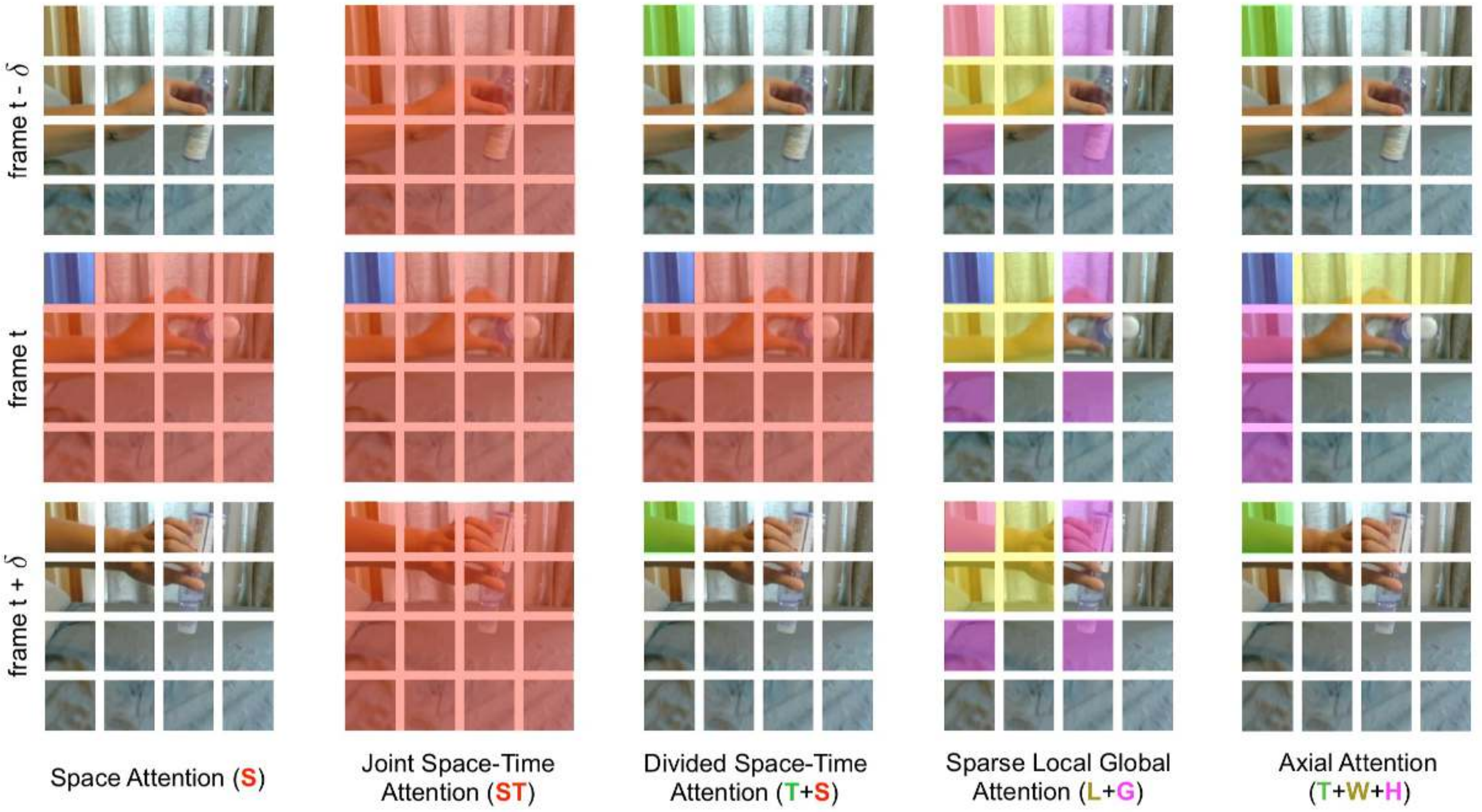


Figure from: Bertasius et al. Is Space-Time Attention All You Need for Video Understanding? ArXiv 2021

MotionFormer

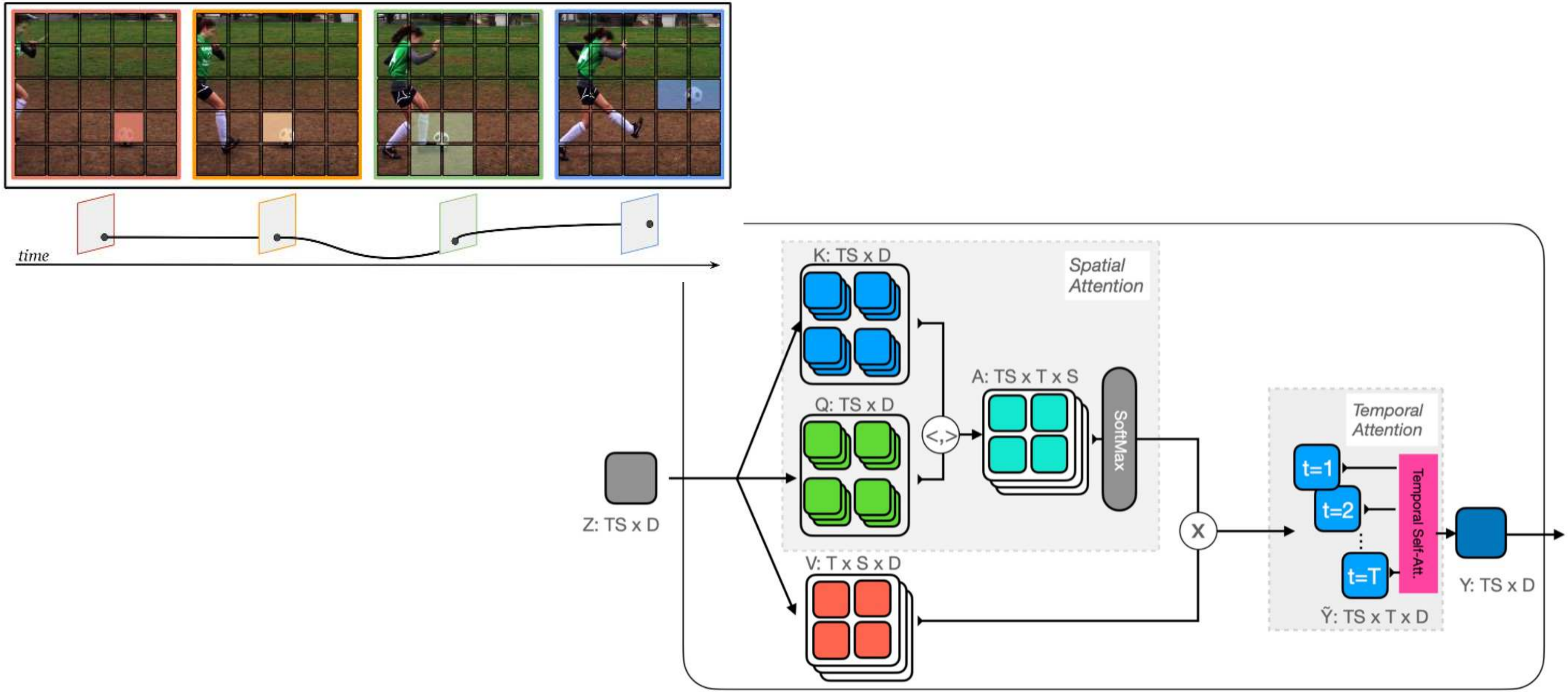


Figure from: Patrick et al. Keeping Your Eye on the Ball: Trajectory Attention in Video Transformers. NeurIPS 2021

VideoMAE

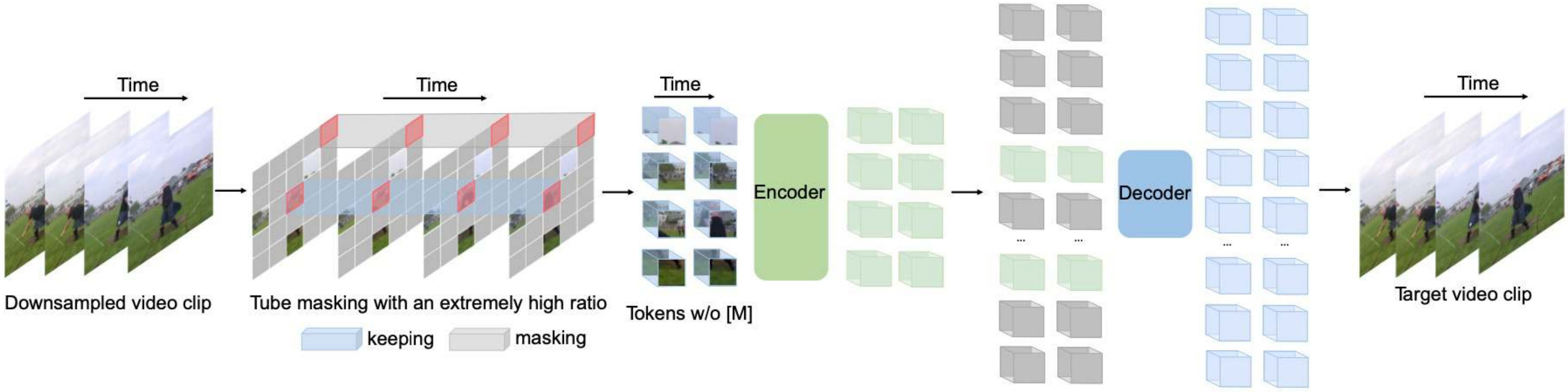


Figure from: Tong et al (2022). VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. NeurIPS 2022

OmniVore

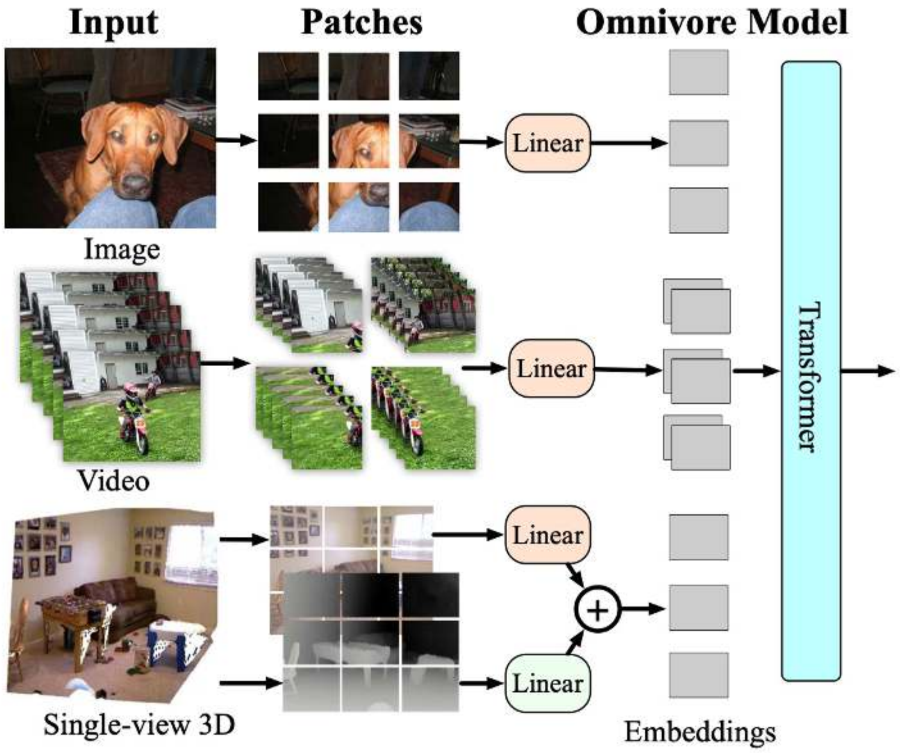
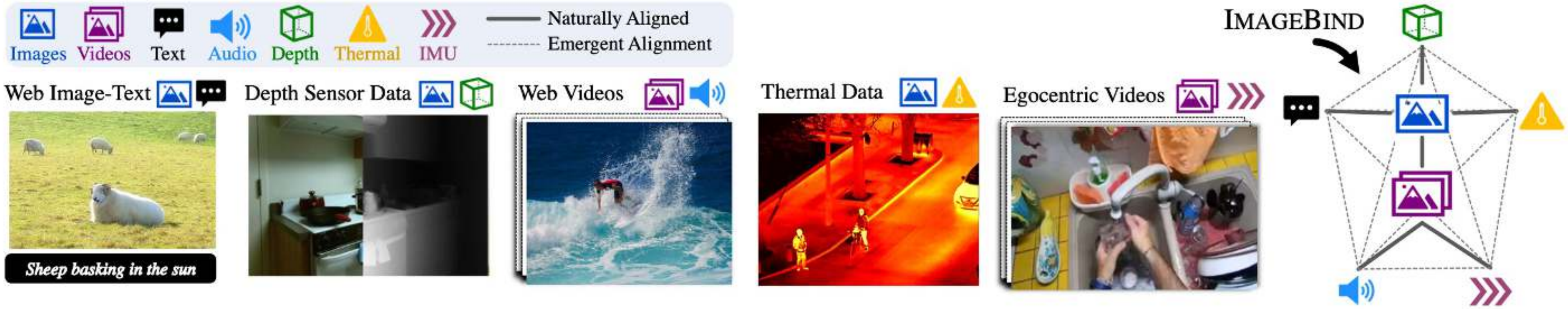


Figure 2. Multiple visual modalities in the OMNIVORE model.

Figure from: Gridhar et al (2022). OMNIVORE: A Single Model for Many Visual Modalities. CVPR

ImageBind



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

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

Flamingo

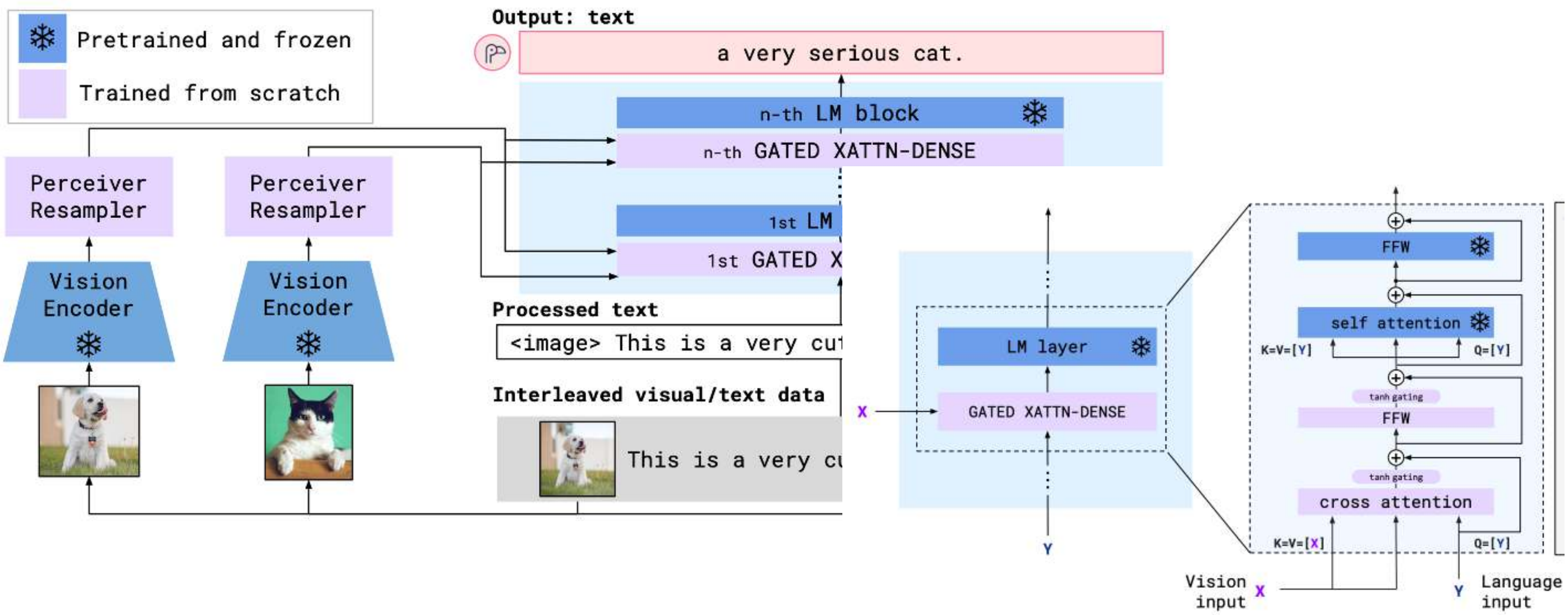


Figure from: Alayrac et al (2022). Flamingo: a Visual Language Model for Few-Shot Learning. NeurIPS

InternVideo

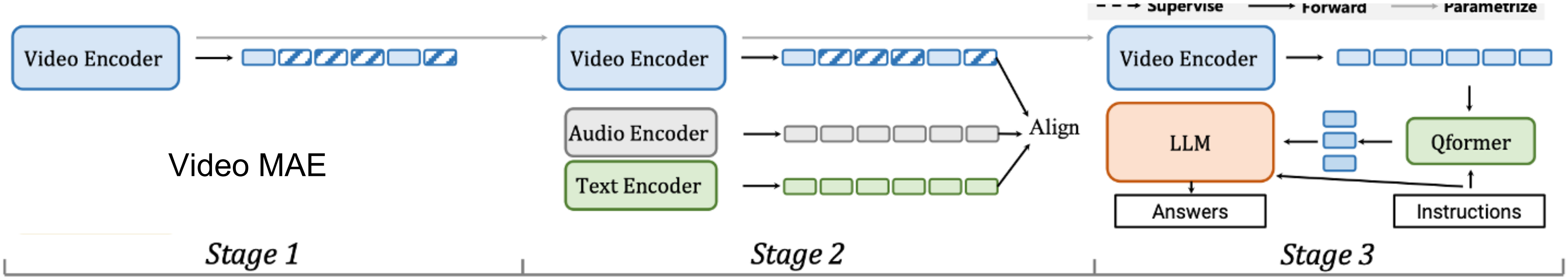
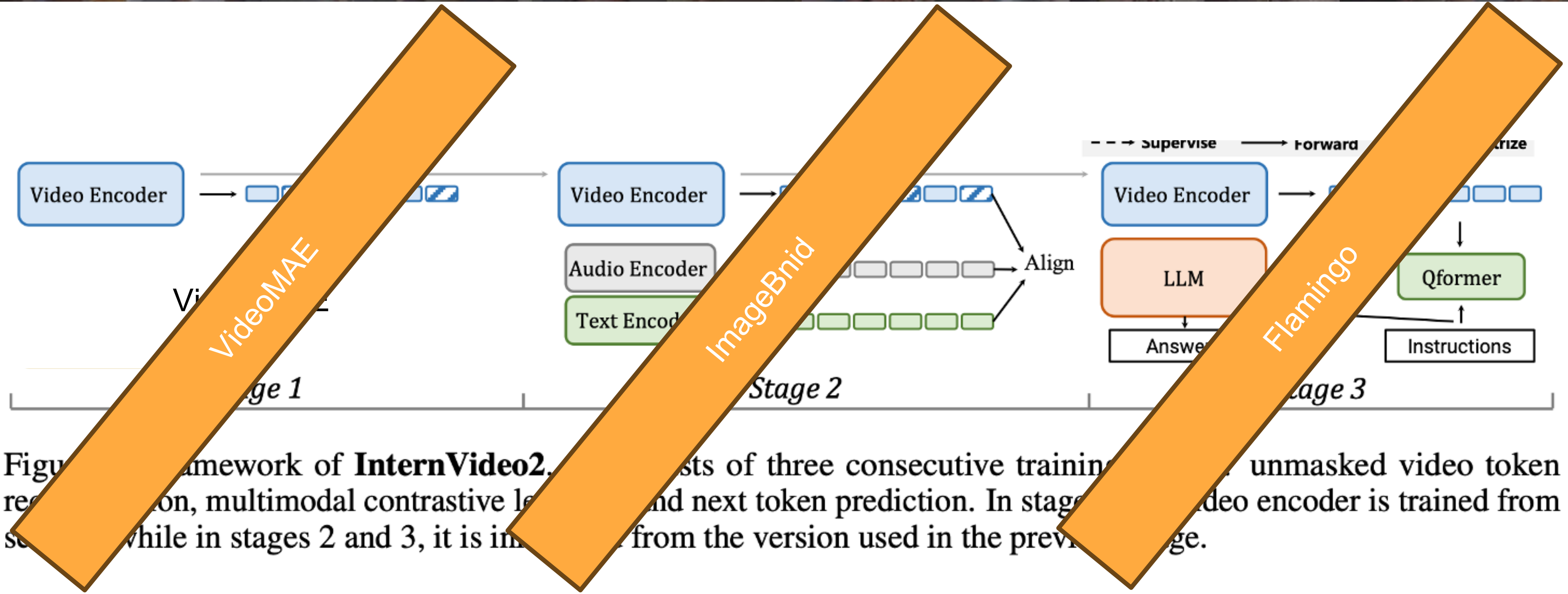


Figure 2: Framework of **InternVideo2**. It consists of three consecutive training phases: unmasked video token reconstruction, multimodal contrastive learning, and next token prediction. In stage 1, the video encoder is trained from scratch, while in stages 2 and 3, it is initialized from the version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

Figure from: Wang et al (2024). InternVideo2: Scaling Video Foundation Models for Multimodal Video Understanding. ArXiv

InternVideo



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

Figure from: Wang et al (2024). InternVideo2: Scaling Video Foundation Models for Multimodal Video Understanding. ArXiv

We are still lacking the right models for video understanding

From Clip to Video

Video

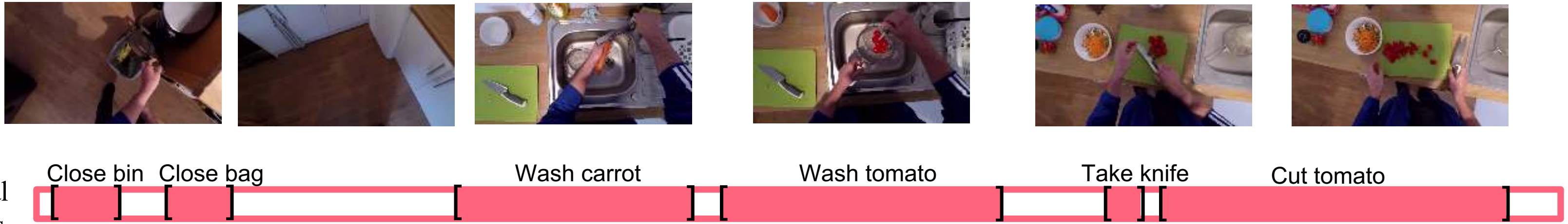
Wash carrot

From Clip to Video

From Clip to Video

Video

Visual labels



Most models work only within the clip… [ignoring the context]

Long-Term Feature Bank

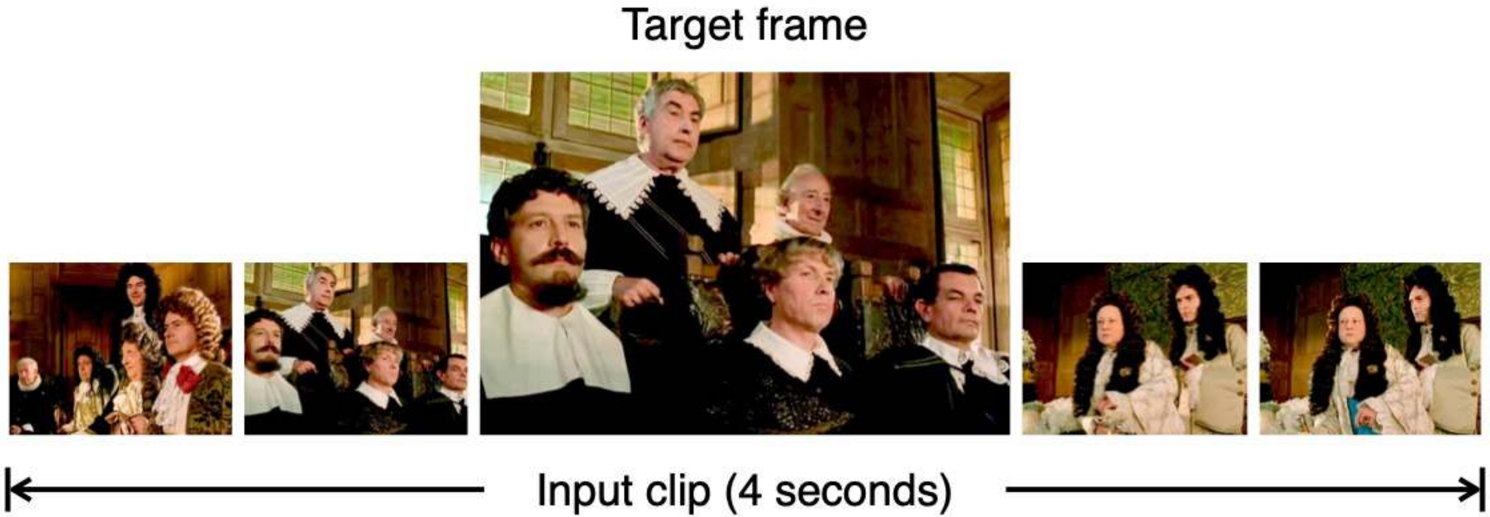
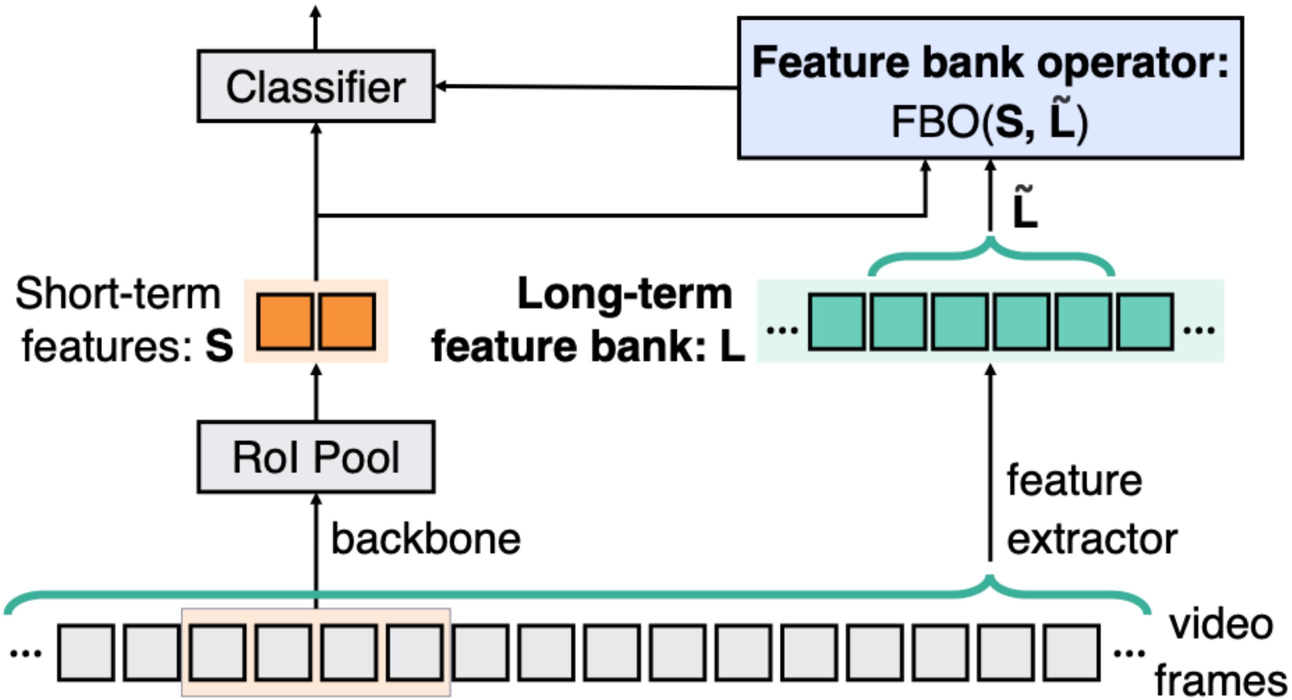


Figure 1. What are these people doing? Current 3D CNN video models operate on short clips spanning only ∼4 seconds. Without observing longer-term context, recognition is difficult. (Video from the AVA dataset [14]; see next page for the answer.)

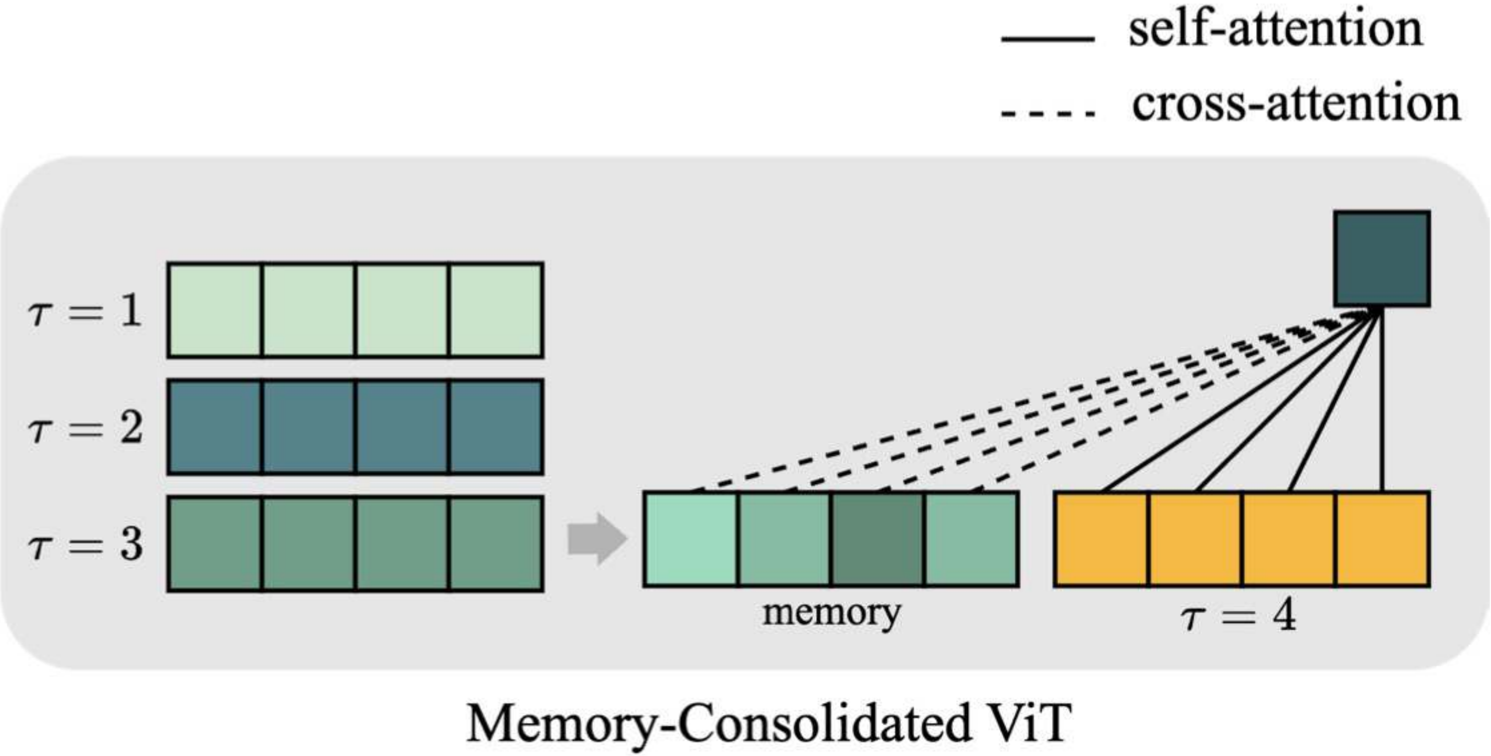
Long-Term Feature Bank



(b) 3D CNN with a Long-Term Feature Bank (Ours)

Wu et al (2019). Long-Term Feature Banks for Detailed Video Understanding. CVPR

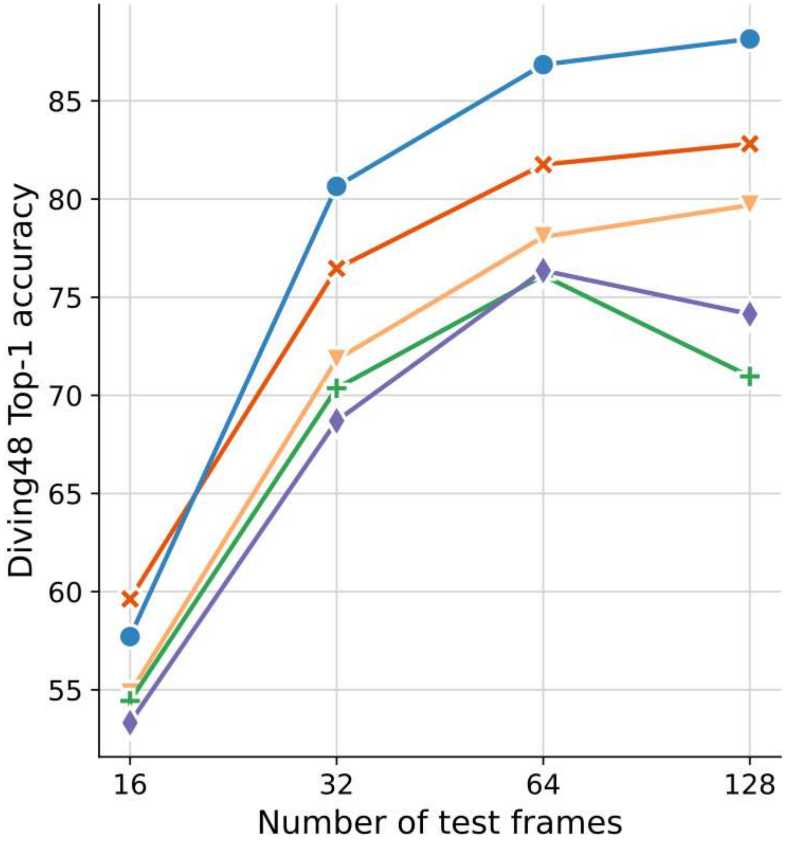
Memory Consolidation



University of BRISTOL

Balazevic´ et al (2024). Memory Consolidation Enables Long-Context Video Understanding. ICML

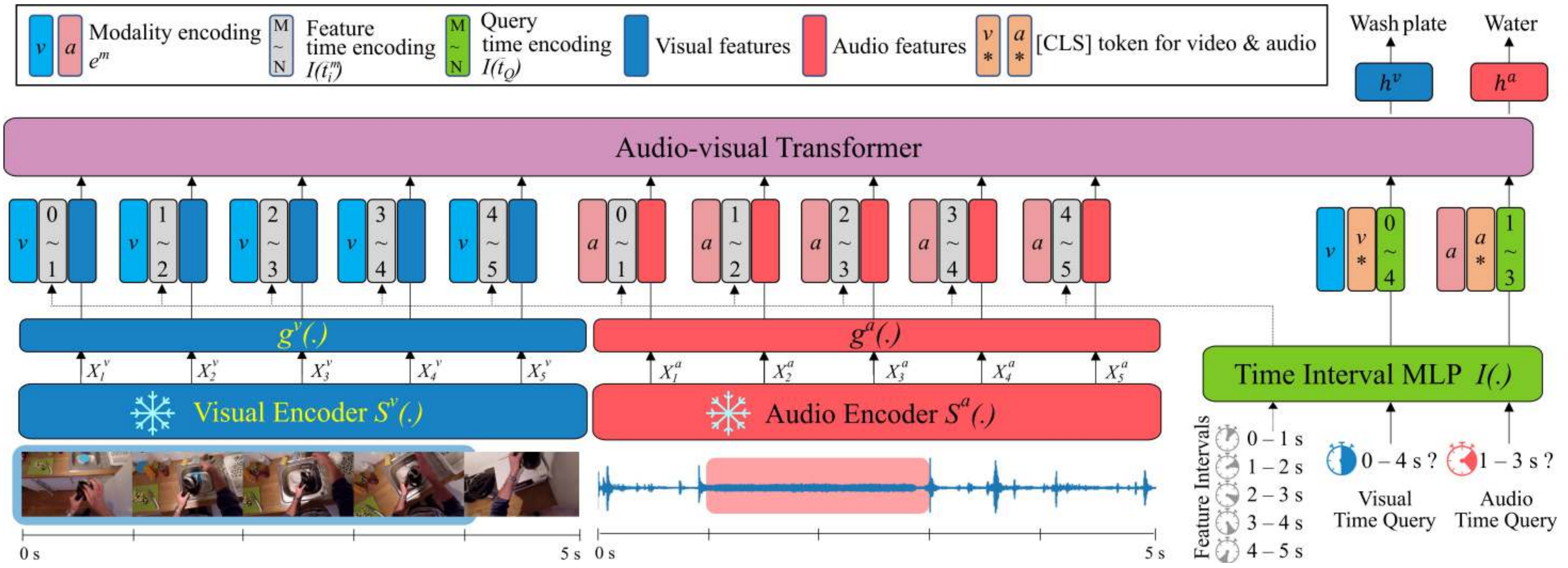
Memory Consolidation



Balazevic´ et al (2024). Memory Consolidation Enables Long-Context Video Understanding. ICML

TIM: A Time-Interval Audio-Visual Machine

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




Do you think video is a unique modality?!!

On Tasks…

Two types of video understanding tasks

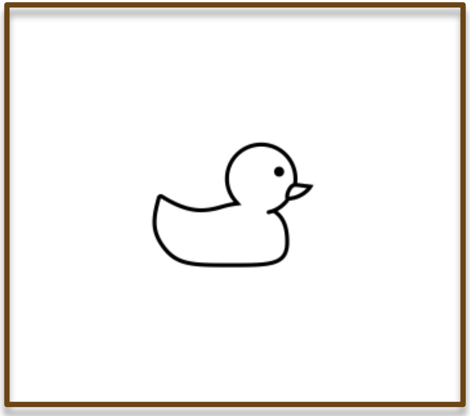
Video Understanding Tasks

- Analogous to Image-based Tasks
- Temporal tasks (not relevant for images)

Analogous Tasks

Image

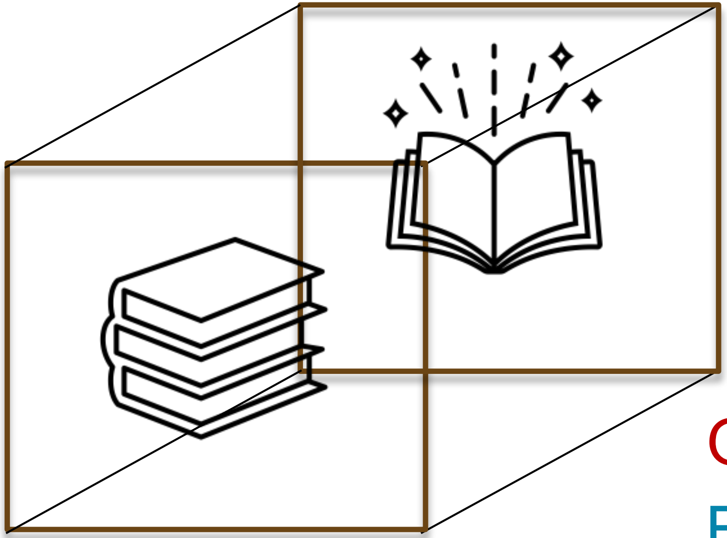
- Object Recognition



Duck

Video

- Action Recognition



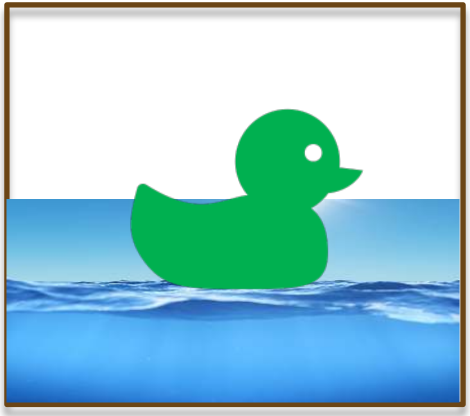
Open

Book

Analogous Tasks

Image

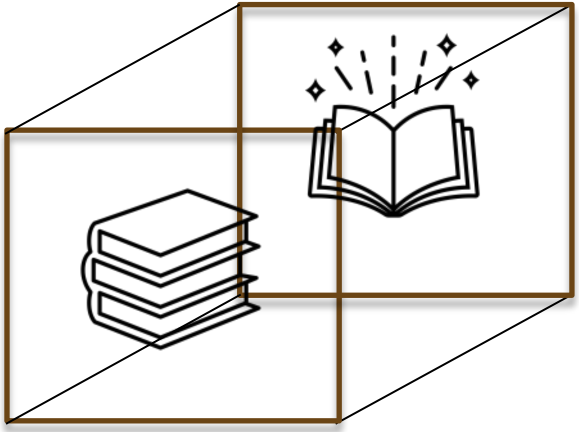
- Image Captioning



A green duck swimming
In clear water

Video

- Video Captioning

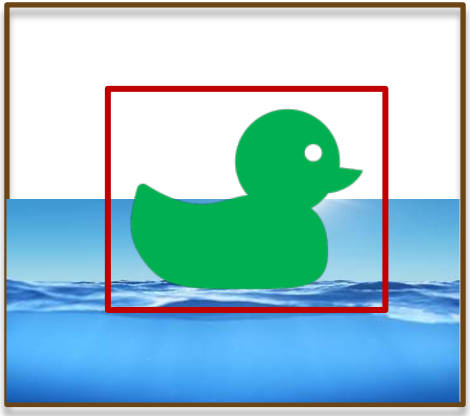


A book picked from top of the pile
and opened to a page in the middle

Analogous Tasks

Image

- Object Detection



Duck

Video

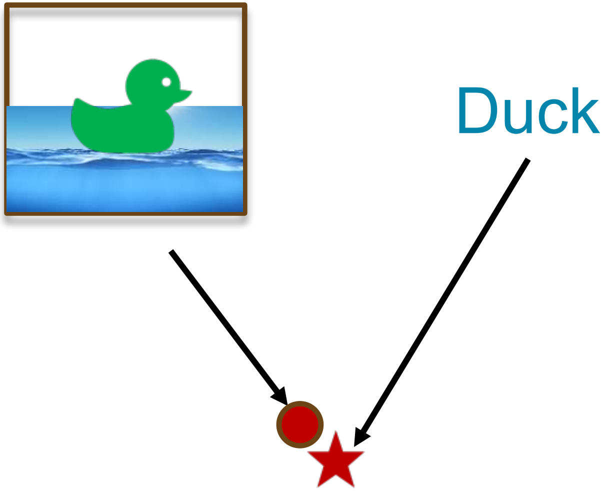
- Action Detection

Open Book

Analogous Tasks

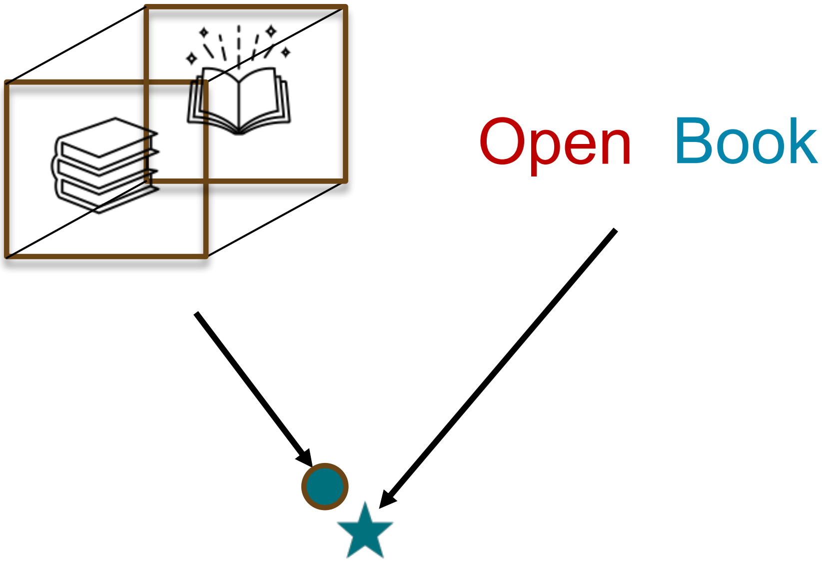
Image

- Image Retrieval



Video

- Video Retrieval



What is a Cross-Modal Video Retrieval?

Video-to-Text Retrieval Task

Ranked Text – Gallery (or Retrieval Set)

Q

put garlic down

Text-to-Video Retrieval Task

Ranked Video – Gallery (or Retrieval Set)

Q put garlic down

with: Michael Wray
Gabriela Csurka
Diane Larlus

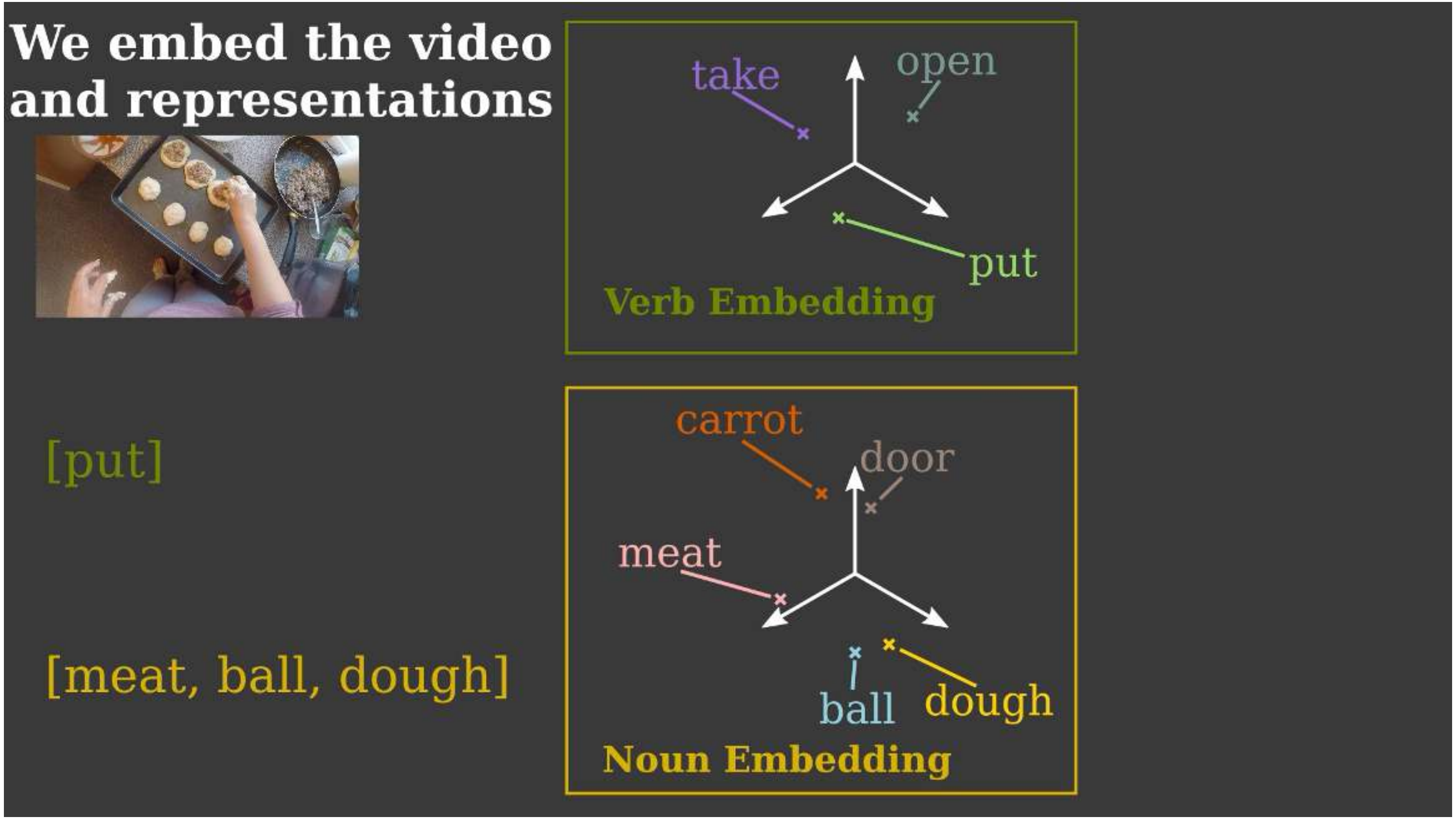
In this work we focus on
Fine-Grained Action Retrieval

I put meat on a ball of dough ⟷

M Wray, D Larlus, G Csurka, D Damen (2019). Fine-Grained Action Retrieval through Multiple Parts-of-Speech Embeddings. ICCV

Fine-Grained Action Retrieval

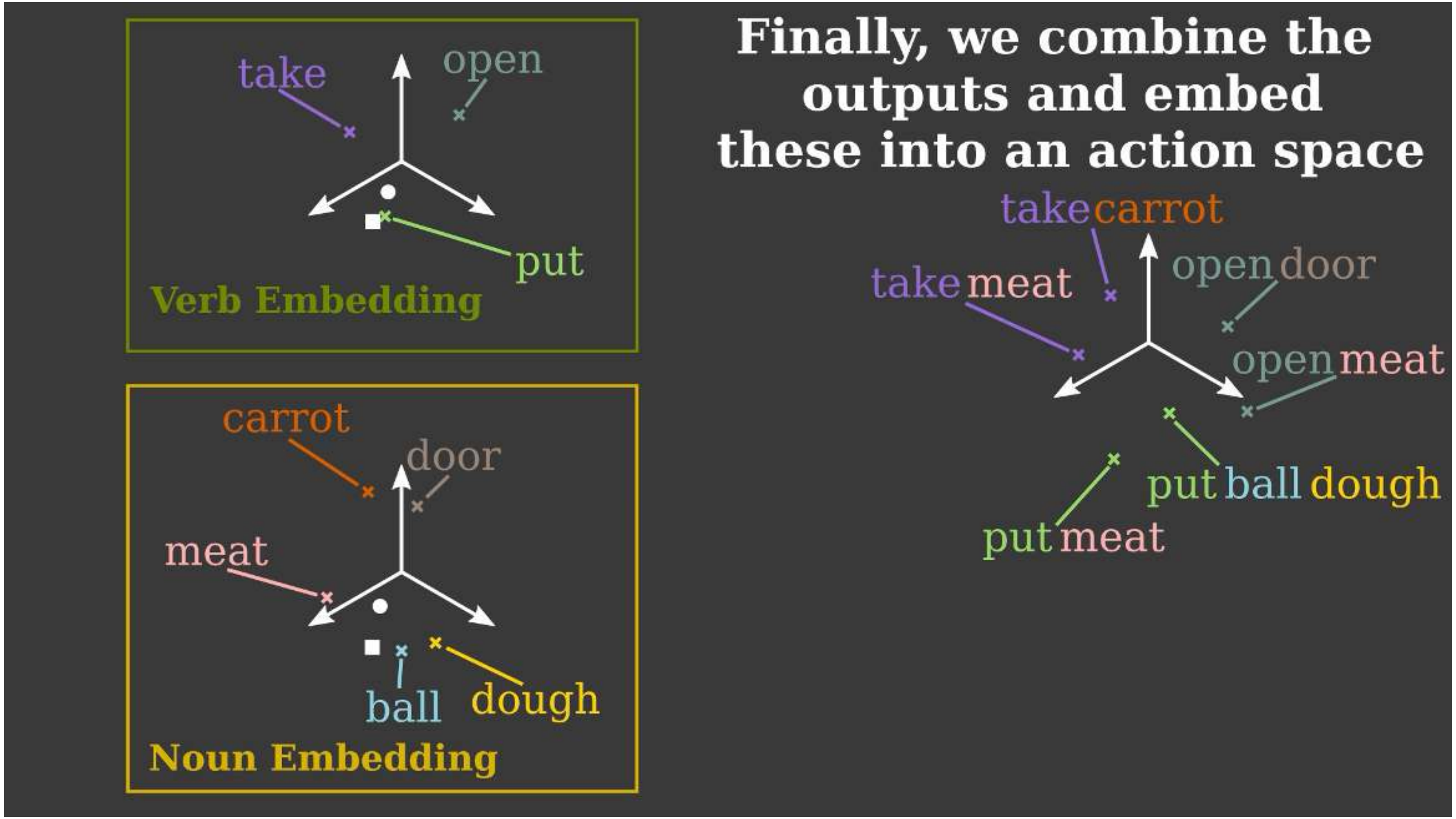
with: Michael Wray
Gabriela Csurka
Diane Larlus



M Wray, D Larlus, G Csurka, D Damen (2019). Fine-Grained Action Retrieval through Multiple Parts-of-Speech Embeddings. ICCV

Fine-Grained Action Retrieval

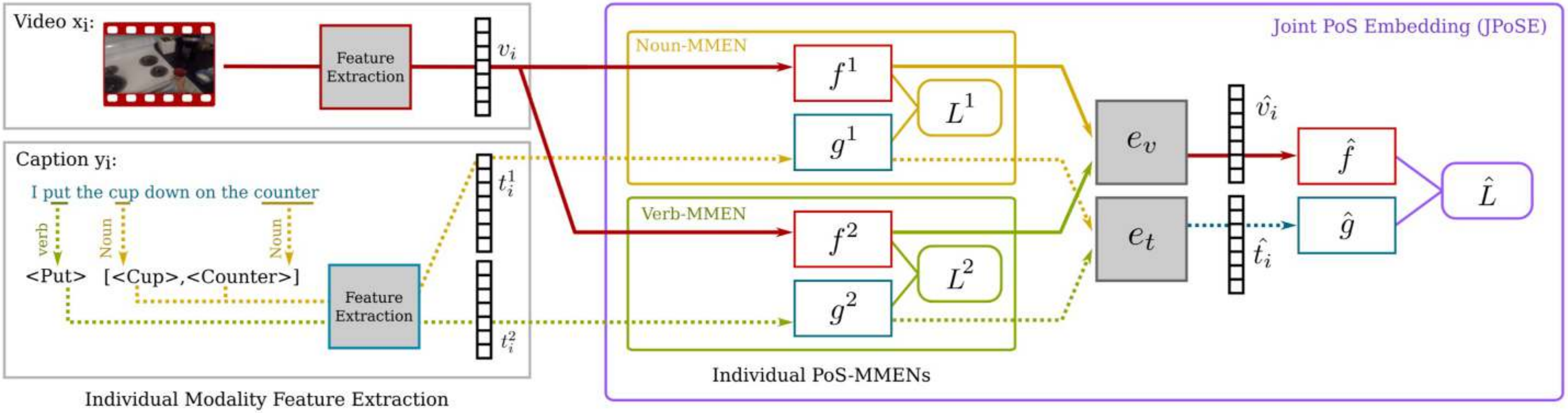
with: Michael Wray
Gabriela Csurka
Diane Larlus



M Wray, D Larlus, G Csurka, D Damen (2019). Fine-Grained Action Retrieval through Multiple Parts-of-Speech Embeddings. ICCV

Fine-Grained Action Retrieval

with: Michael Wray
Gabriela Csurka
Diane Larlus



M Wray, D Larlus, G Csurka, D Damen (2019). Fine-Grained Action Retrieval through Multiple Parts-of-Speech Embeddings. ICCV

Fine-Grained Action Retrieval

with: Michael Wray
Gabriela Csurka
Diane Larlus

Maximum activation examples for a neuron in a noun PoS Embedding (Cutting Board) - Figure 4

M Wray, D Larlus, G Csurka, D Damen (2019). Fine-Grained Action Retrieval through Multiple Parts-of-Speech Embeddings. ICCV

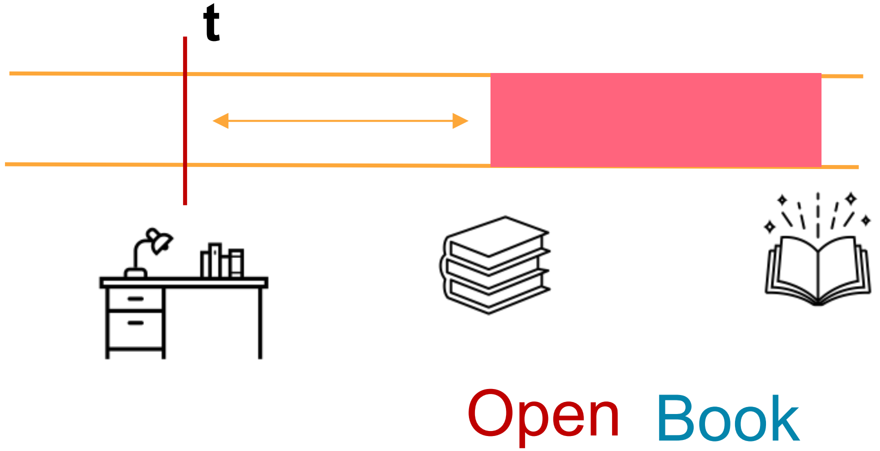
Non-Analogous Tasks

Image

Video

- Action Anticipation

What will happen after 1 second?



University of BRISTOL

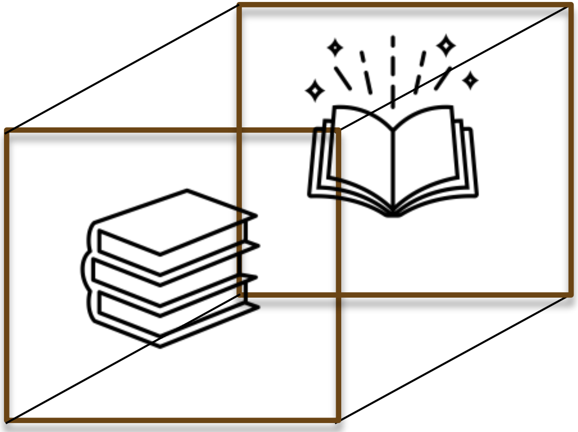
Non-Analogous Tasks

Image

Video

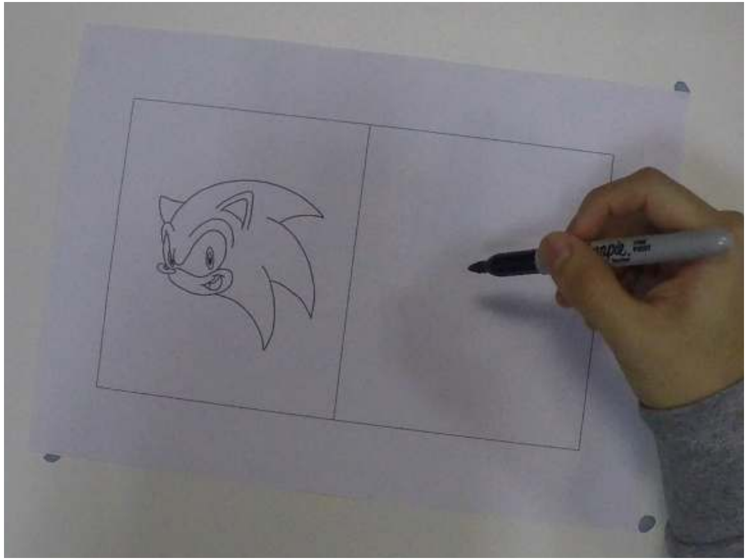
- Skill Understanding

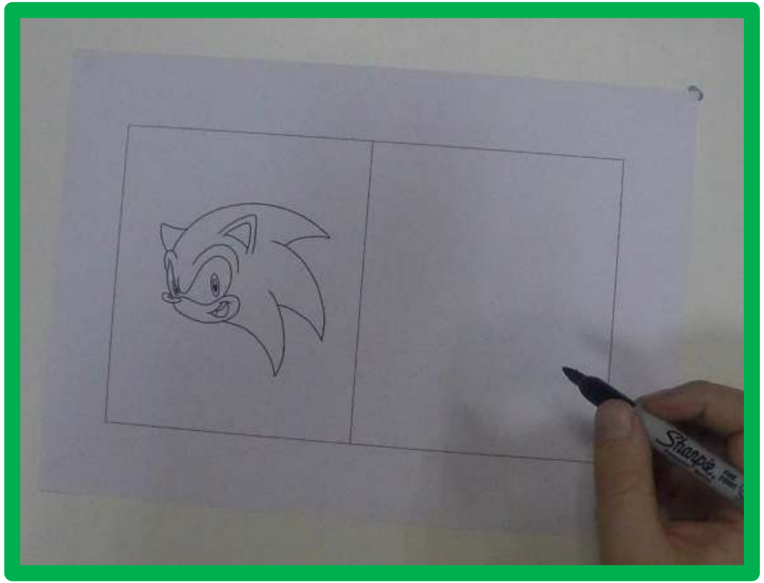
How did you open the book?



with: Hazel Doughty
Walterio Mayol-Cuevas

Pairwise annotations of videos, indicating higher skill or no skill preference



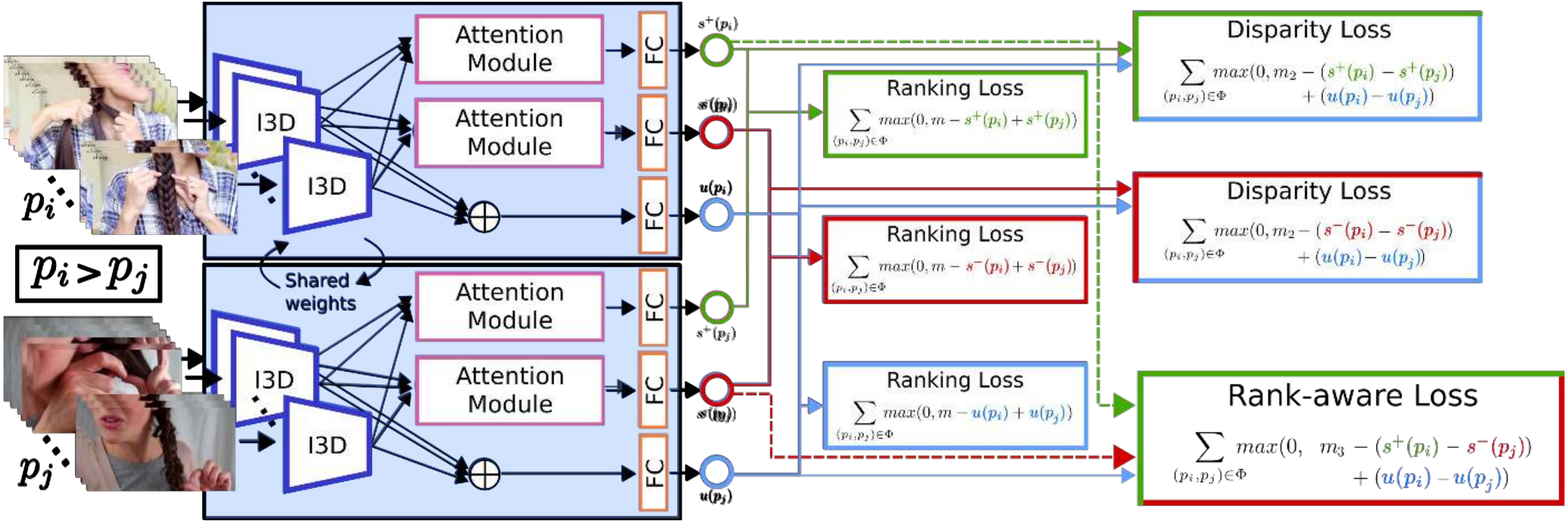


University of BRISTOL

H Doughty, W Mayol-Cuevas, D Damen (2019). The Pros and Cons: Rank-aware Temporal Attention for Skill Determination in Long Videos. *Computer Vision and Pattern Recognition (CVPR)*

Skill determination in video

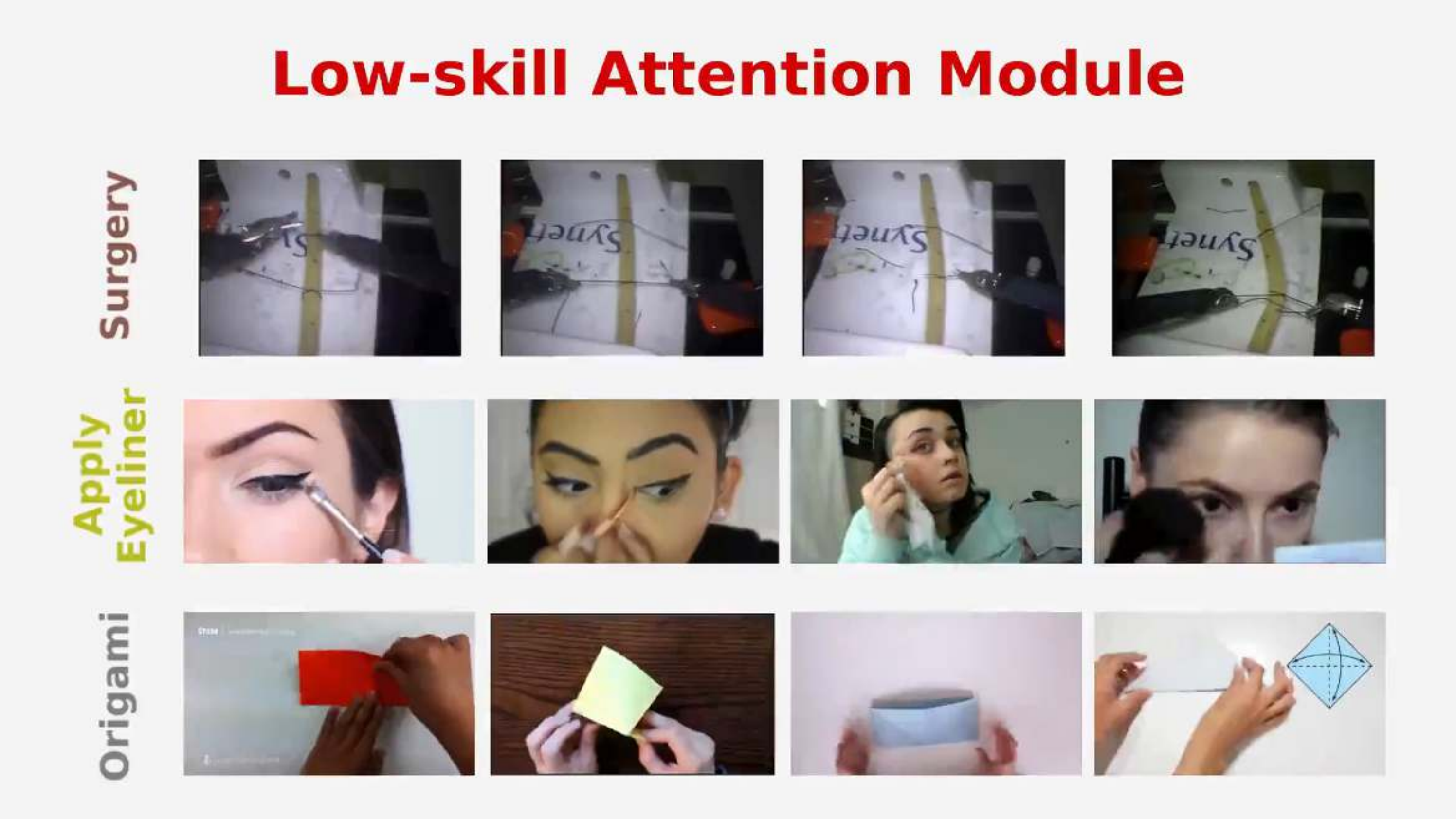
with: Hazel Doughty
Walterio Mayol-Cuevas



H Doughty, W Mayol-Cuevas, D Damen (2019). The Pros and Cons: Rank-aware Temporal Attention for Skill Determination in Long Videos. *Computer Vision and Pattern Recognition (CVPR)*

Skill determination in video

with: Hazel Doughty
Walterio Mayol-Cuevas



University of BRISTOL

H Doughty, W Mayol-Cuevas, D Damen (2019). The Pros and Cons: Rank-aware Temporal Attention for Skill Determination in Long Videos. *Computer Vision and Pattern Recognition (CVPR)*

Skill determination in video

with: Hazel Doughty
Walterio Mayol-Cuevas

University of BRISTOL

H Doughty, W Mayol-Cuevas, D Damen (2019). The Pros and Cons: Rank-aware Temporal Attention for Skill Determination in Long Videos. *Computer Vision and Pattern Recognition (CVPR)*

Analogous Tasks

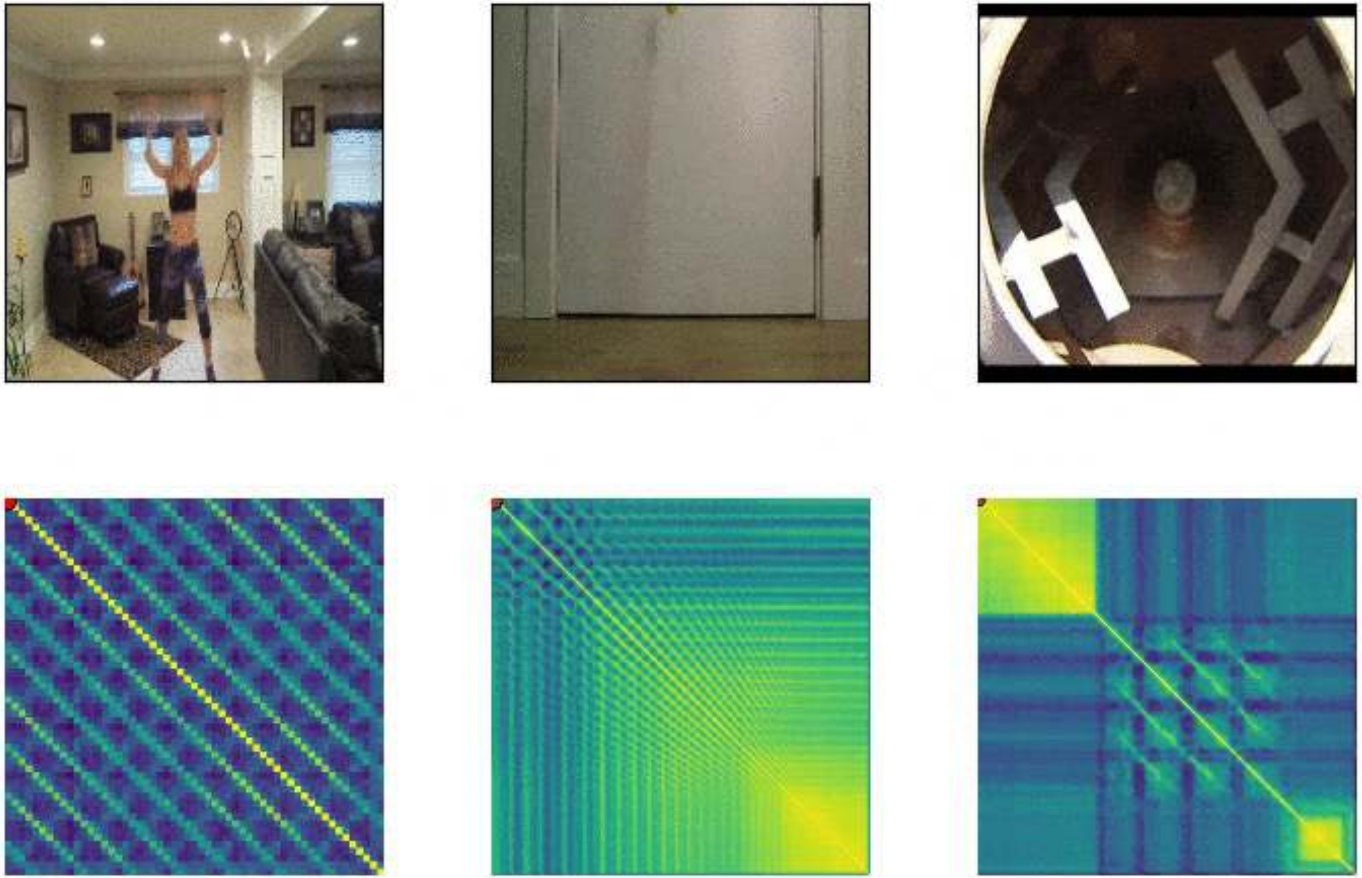
Image

- Object Counting

Video

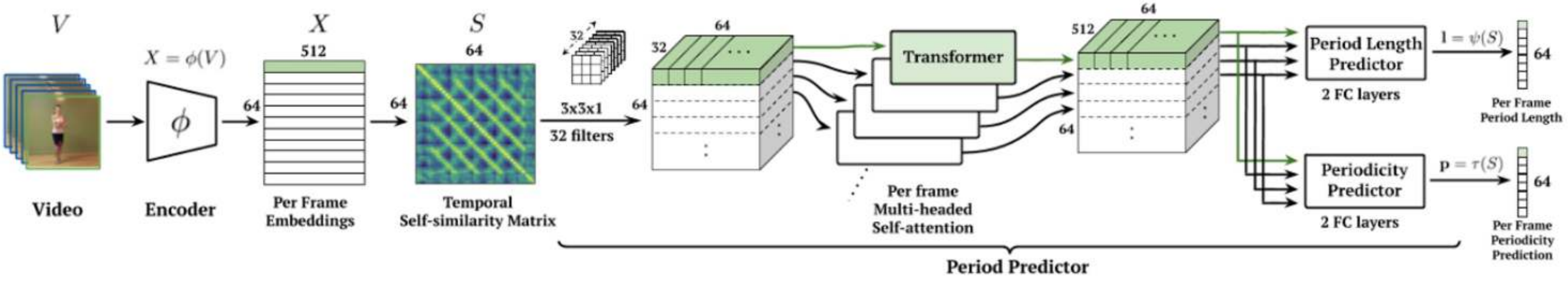
- Action Counting

Countix



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

Countix



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

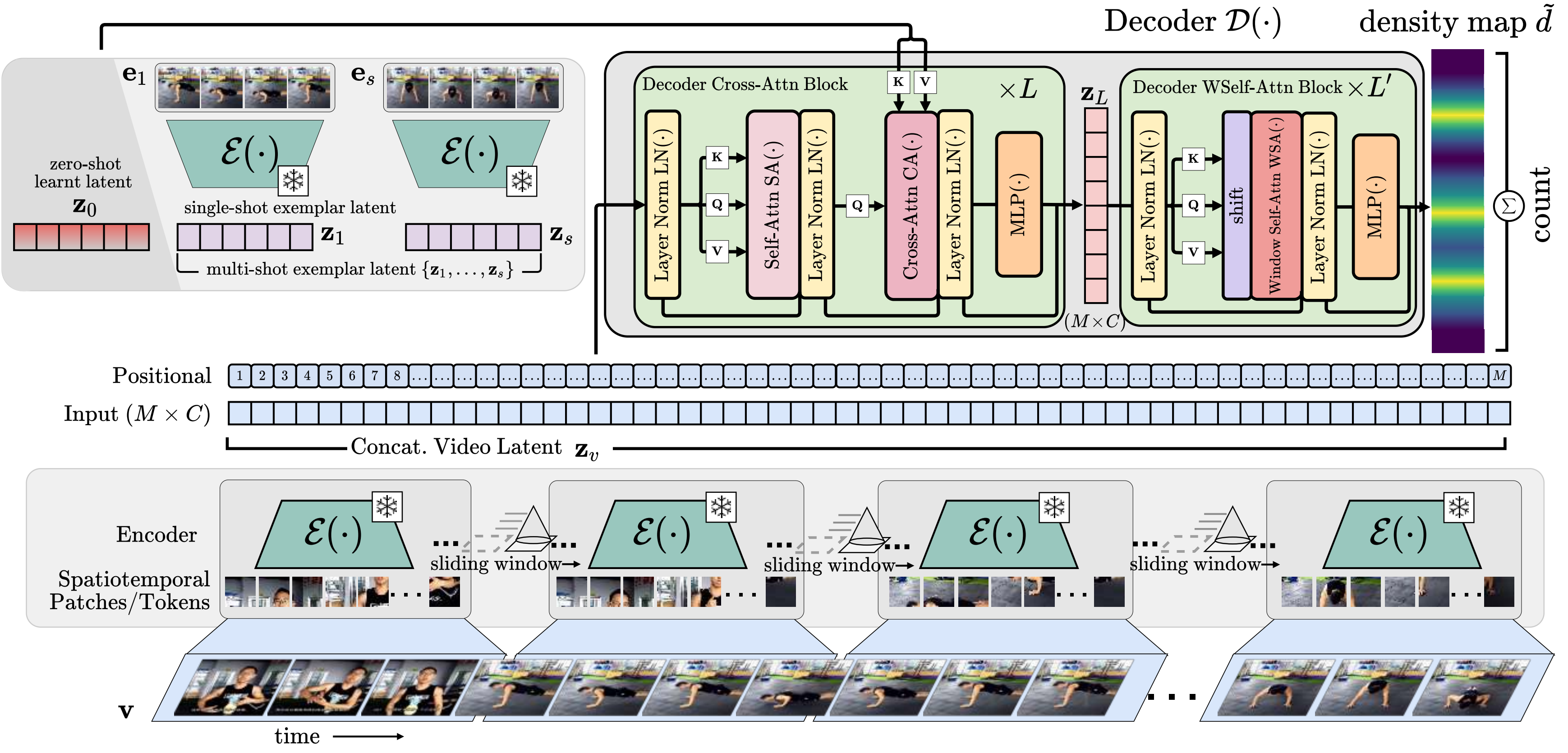
Every Shot Counts

with: Saptarshi Sinha
Alexandros Stergiou

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

Every Shot Counts

with: Saptarshi Sinha
Alexandros Stergiou



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

Every Shot Counts

with: Saptarshi Sinha
Alexandros Stergiou

(a) RepCount

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

(c) UCFRep

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

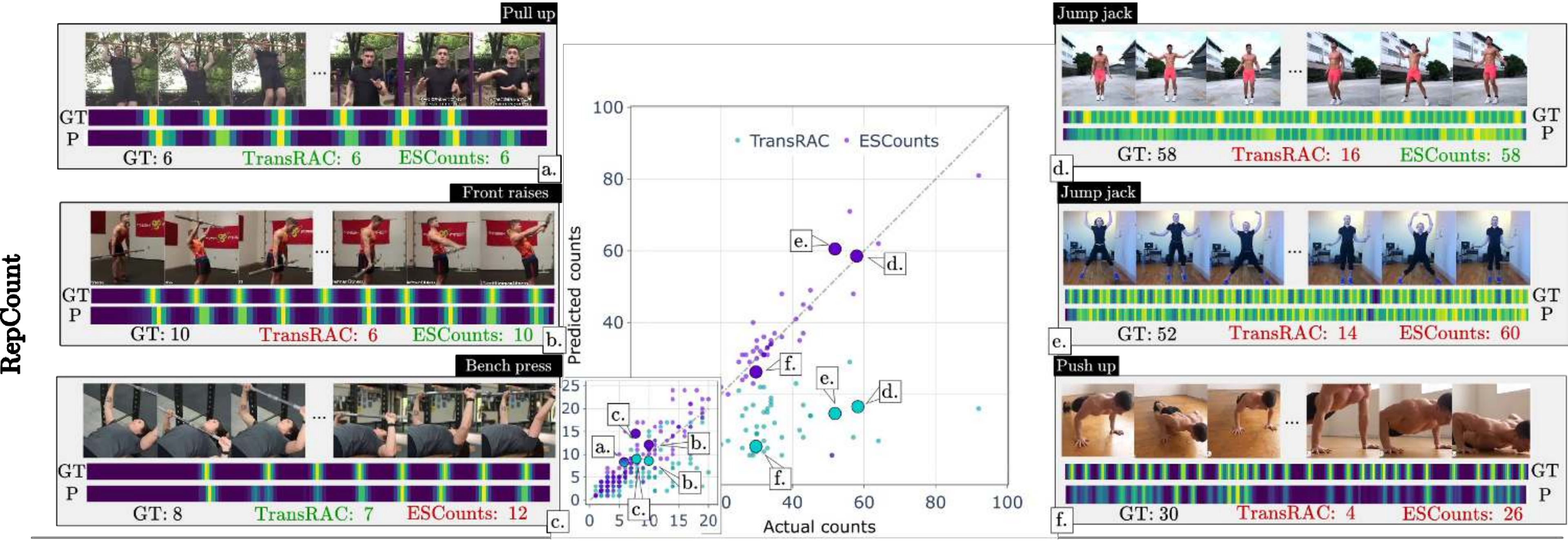
(b) Countix

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

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

Every Shot Counts

with: Saptarshi Sinha
Alexandros Stergiou



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

Analogous Tasks

Image

- Text-to-image Generation

Stable Diffusion

Video

- Text-to-Video Generation

SORA

Text-to-Video Generation

Prompt: Historical footage of California during the gold rush.

Text-to-Video Generation

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

Generative Video approaches do not yet understand physics, actions or action consequences…

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

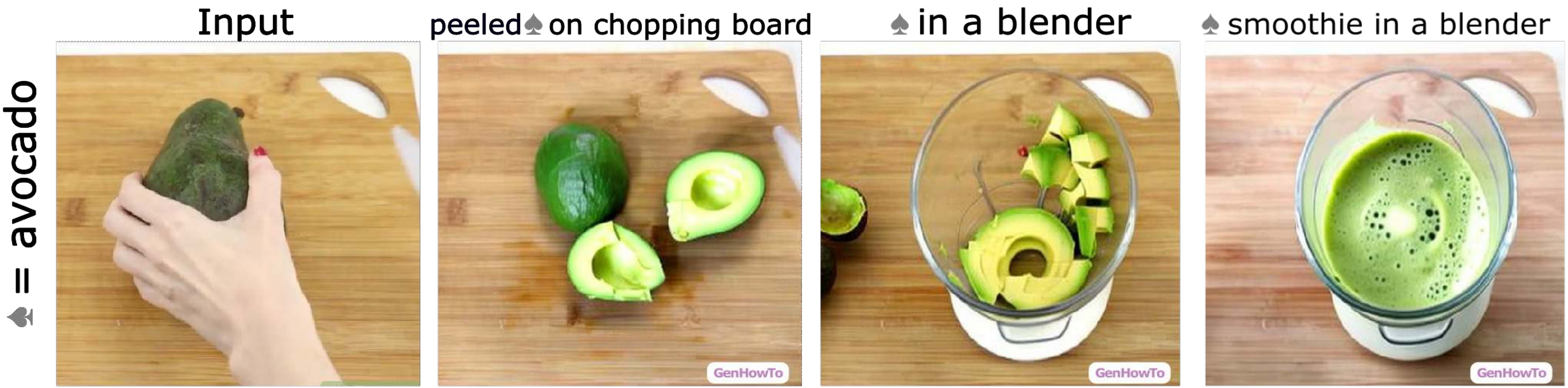
University of BRISTOL

CTU CZECH TECHNICAL UNIVERSITY IN PRAGUE

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

with: Tomas Soucek
Ivan Laptev
Michael Wray
Josef Sivic

- Hands transform objects….



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

GenHowTo…

with: Tomas Soucek Michael Wray
Ivan Laptev Josef Sivic

GenHowTo…

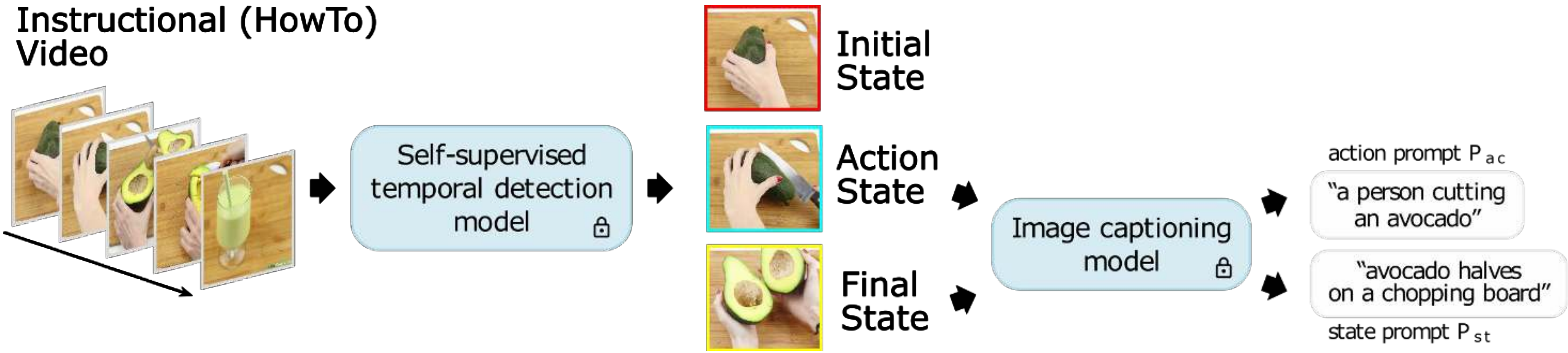
with: Tomas Soucek Michael Wray
Ivan Laptev Josef Sivic

- Two contributions…. Dataset & Method

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

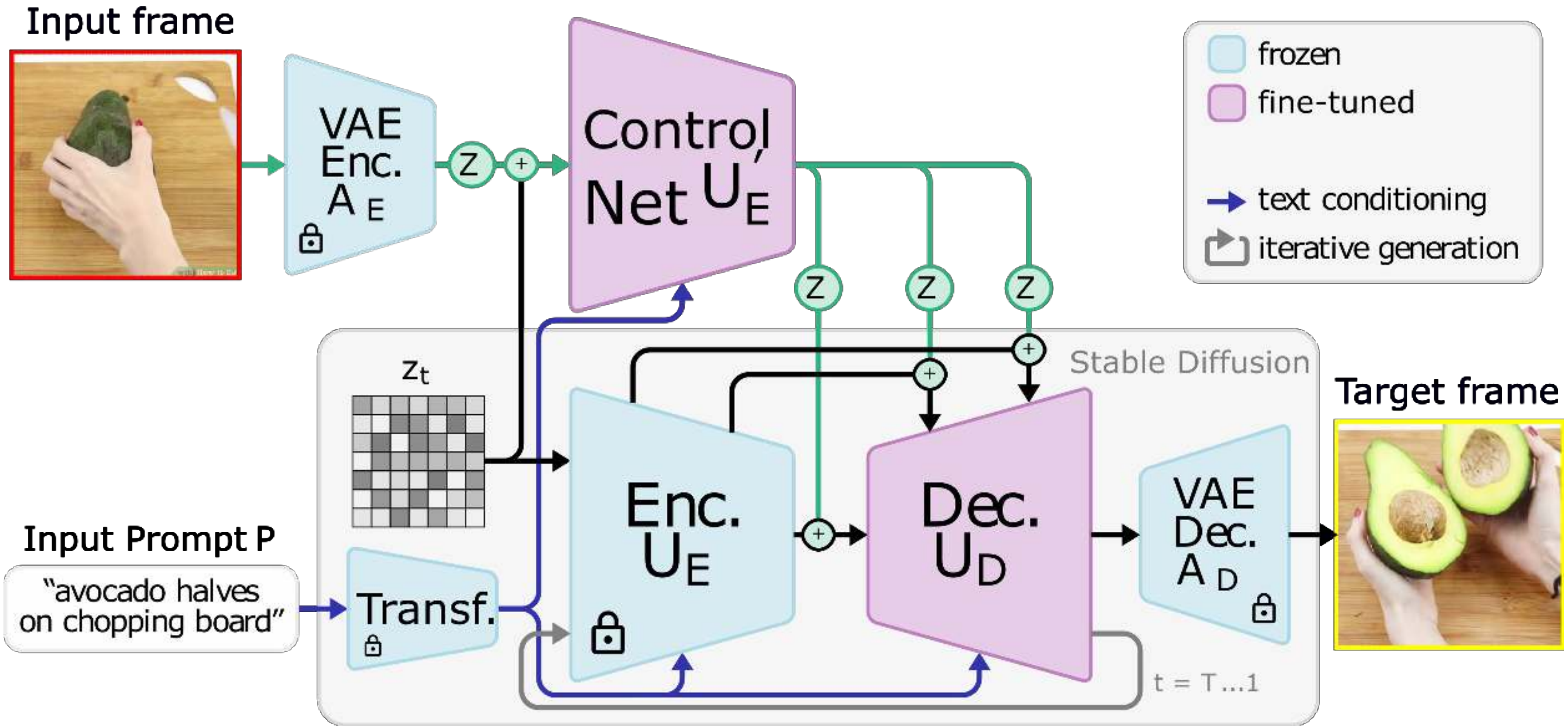
with: Tomas Soucek
Ivan Laptev
Michael Wray
Josef Sivic

- 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.

- 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 Michael Wray
Ivan Laptev Josef Sivic

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

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

	Method	$Acc_{ac}\uparrow$	$Acc_{st}\uparrow$
	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

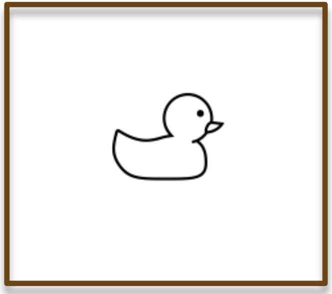
T Soucek et al (2023). GenHowTo: Learning to Generate Actions and State Transformations from Instructional Videos. ArXiv

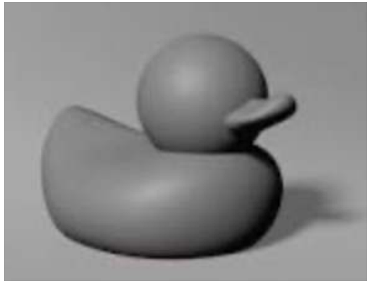
It is more important to understand consequences of actions that to generate smooth motions

Analogous Tasks

Image

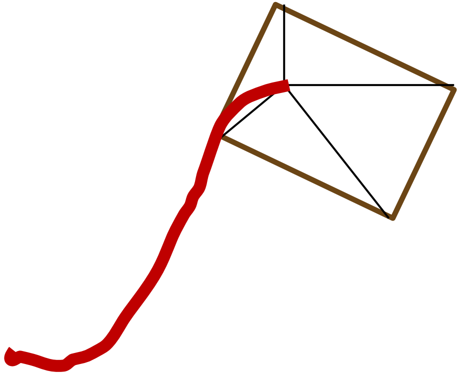
- Image – to – 3D





Video

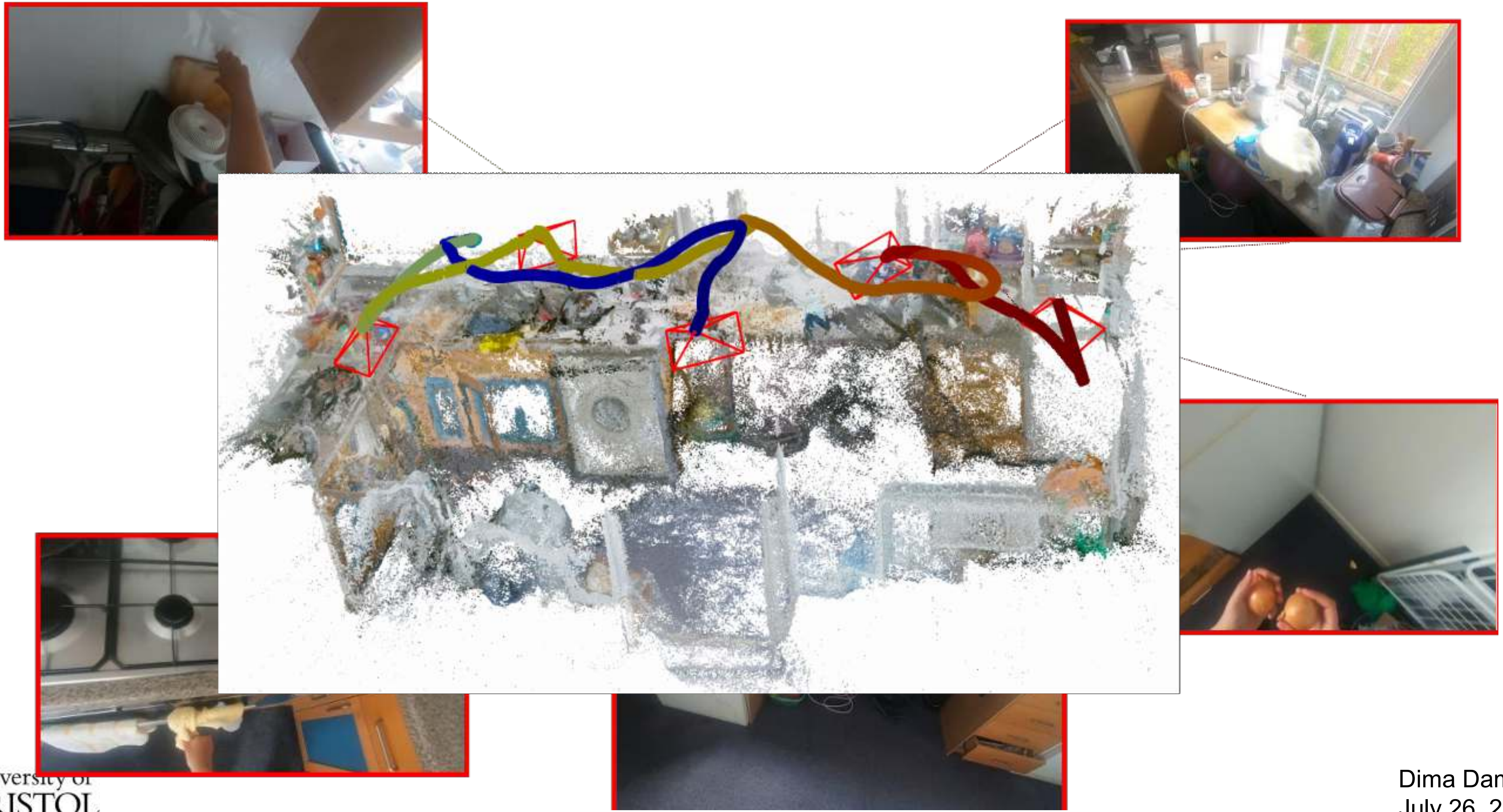
- Video – to - 3D



Almost all video-to-3D focuses on a static scene

EPIC Fields

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



EPIC-KITCHENS

with: Chiara Plizzari Shubham Goel
Toby Perrett Angjoo Kanazawa

Spatial Cognition from Egocentric Video:

Out of Sight, Not Out of Mind

Chiara Plizzari Shubham Goel Toby Perrett Jacob Chalk

Angjoo Kanazawa Dima Damen

http://dimadamen.github.io/OSNOM

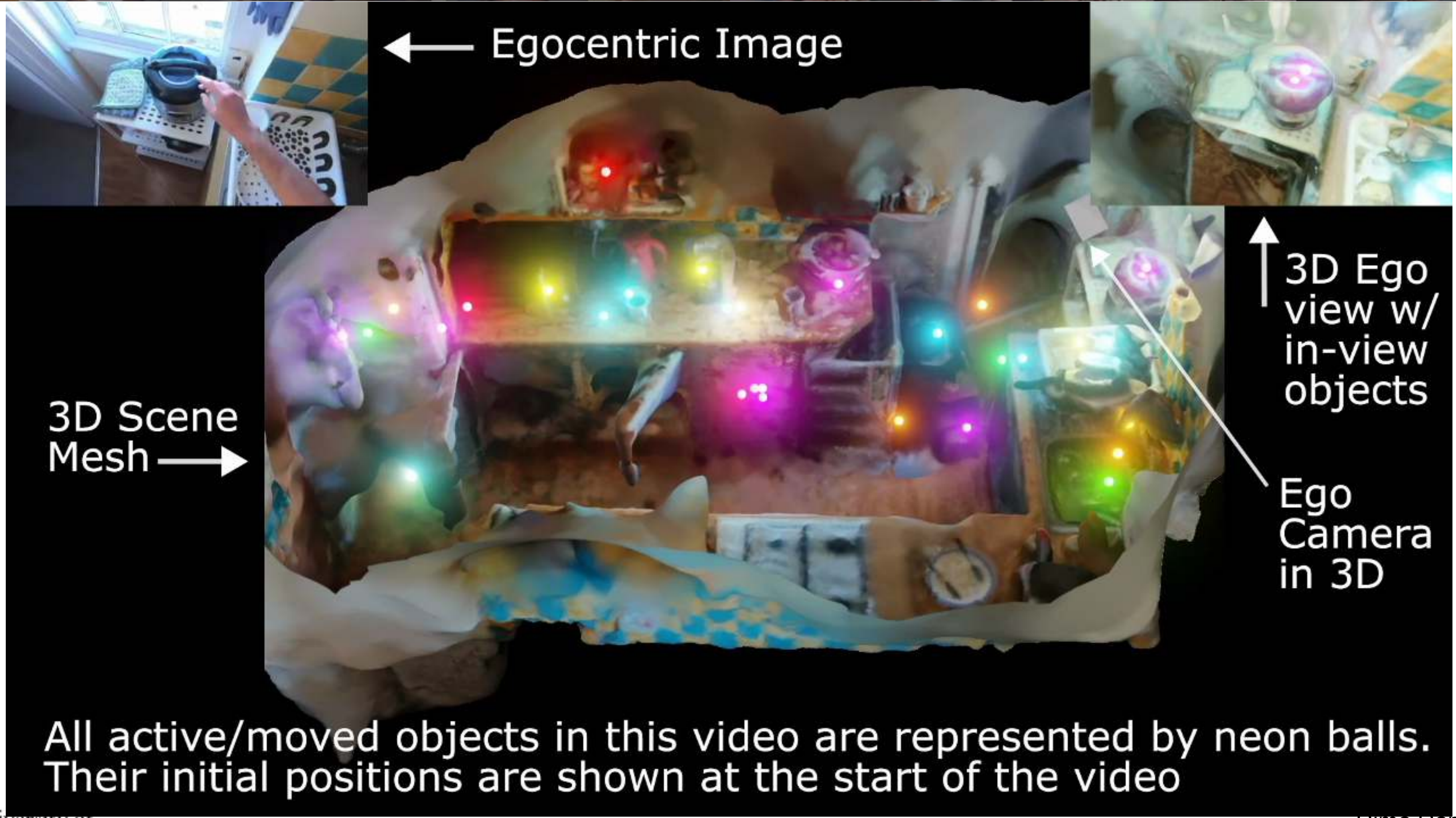
Politecnico di Torino 1859

Berkeley UNIVERSITY OF CALIFORNIA

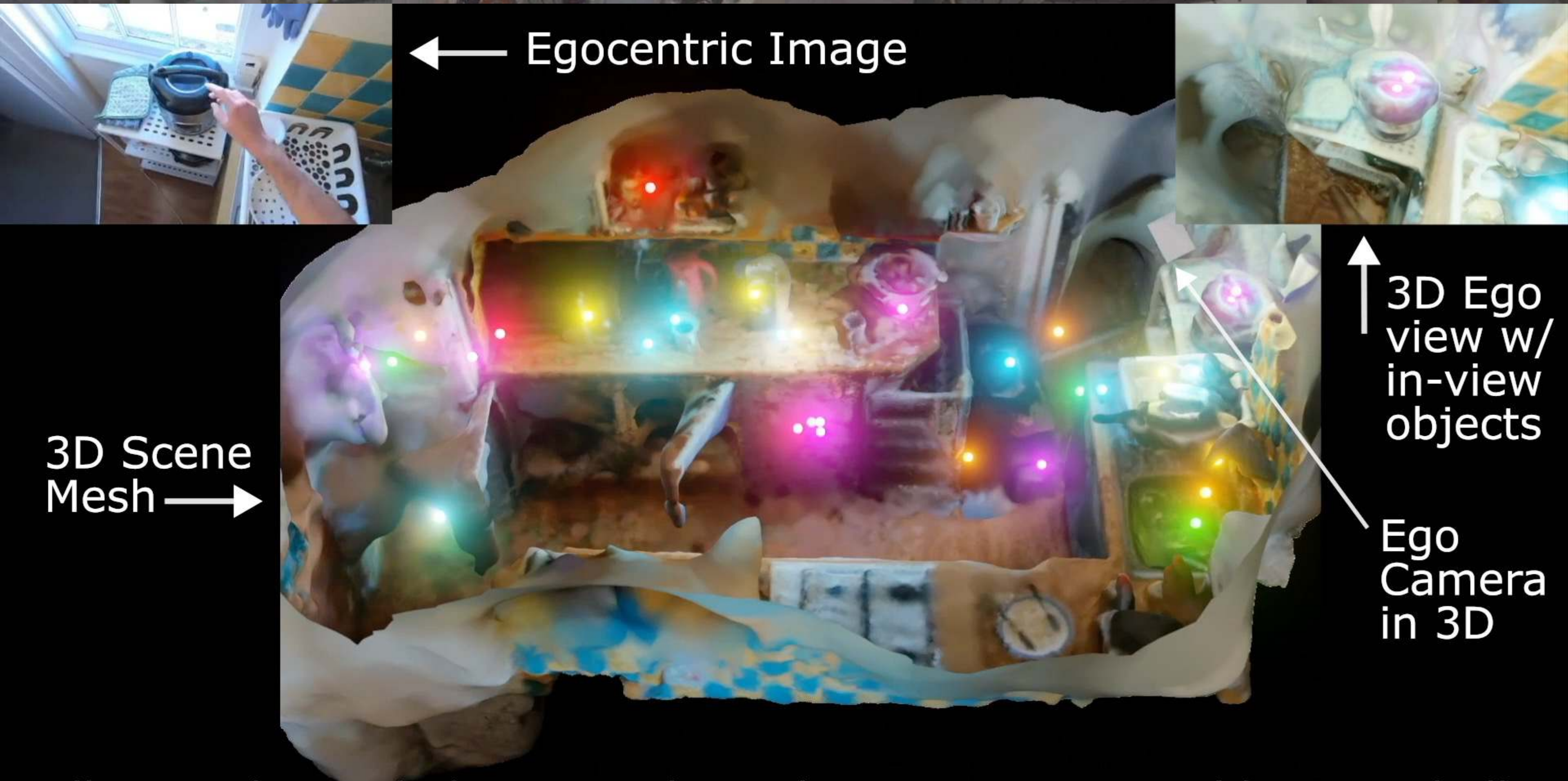
University of BRISTOL

Out of Sight, not Out of Mind

with: Chiara Plizzari Shubham Goel
Toby Perrett Angjoo Kanazawa



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



All active/moved objects in this video are represented by neon balls.
Their initial positions are shown at the start of the video

Non-Analogous Tasks

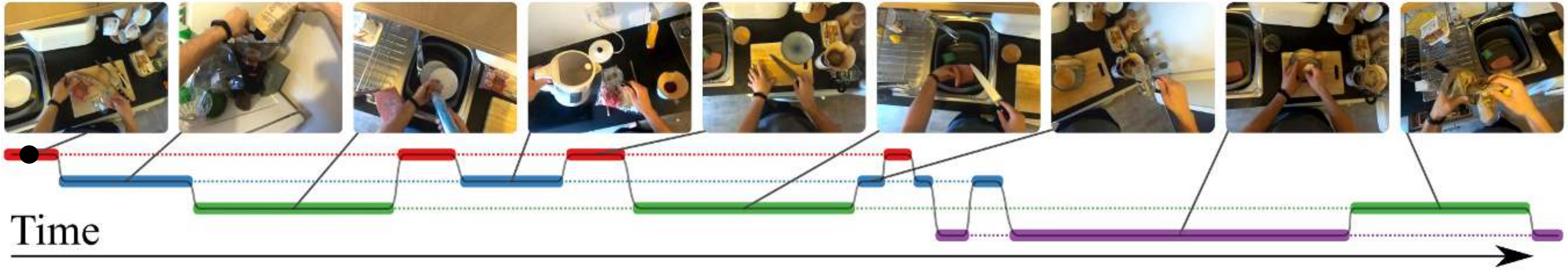
Image

Video

- Understanding Goals in Long Videos

Goals…

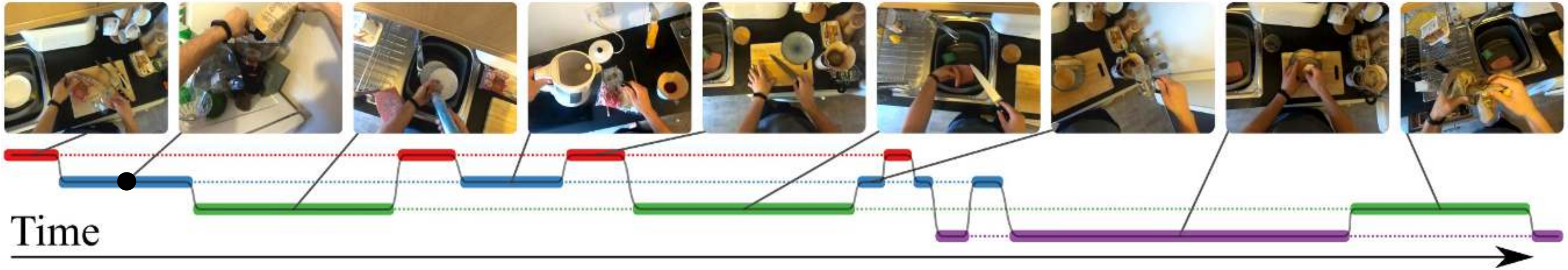
with: Will Price
Carl Vondrick



W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

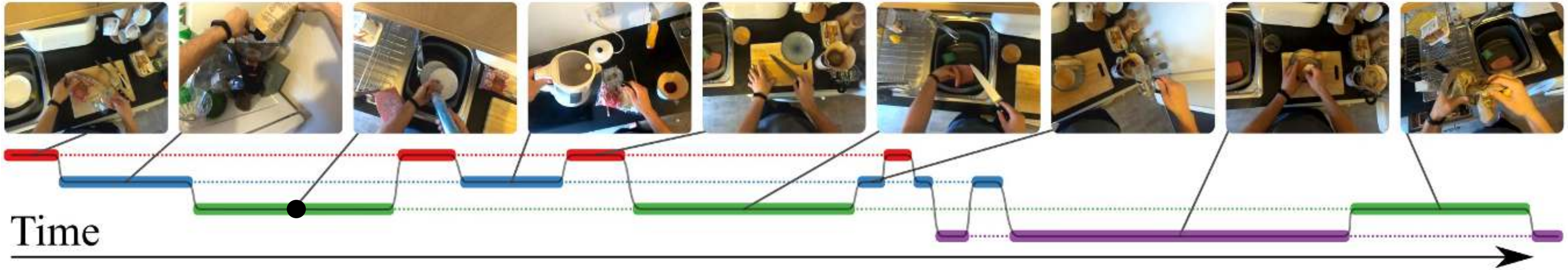
UnweaveNet

with: Will Price
Carl Vondrick



UnweaveNet

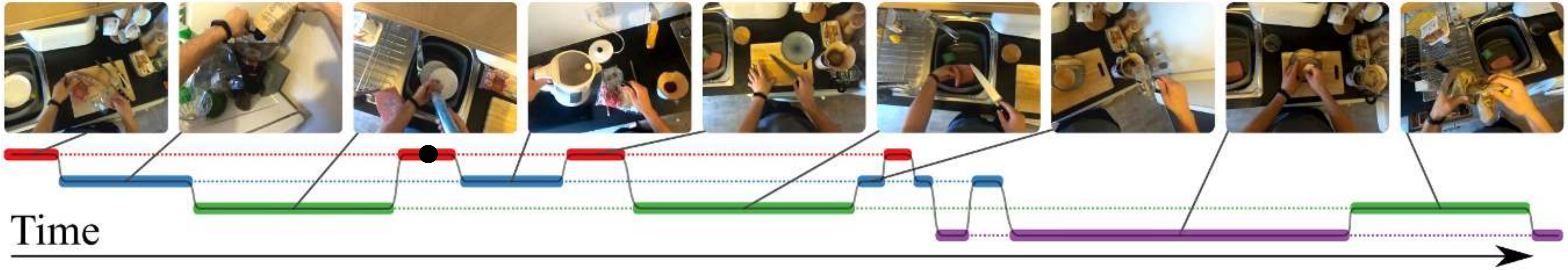
with: Will Price
Carl Vondrick



W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

UnweaveNet

with: Will Price
Carl Vondrick

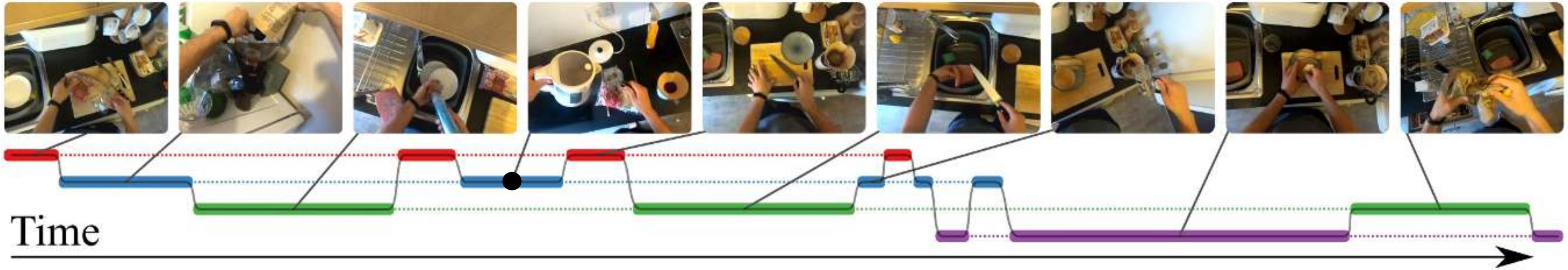


University of BRISTOL

W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

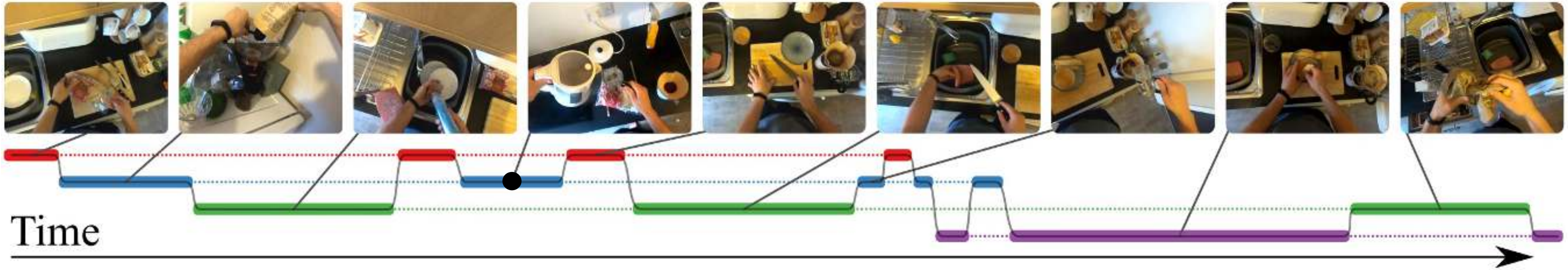
UnweaveNet

with: Will Price
Carl Vondrick



UnweaveNet

with: Will Price
Carl Vondrick



W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

UnweaveNet

with: Will Price
Carl Vondrick

W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

UnweaveNet

with: Will Price
Carl Vondrick

UnweaveNet

with: Will Price
Carl Vondrick

W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

UnweaveNet

with: Will Price
Carl Vondrick

W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

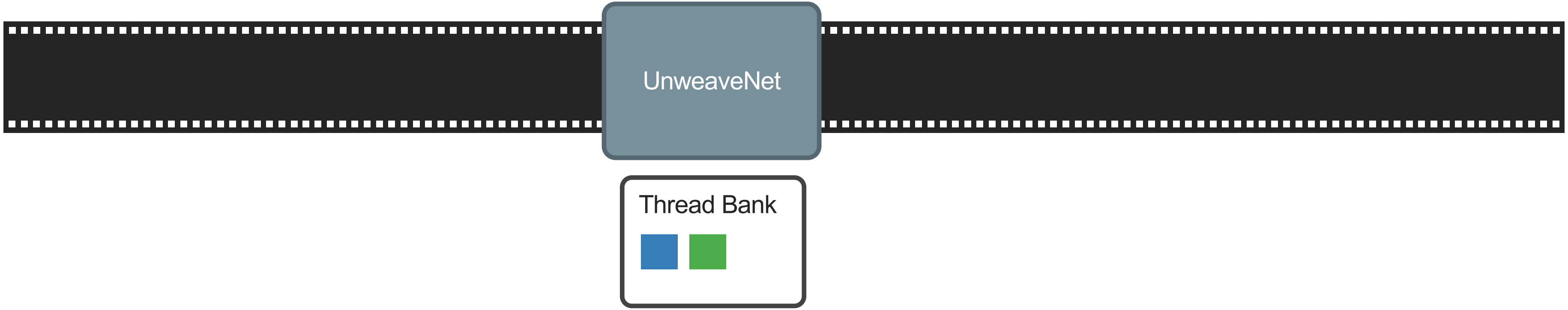
UnweaveNet

with: Will Price
Carl Vondrick

W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

UnweaveNet

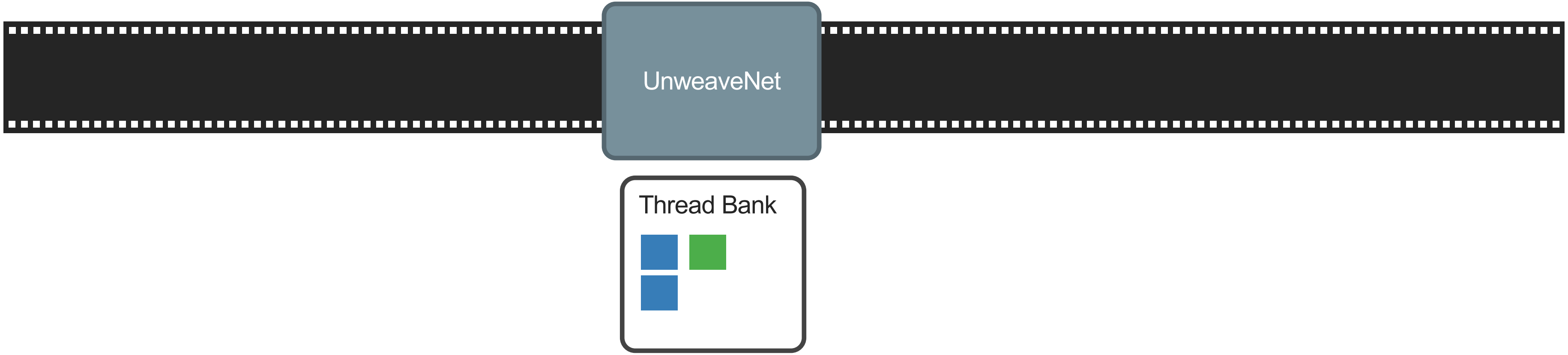
with: Will Price
Carl Vondrick



W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

UnweaveNet

with: Will Price
Carl Vondrick

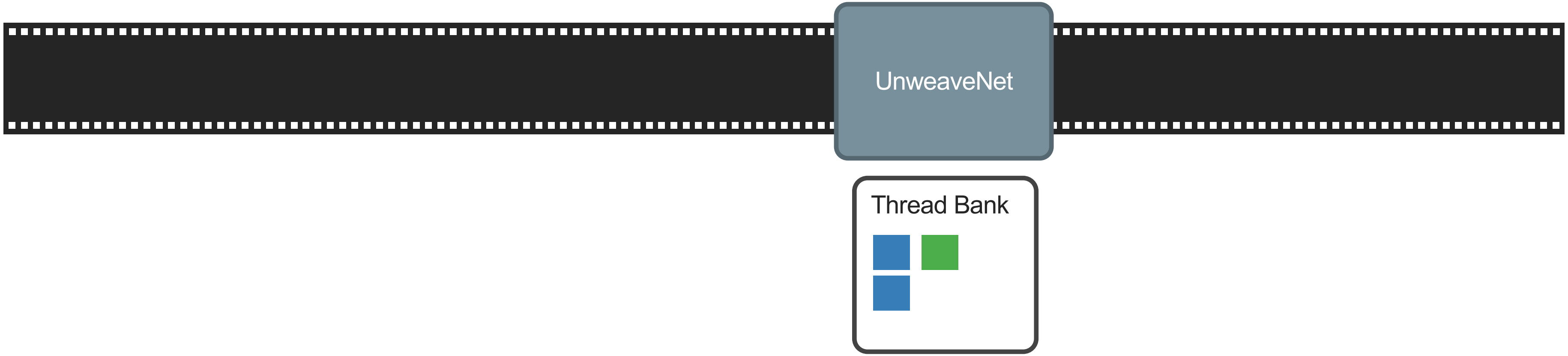


University of BRISTOL

W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

UnweaveNet

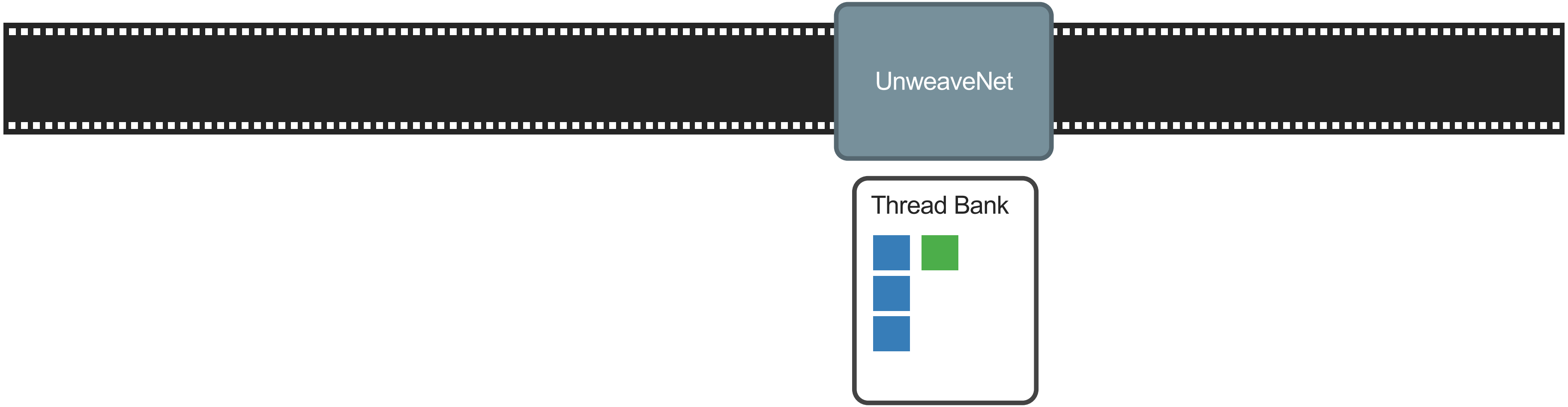
with: Will Price
Carl Vondrick



W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

UnweaveNet

with: Will Price
Carl Vondrick



W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

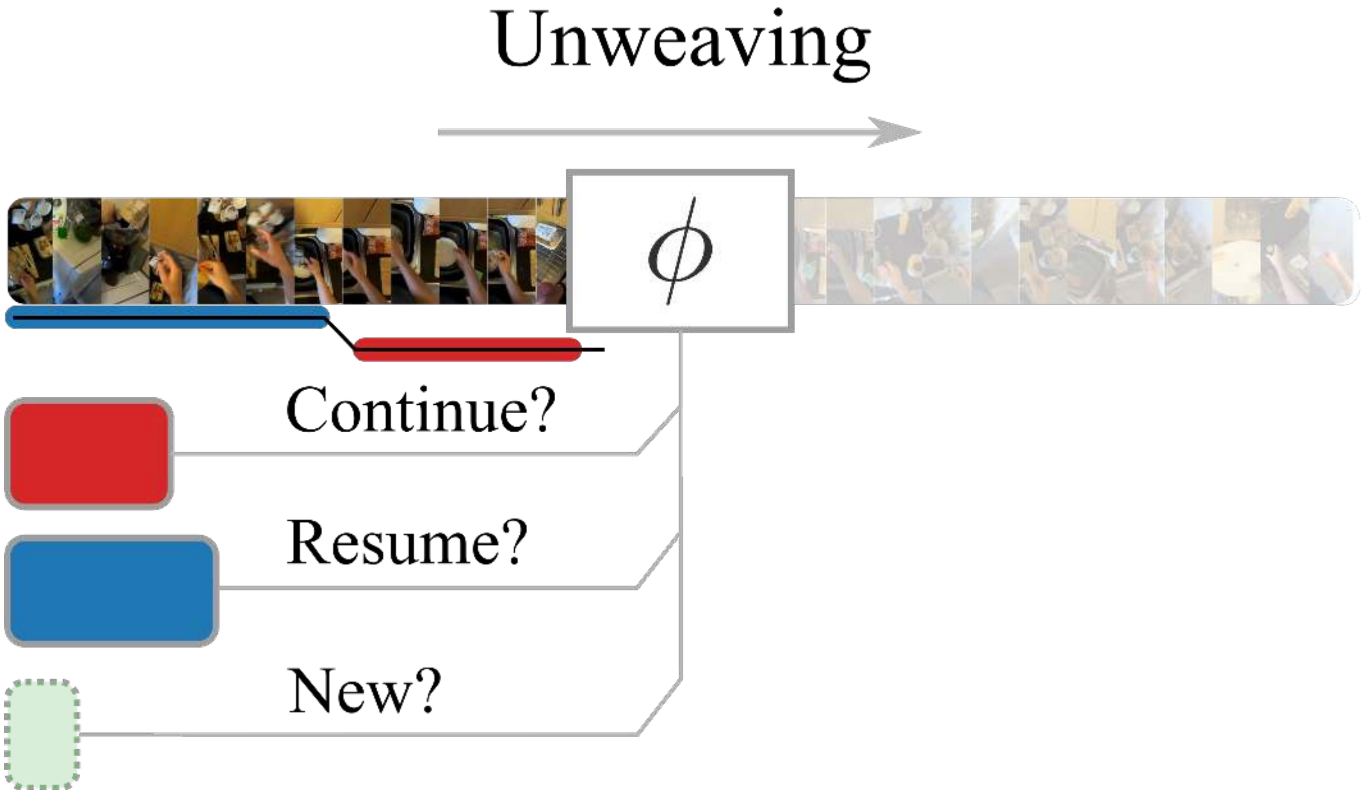
UnweaveNet

with: Will Price
Carl Vondrick

UnweaveNet

with: Will Price
Carl Vondrick

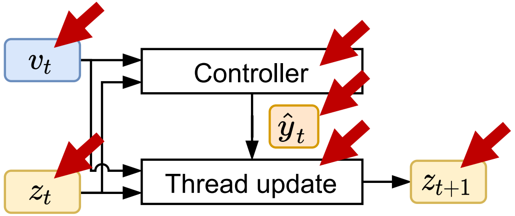
with: Will Price
Carl Vondrick



W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

UnweaveNet

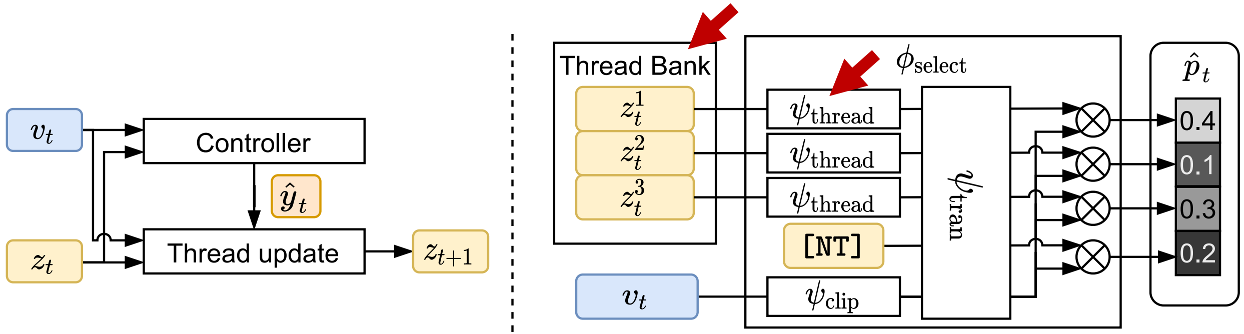
with: Will Price
Carl Vondrick



(a) UnweaveNet Overview

UnweaveNet

with: Will Price
Carl Vondrick



(a) UnweaveNet Overview

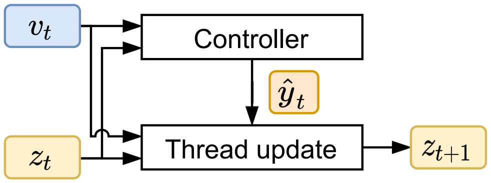
(b) Controller Architecture

Two learnt embeddings

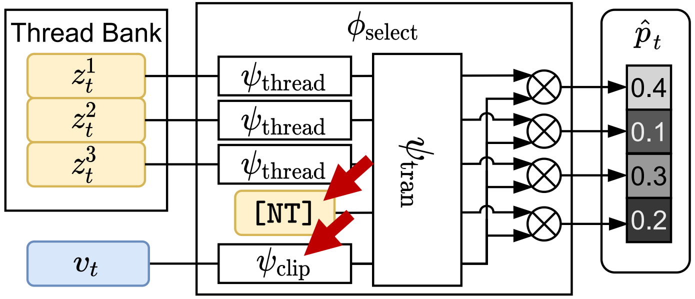
$$\psi_{\text{thread}} : \mathbb{R}^D \rightarrow \mathbb{R}^E$$

UnweaveNet

with: Will Price
Carl Vondrick



(a) UnweaveNet Overview



(b) Controller Architecture

Two learnt embeddings

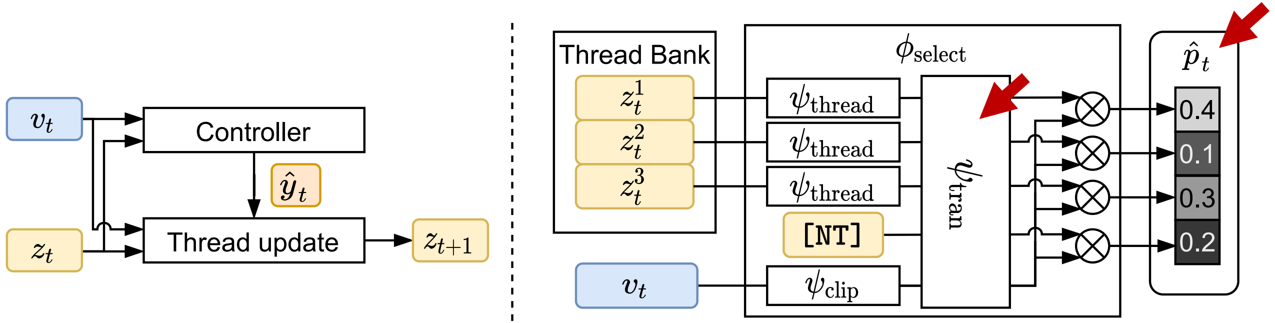
$$\psi_{\text{thread}} : \mathbb{R}^D \rightarrow \mathbb{R}^E$$

$$\psi_{\text{clip}} : \mathbb{R}^C \rightarrow \mathbb{R}^E$$

Learnt Encoding $\quad [\mathtt{NT}] \in \mathbb{R}^E$

UnweaveNet

with: Will Price
Carl Vondrick



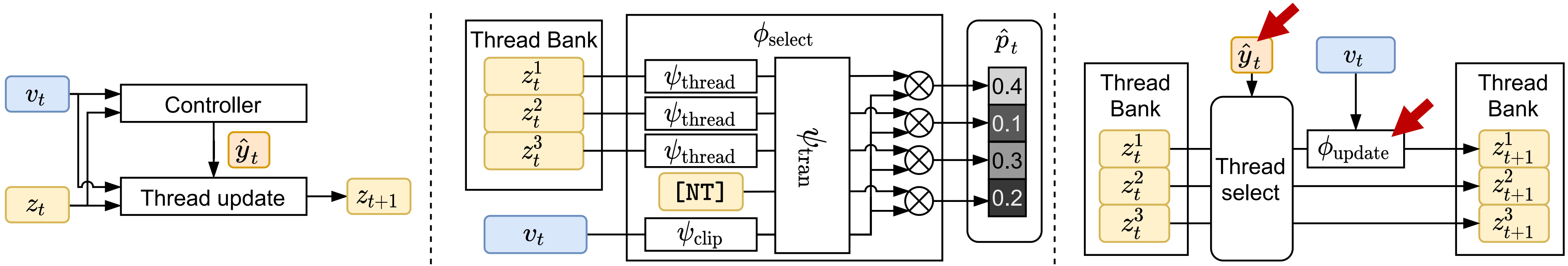
(a) UnweaveNet Overview

(b) Controller Architecture

W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

UnweaveNet

with: Will Price
Carl Vondrick



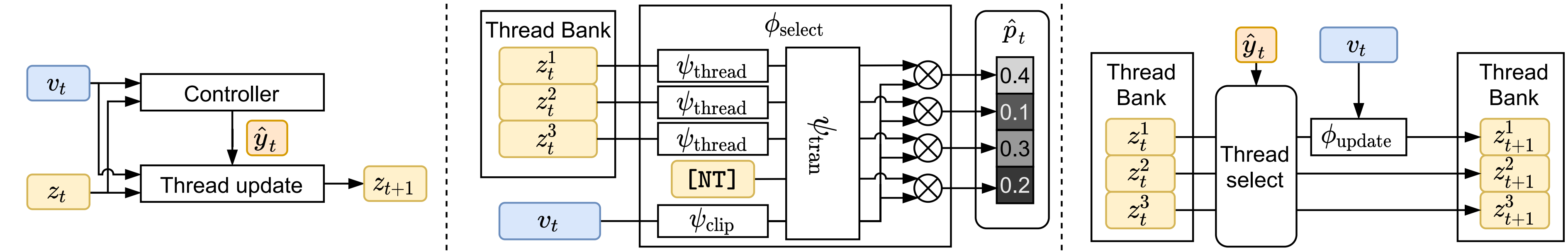
(a) UnweaveNet Overview

(b) Controller Architecture

(c) Thread bank update

UnweaveNet

with: Will Price
Carl Vondrick



(a) UnweaveNet Overview (b) Controller Architecture (c) Thread bank update

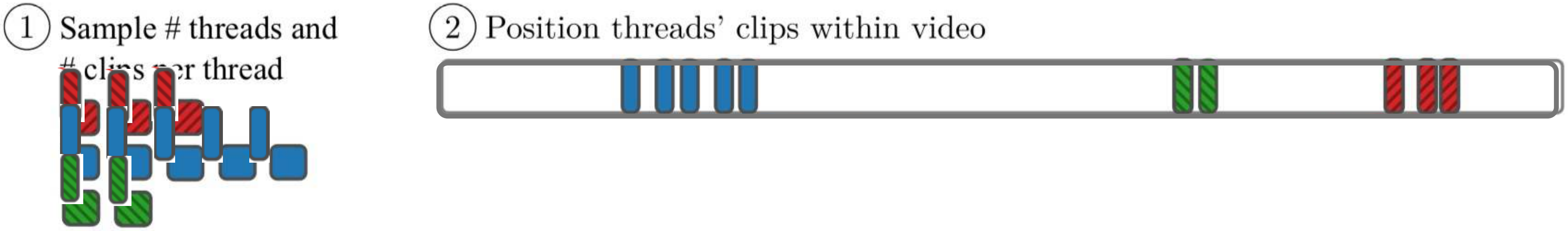
- Trained **end-to-end** including the backbone for clip features
- decisions made by ϕ_{select} are supervised using **teacher forcing**
  - at each time step, z_t is populated according to the ground-truth assignments $y_{1:t-1}$
  - A loss is then imposed on the output p_t given the correct decision y_t with focal hyperparameter γ due to the imbalance in decisions

W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

UnweaveNet

with: Will Price
Carl Vondrick

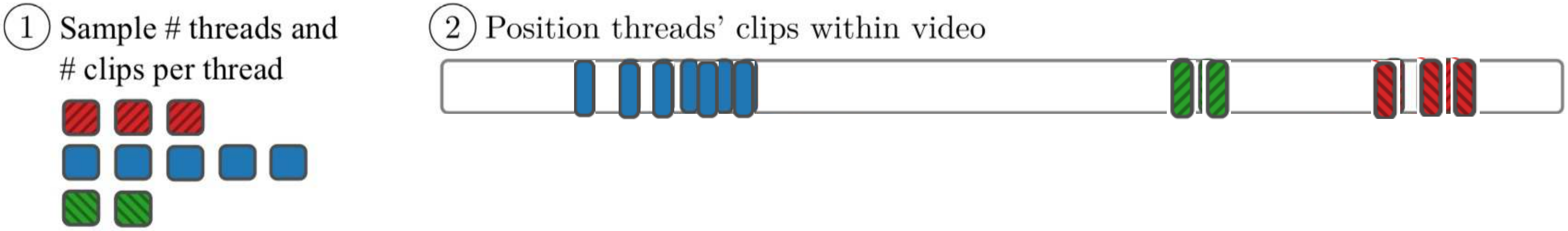
- We propose self-supervised pretraining for UnweaveNet that samples threads from different parts of a long video and synthetically forms woven activity stories.



UnweaveNet

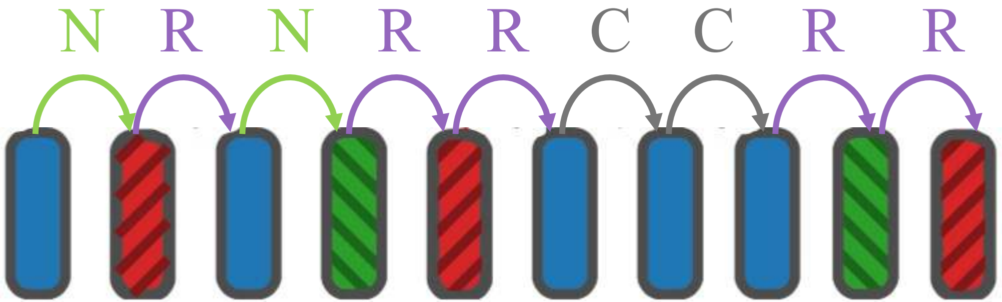
with: Will Price
Carl Vondrick

- We propose self-supervised pretraining for UnweaveNet that samples threads from different parts of a long video and synthetically forms woven activity stories.



with: Will Price
Carl Vondrick

- 1000s of synthetic stories from training set.



UnweaveNet

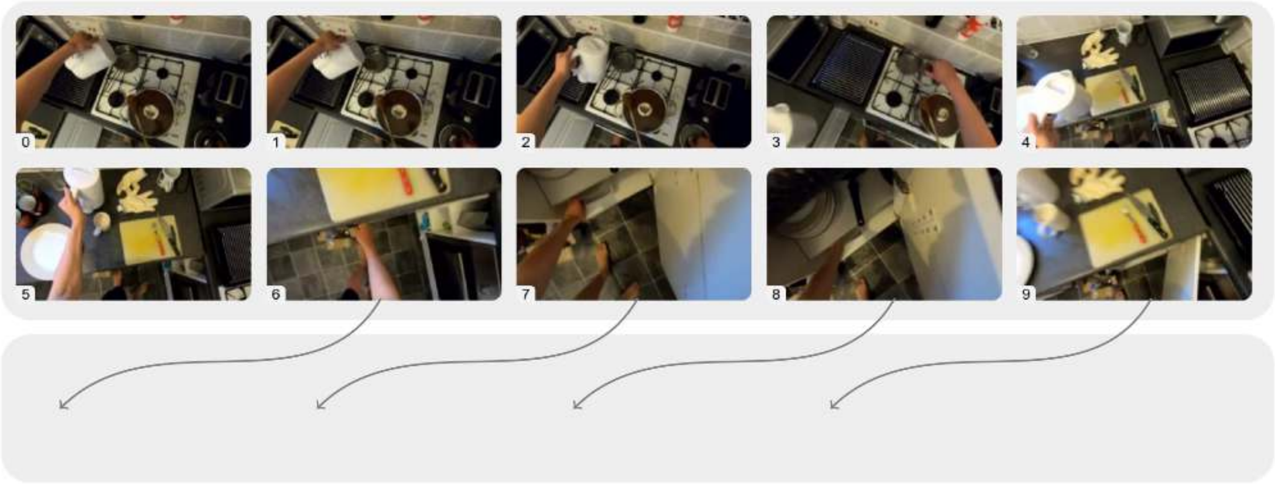
with: Will Price
Carl Vondrick

- Labelled Sequences

Story Annotator

Story ID: c7c7f261-e5c7-4641-b0cf-b2b4e8c8ef3d
Video ID: P06_103
Start time: 7567
Splits: train+val
Author:

Skip (space) Next (enter)



	# Threads		
Split	1	2	3
Train	718	201	32
Val	211	94	46
Test	50	50	50
Total	979	345	128

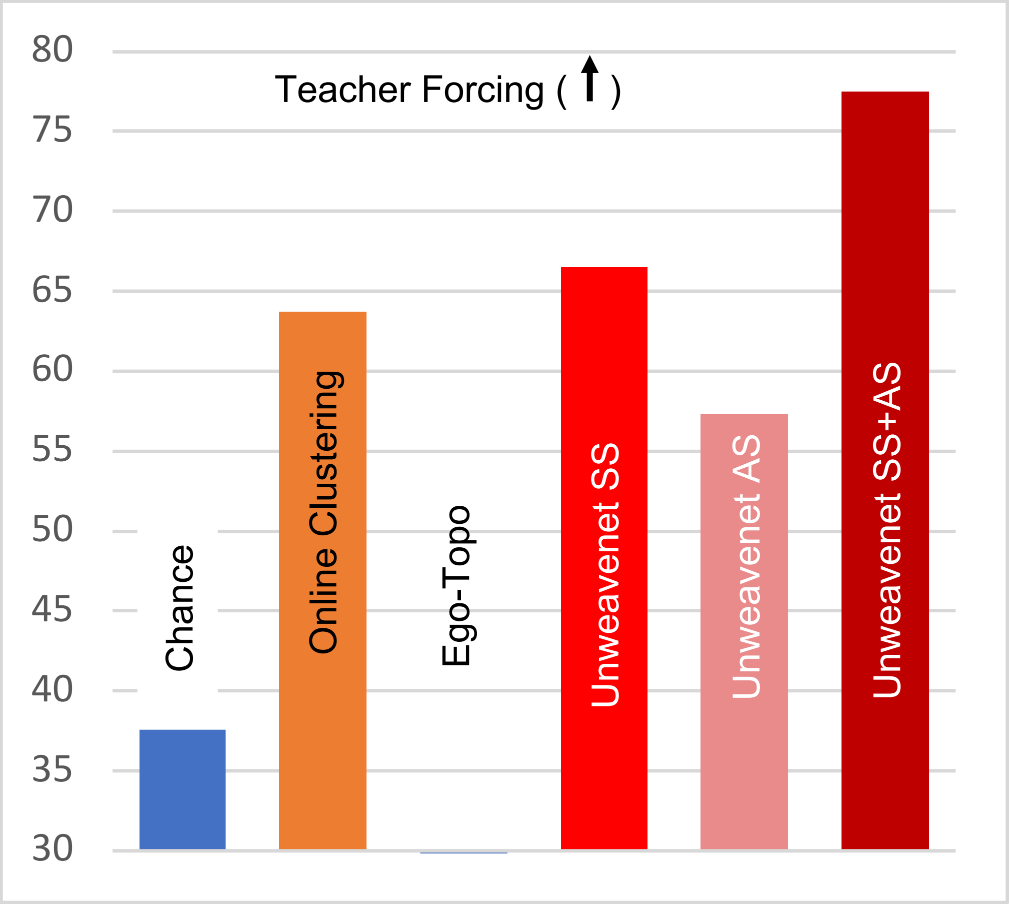
Table 1. EPIC-KITCHENS activity-story dataset by # of threads.

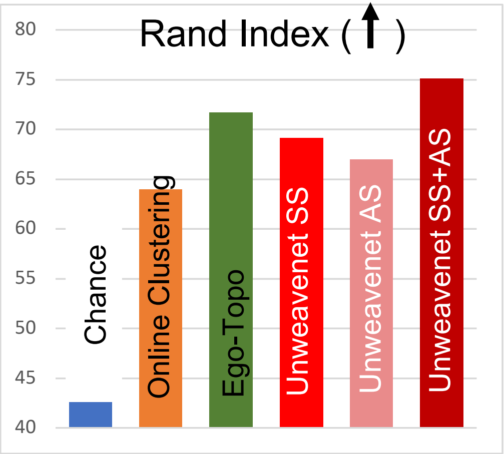
University of BRISTOL

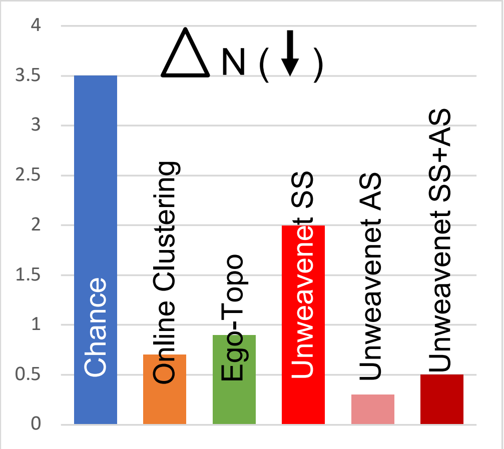
W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

UnweaveNet

with: Will Price
Carl Vondrick







University of BRISTOL

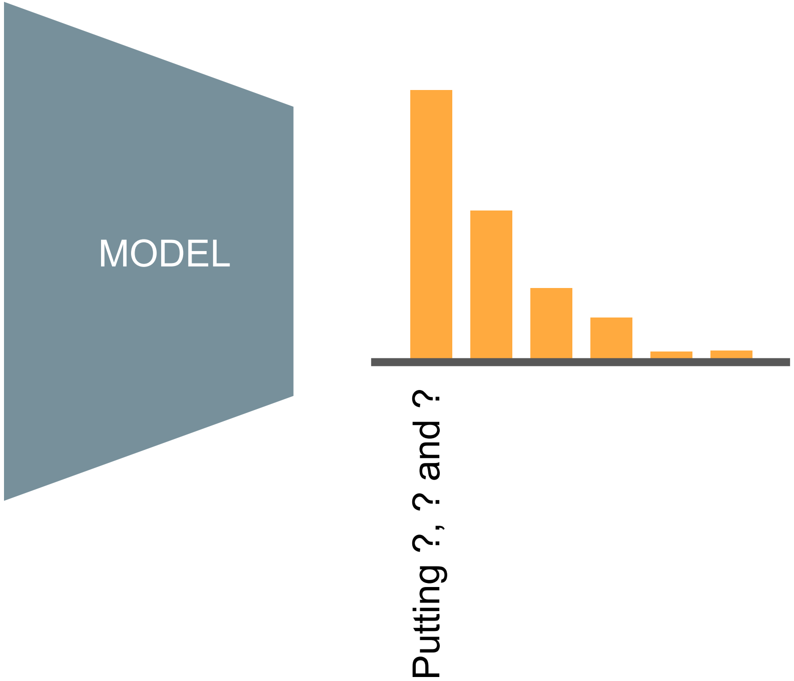
W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

VU - An Egocentric Perspective

Explainable?

Frame Attributions in Video Models

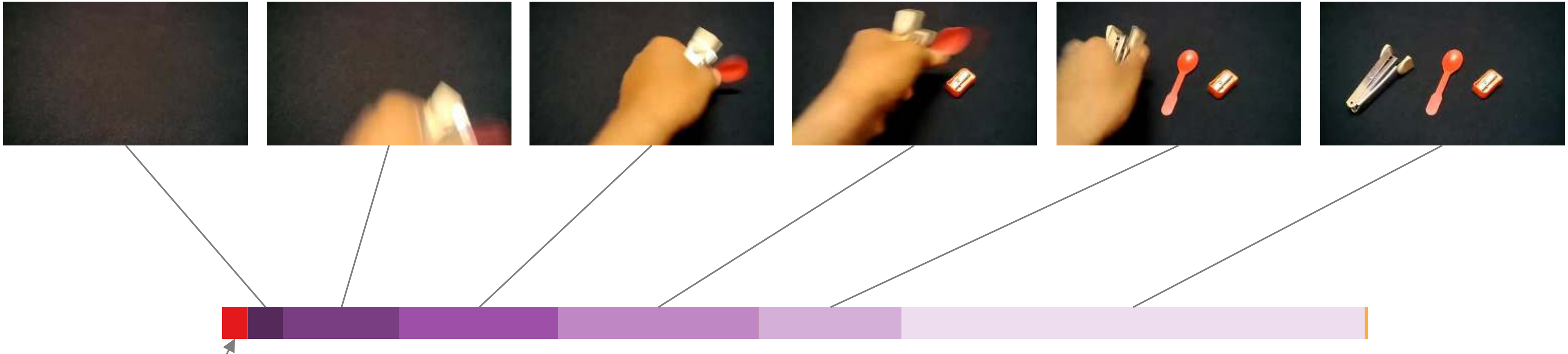
with: Will Price



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

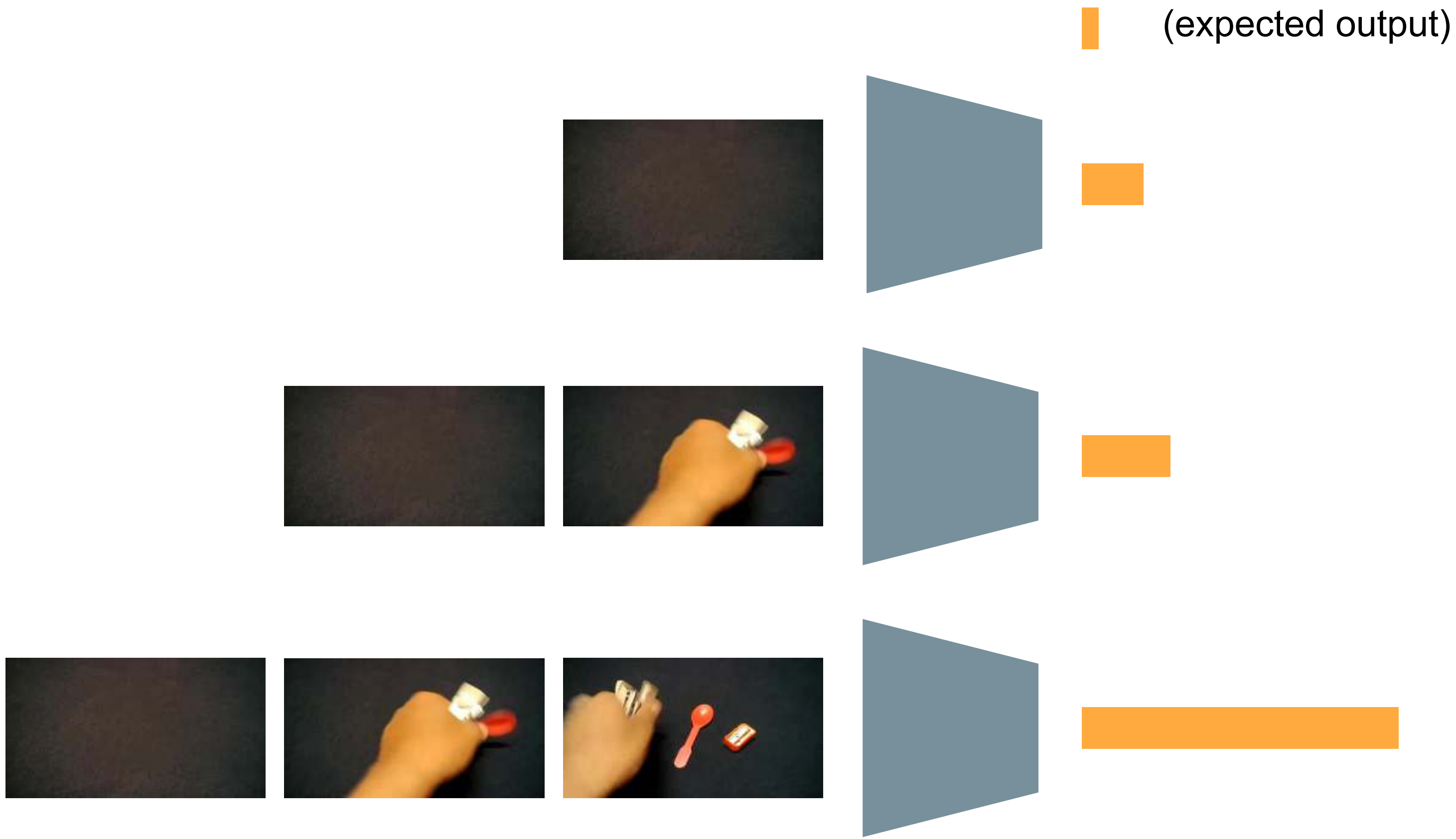


Expected output
(Prior probability for
classification model)

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

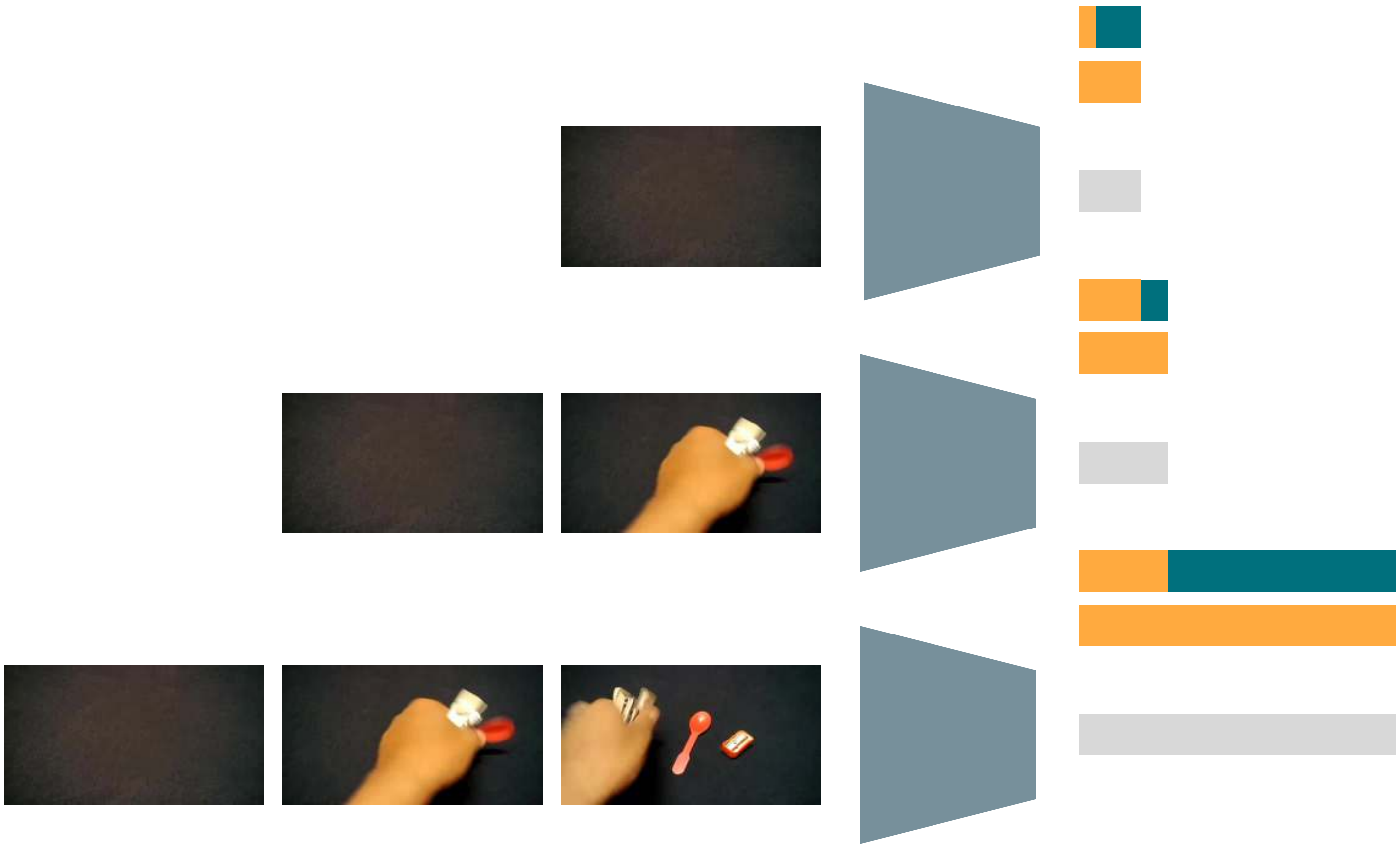
Frame Attributions in Video Models

with: Will Price



Frame Attributions in Video Models

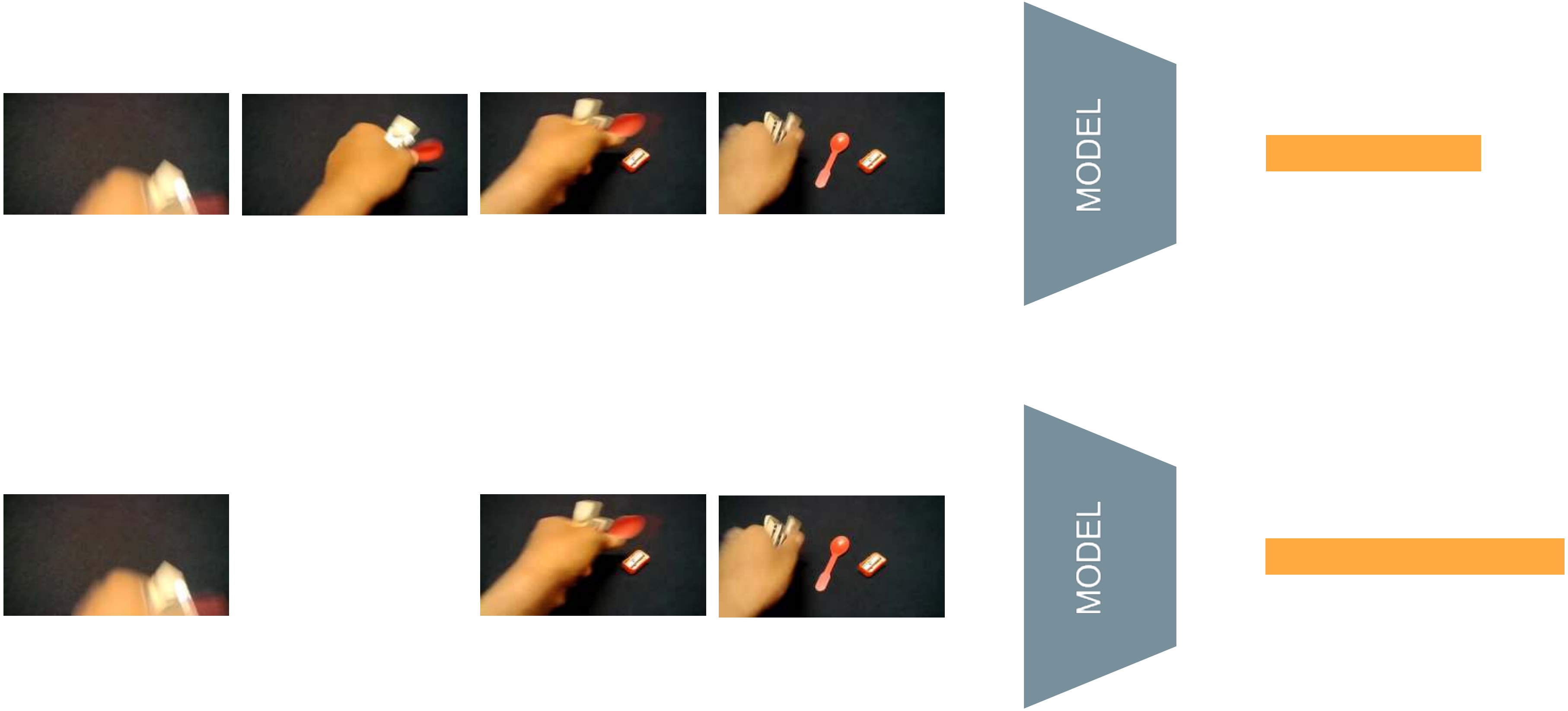
with: Will Price



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

Frame Attributions in Video Models

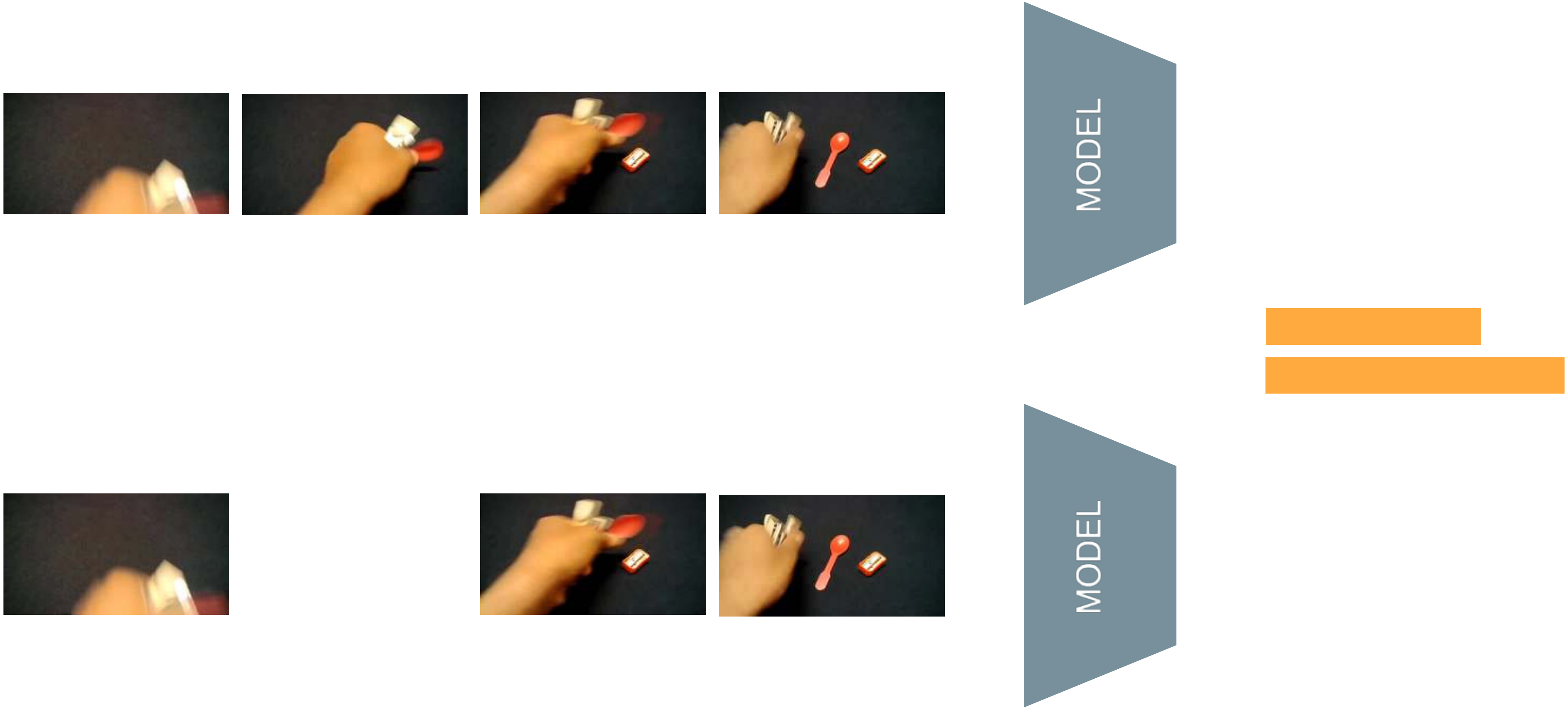
with: Will Price



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

Frame Attributions in Video Models

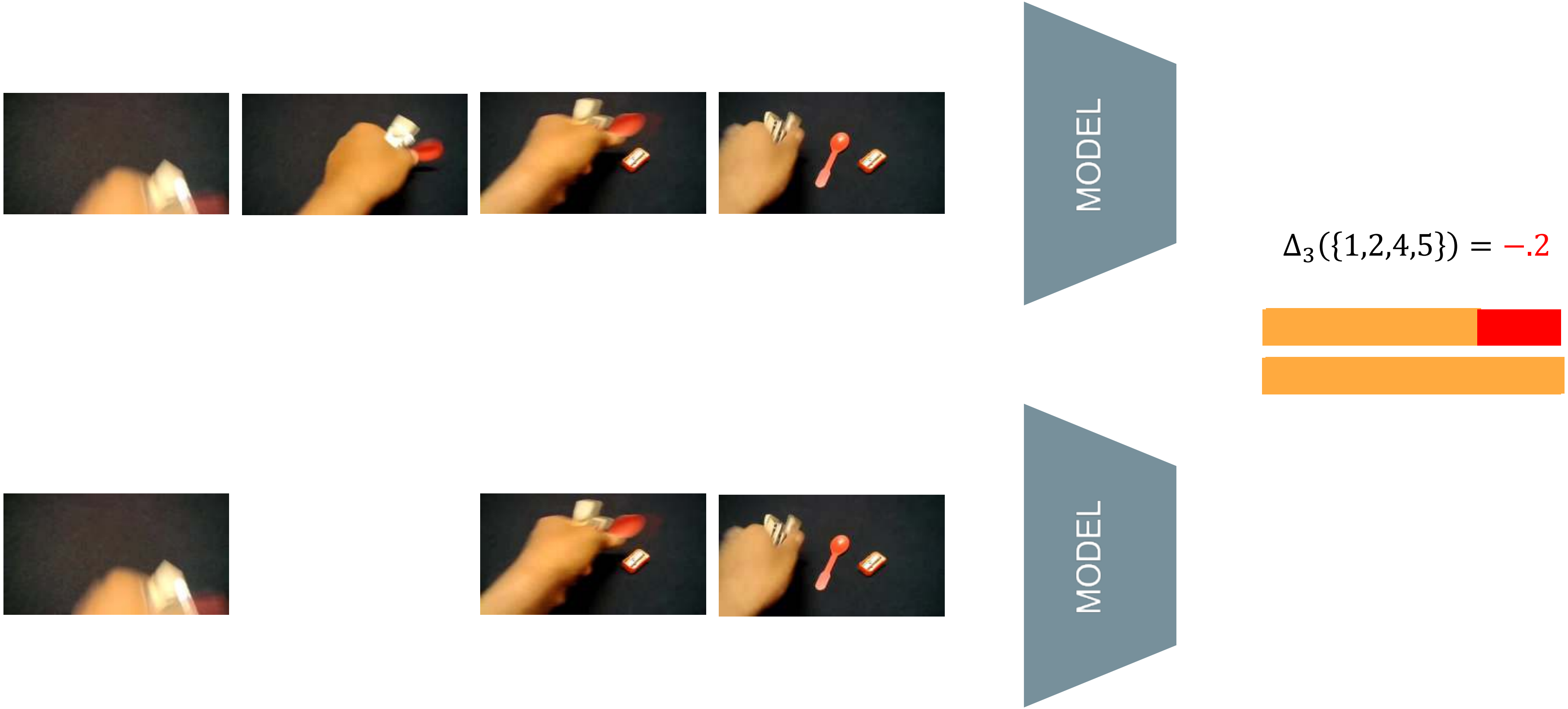
with: Will Price



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

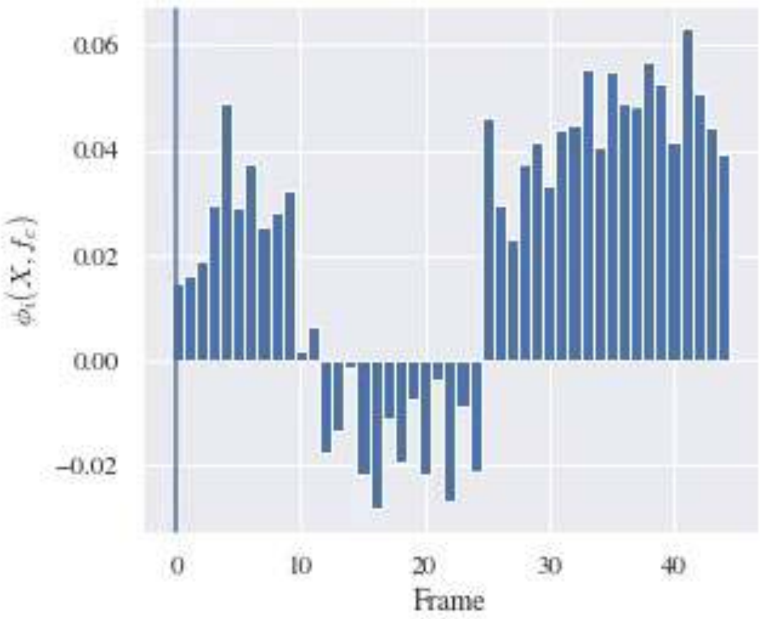


$$\Delta_3(\{1,2,4,5\}) = -.2$$

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



Showing that something is empty

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: open

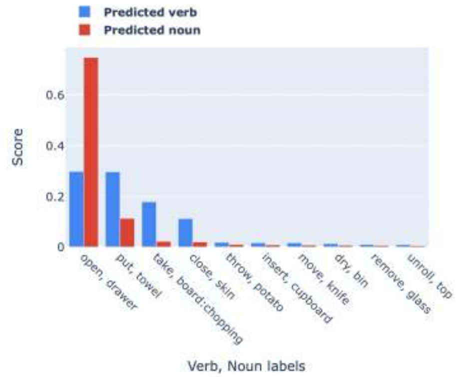
Select a noun: drawer

Select a video: P01_103_84

Select number of frames: 1 2 3 4 5 6 7 8

Original Video:

Model Predictions



ESV Predictions

Selected frame: 2

Selected Verb: 3, Selected Noun: 8, Video P01_103_84

University of BRISTOL

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

with: Will Price
Tom Stark

ESVs Dashboard for Epic

Select a verb

cut

Select a noun

tomato

Select a video

P01_17_126

Select number of frames

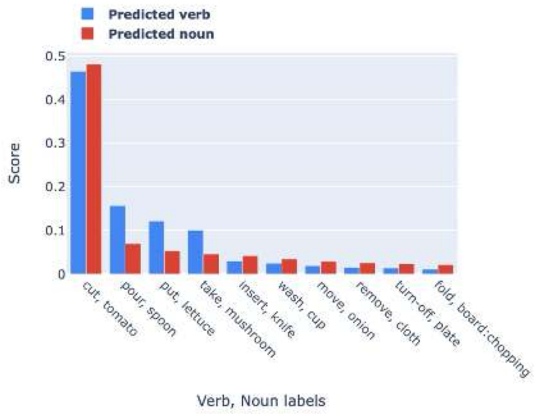
1 2 3 4 5 6 7 8

Original Video:

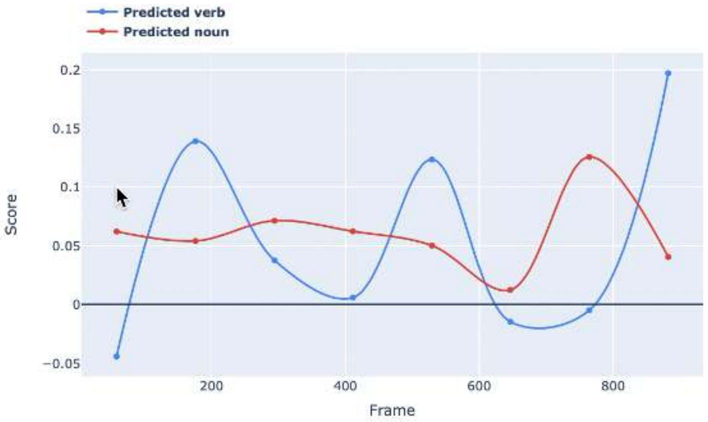
Selected frame: 529

Selected Verb: 7, Selected Noun: 43, Video P01_17_126

Model Predictions



ESV Predictions



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

Summary Wordle

ranking
pixels
models
sampling
streams
temporal
Labels
3D Conv
transformers
Video
bias
semantic
source
Multi-modal
Skill
Tasks
Dataset
Goal
Self-supervision
Domain Adaptation
Frame Attribution

University of BRISTOL

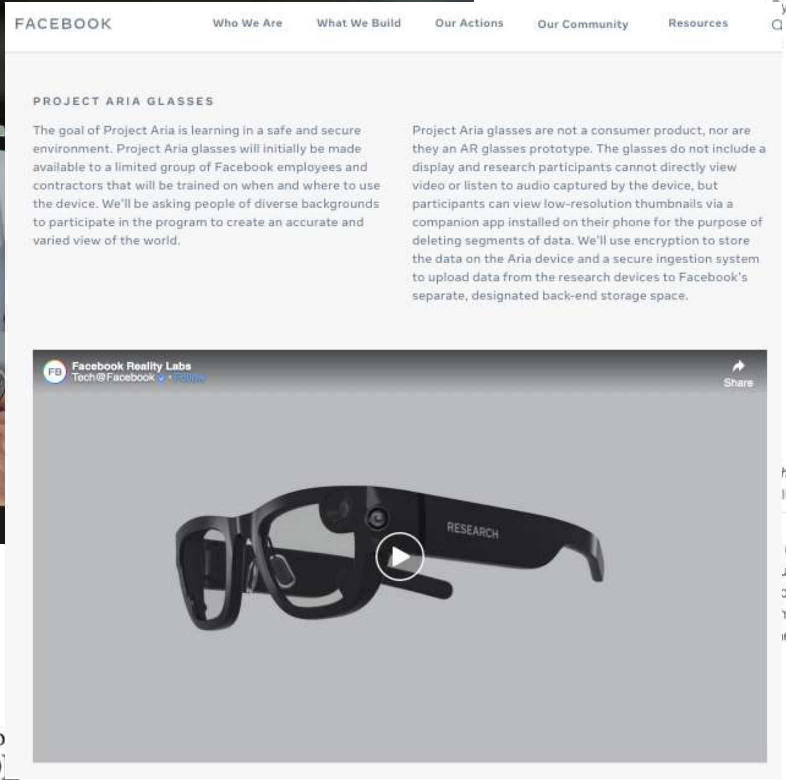
and many more…

Video Understanding

An Egocentric Perspective

University of Bristol

The future is here…



FACEBOOK

Who We Are | What We Build | Our Actions | Our Community | Resources

PROJECT ARIA GLASSES

The goal of Project Aria is learning in a safe and secure environment. Project Aria glasses will initially be made available to a limited group of Facebook employees and contractors that will be trained on when and where to use the device. We'll be asking people of diverse backgrounds to participate in the program to create an accurate and varied view of the world.

Project Aria glasses are not a consumer product, nor are they an AR glasses prototype. The glasses do not include a display and research participants cannot directly view video or listen to audio captured by the device, but participants can view low-resolution thumbnails via a companion app installed on their phone for the purpose of deleting segments of data. We'll use encryption to store the data on the Aria device and a secure ingestion system to upload data from the research devices to Facebook's separate, designated back-end storage space.

Samsung patent application reveals augmented reality headset design

It comes as the Gear VR slowly fades away

The future is here…

Let's start with a show of hands…

Hands-Up if you are ready to wear a head-mounted or glass-mounted camera…

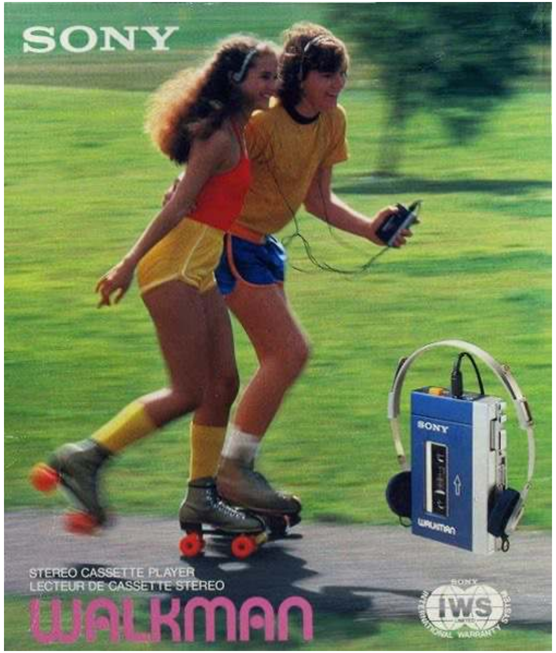
Hands-Up if this is NOT the future…

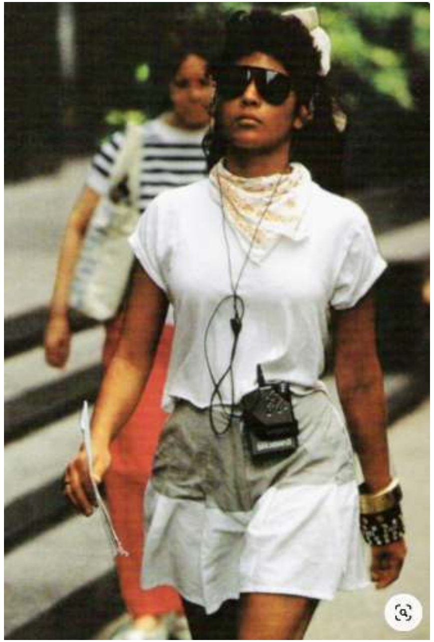
- A world of isolated individuals….
- Dangerous for crossing the road…
- Mind-altering…

45 years ago…

https://humanprogress.org/when-the-walkman-was-banned/

45 years ago…


SONY
SONY
WALKMAN
STEREO CASSETTE PLAYER
LECTEUR DE CASSETTE STEREO
WALKMAN
IWS




PRINCESS

How did we get here?

1907

1970s

..

1987

2004
GoPro

How did we get here?

https://www.youtube.com/watch?v=-wN2p4KkO2A

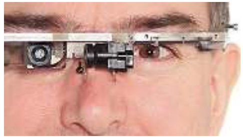
How did we get here?

1907

1970s

1987

1995



1999
Eyetrap

2004
GoPro

2013
Google
Glass

2024

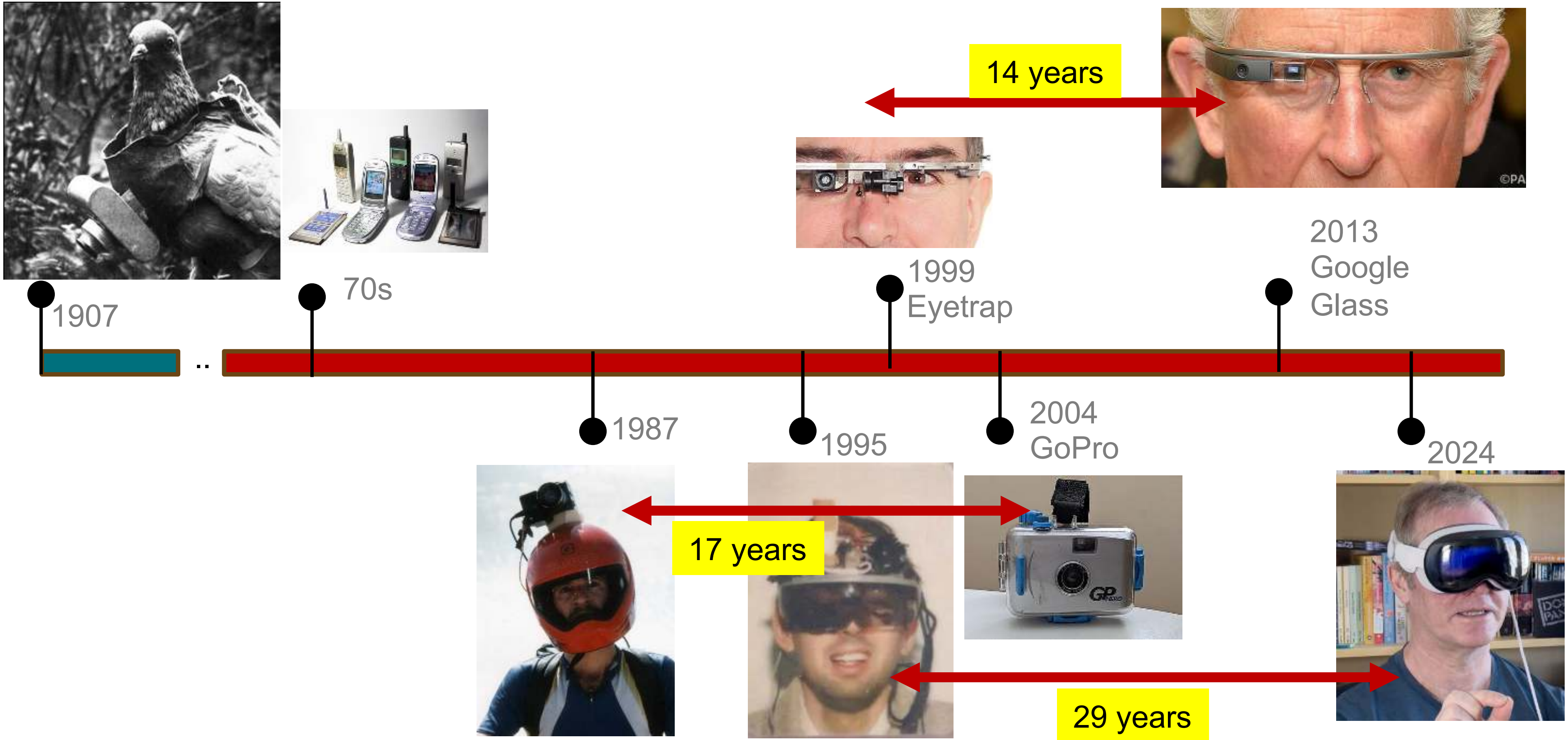
How did we get here?

https://www.youtube.com/watch?v=fCco6FMCRmk

University of BRISTOL

Dima Damen
July 26, 2024

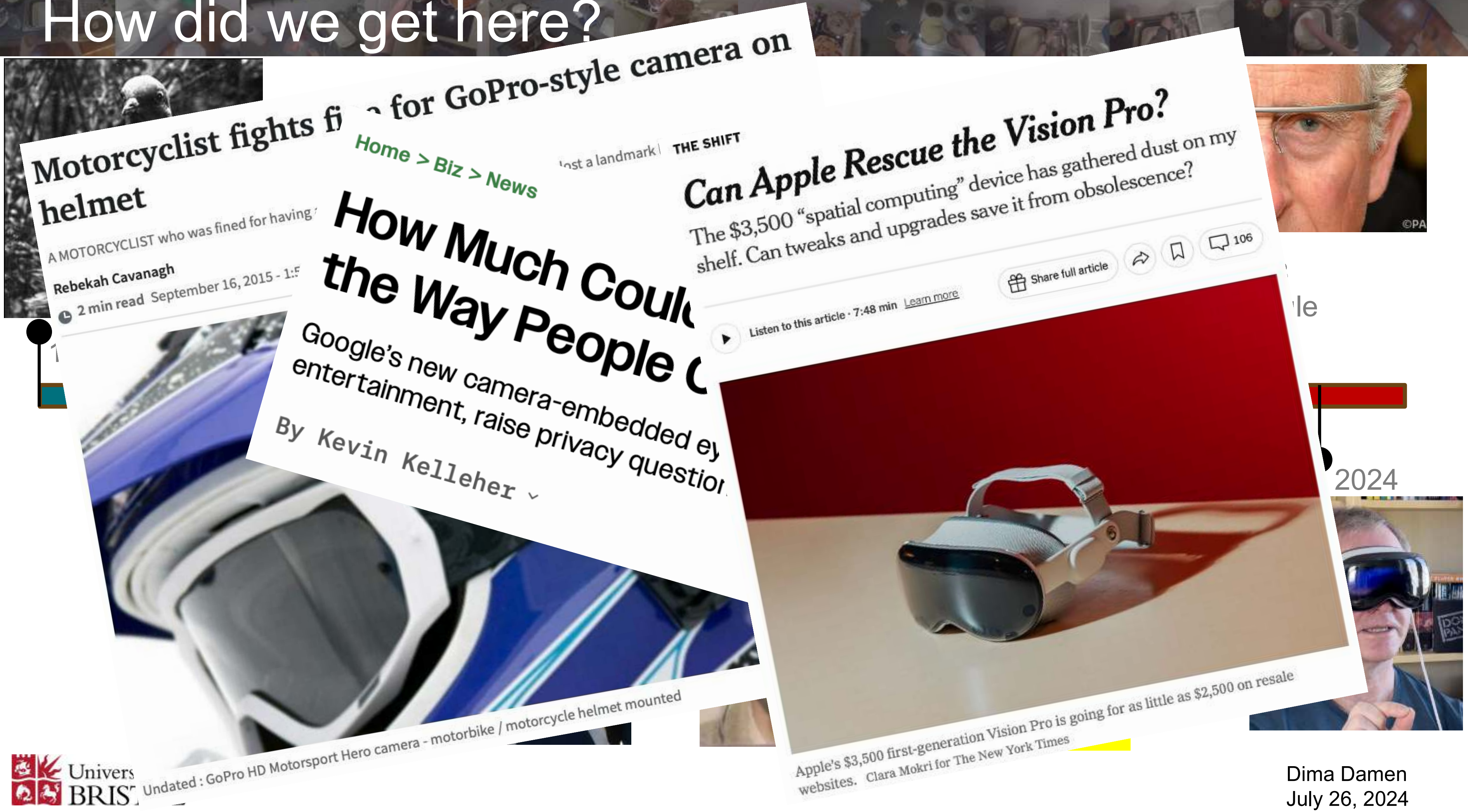
How did we get here?



University of BRISTOL

Dima Damen
July 26, 2024

How did we get here?



Dima Damen
July 26, 2024

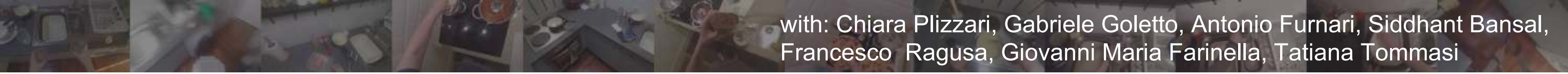
An Outlook into the Future of Egocentric Vision

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

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

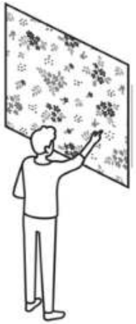
How did we do this?

Dima Damen
July 26, 2024



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

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

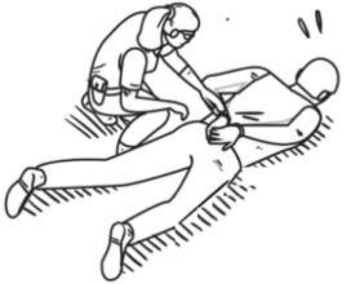


EGO-Designer

EGO-Worker

EGO-Tourist

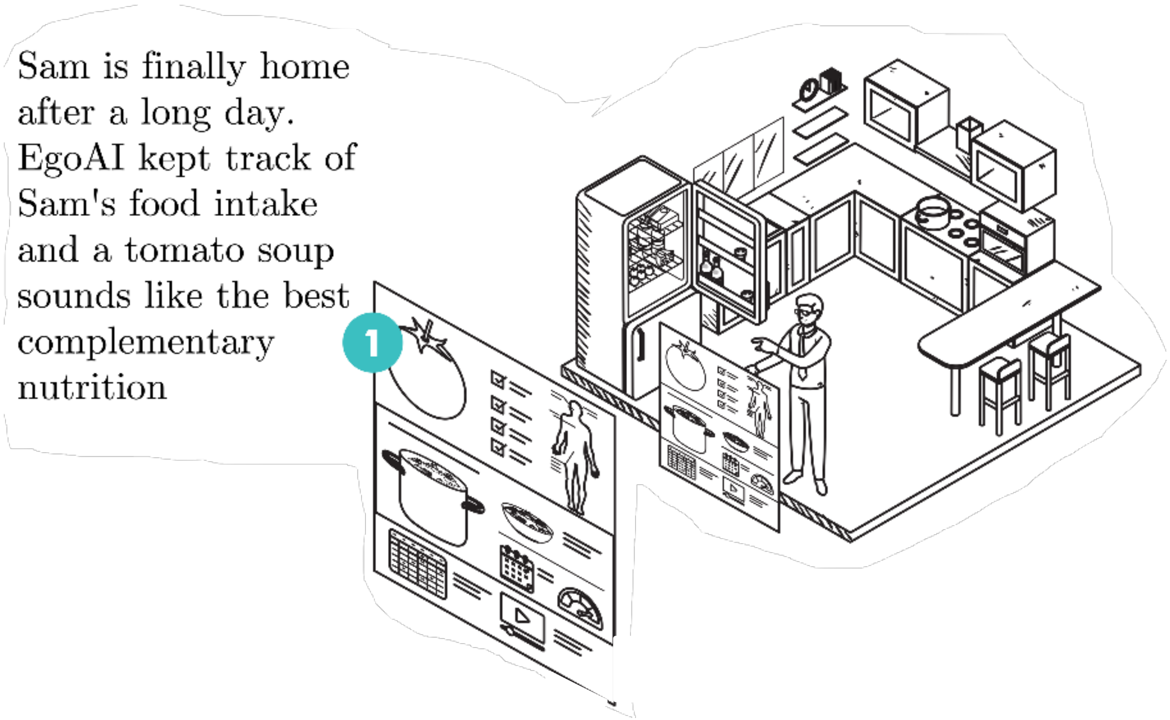
EGO-Home



Ego-Police

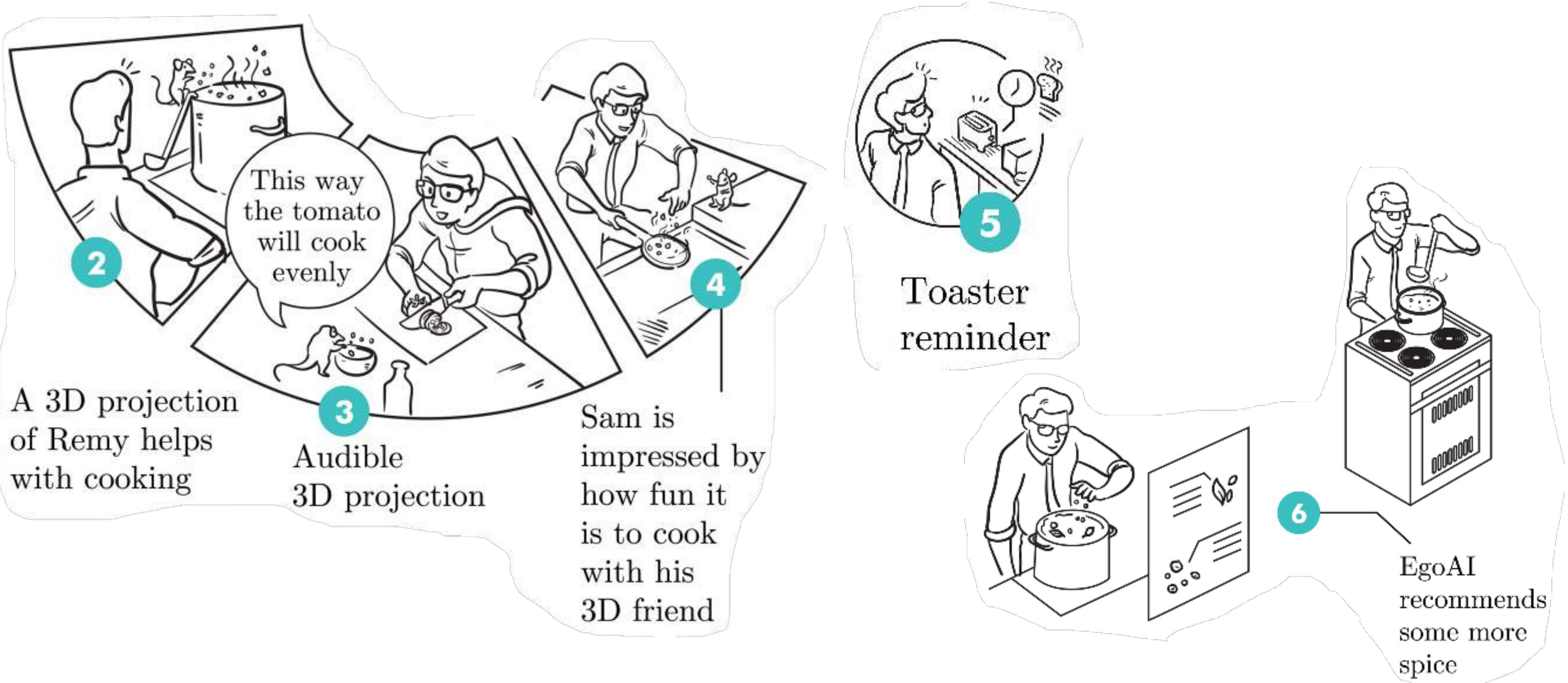
EGO-Home

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



EGO-Home

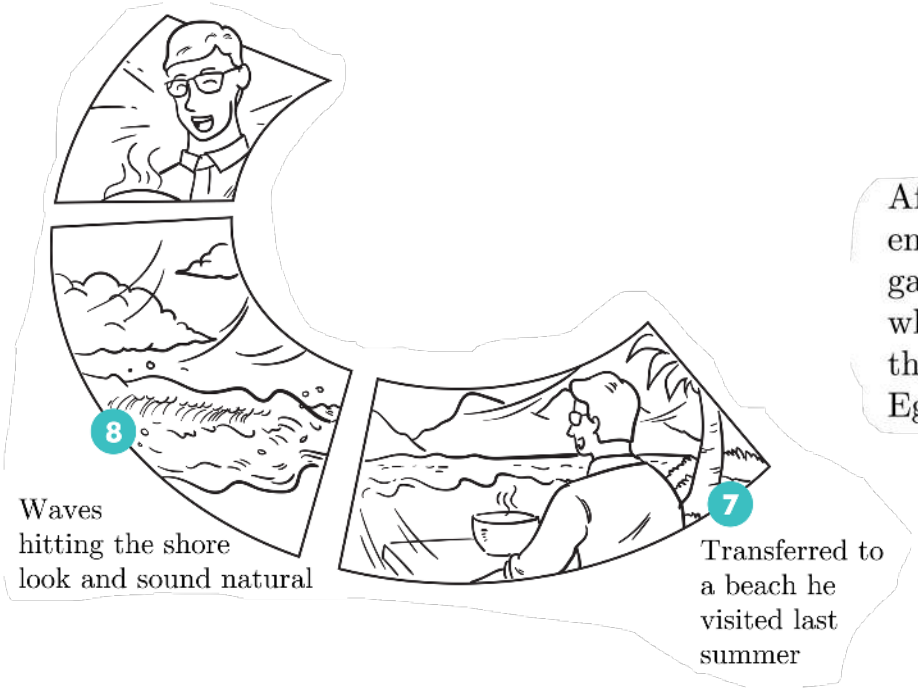
with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi

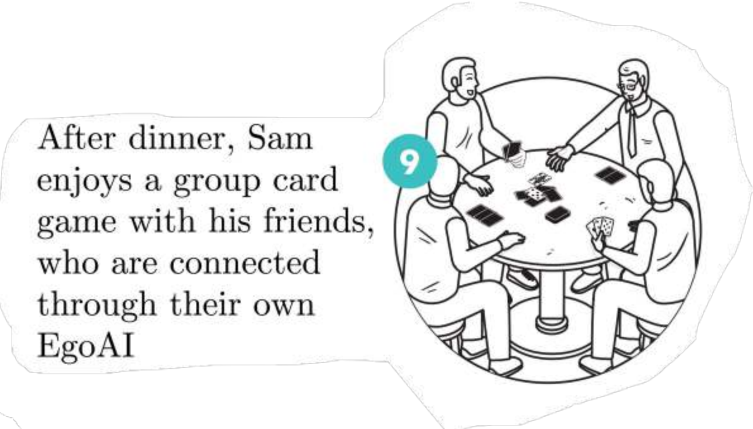


C Plizzari et al (2024). An Outlook into the Future of Egocentric Vision. IJCV

EGO-Home

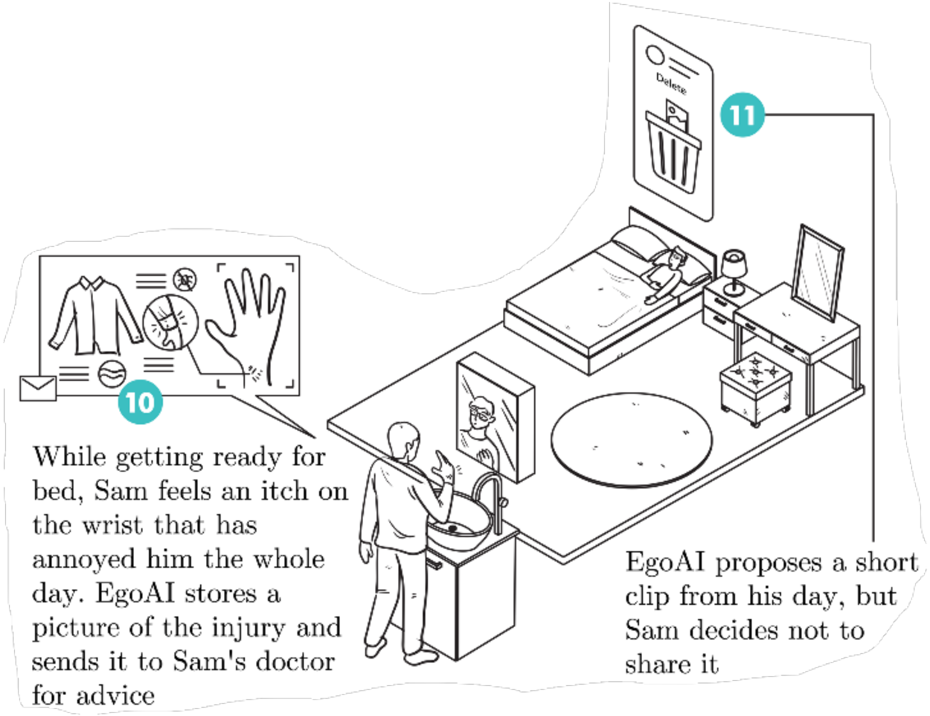
with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi





EGO-Home

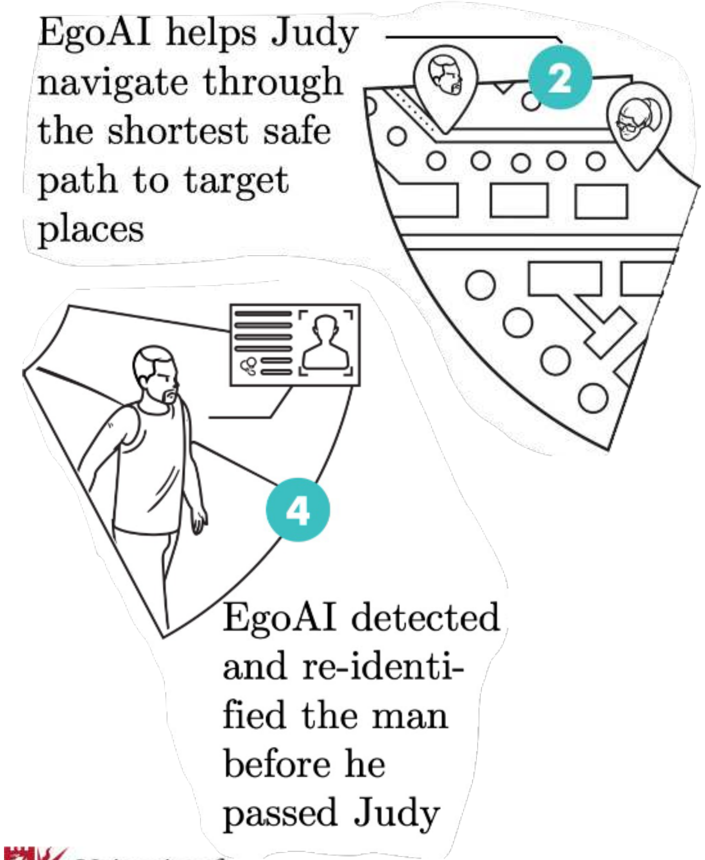
with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi

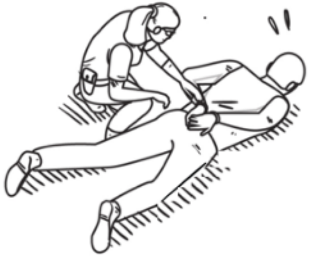


C Plizzari et al (2024). An Outlook into the Future of Egocentric Vision. IJCV

From Stories to Tasks

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



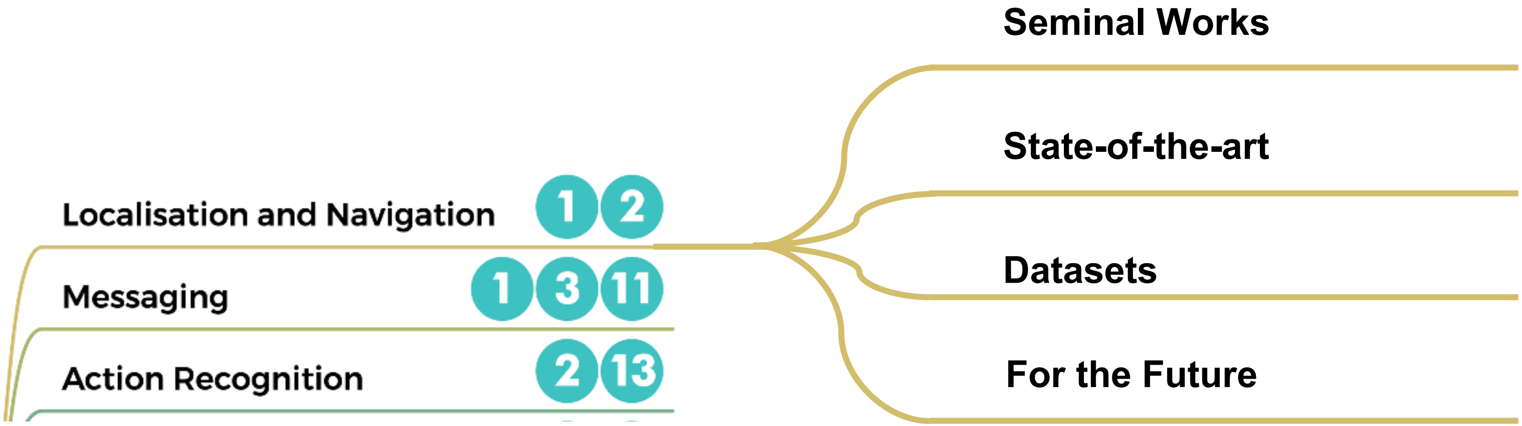


EGO-Police

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

The Survey Part

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



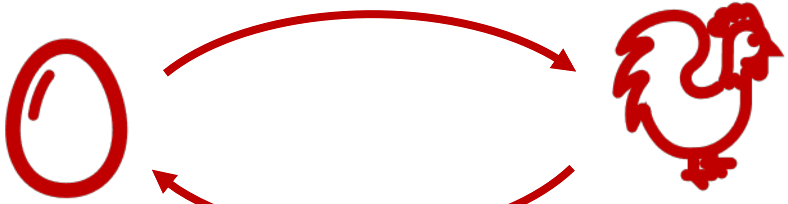
The Survey Part

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

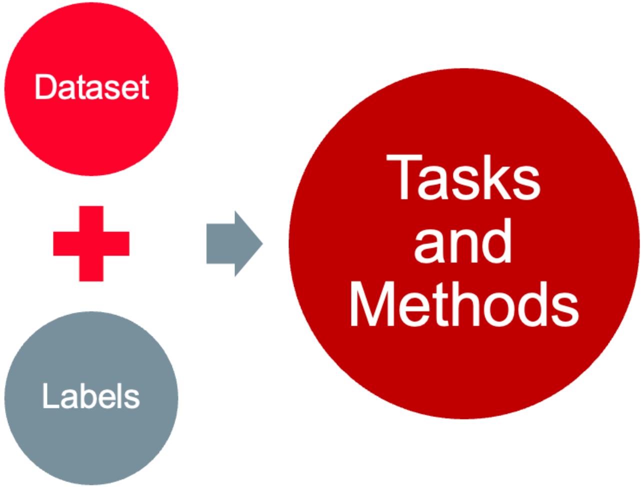
- 12 tasks
- 46 pages (excluding references)
- 462 references

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

In this talk…



Labels **Data**



University of BRISTOL

Thank you

For further info, datasets, code, publications…

http://dimadamen.github.io

@dimadamen

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

Q&A