A fine-grained Perspective onto Object Interactions

University of BRISTOL

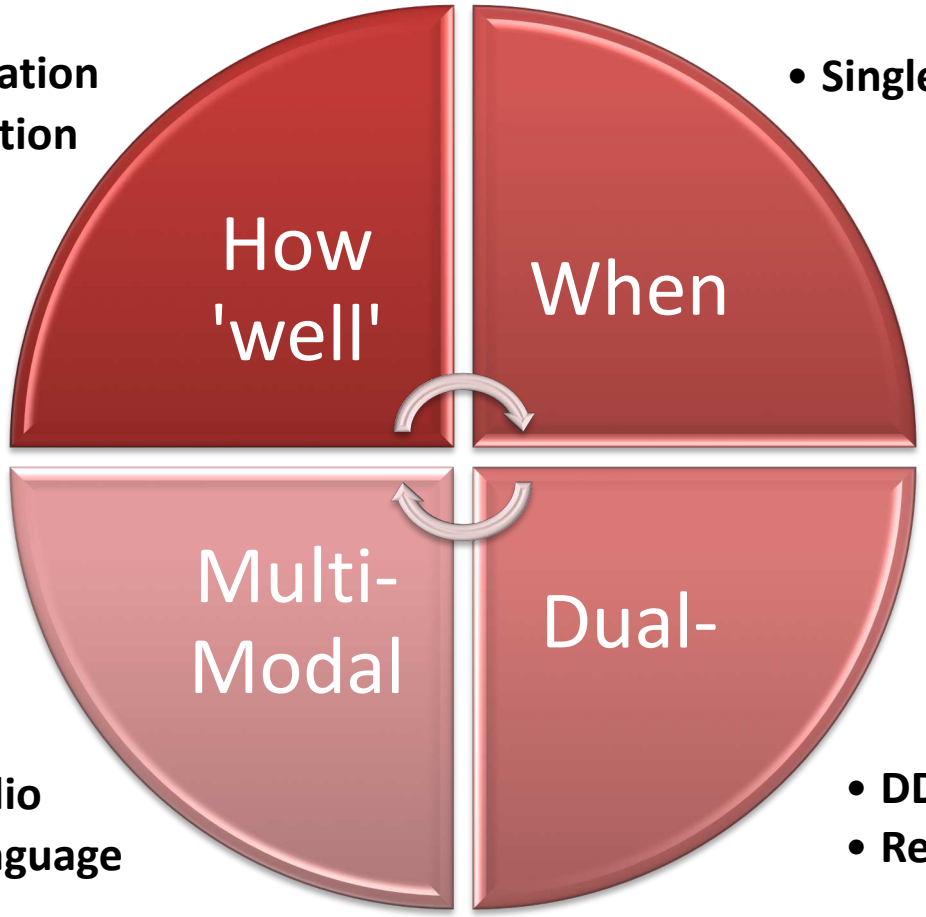
Dima Damen

14 November 2019

Natural interactions

Fine-Grained Object Interactions

- Skill Determination
- Action Completion



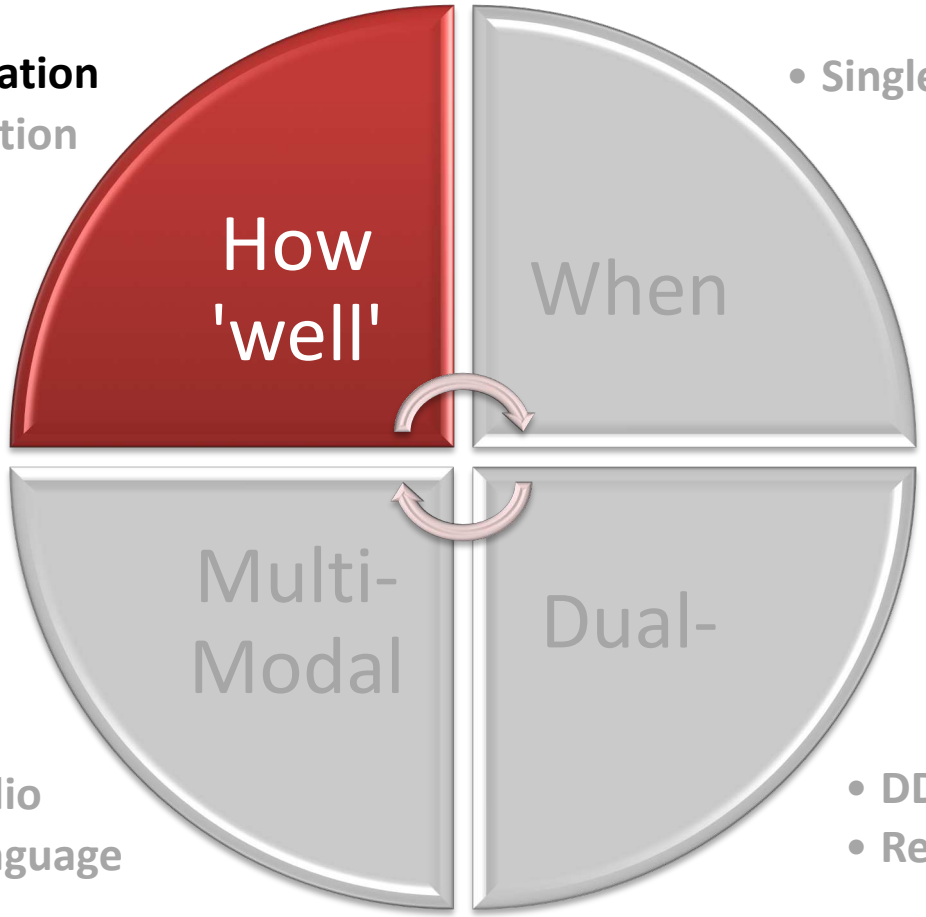
- Single-timestamp

- Vision+Audio
- Vision+Language

- DDLSTM
- Retro-Actions

Fine-Grained Object Interactions

- **Skill Determination**
- Action Completion



- Single-timestamp

- Vision+Audio
- Vision+Language

- DDLSTM
- Retro-Actions

Who's Better? Who's Best? Skill Determination in Video using Deep Ranking

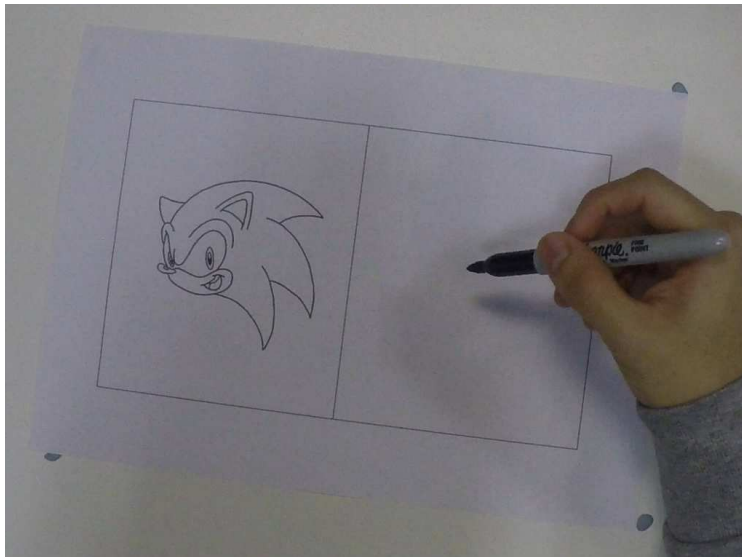
with: Hazel Doughty
Walterio Mayol-Cuevas

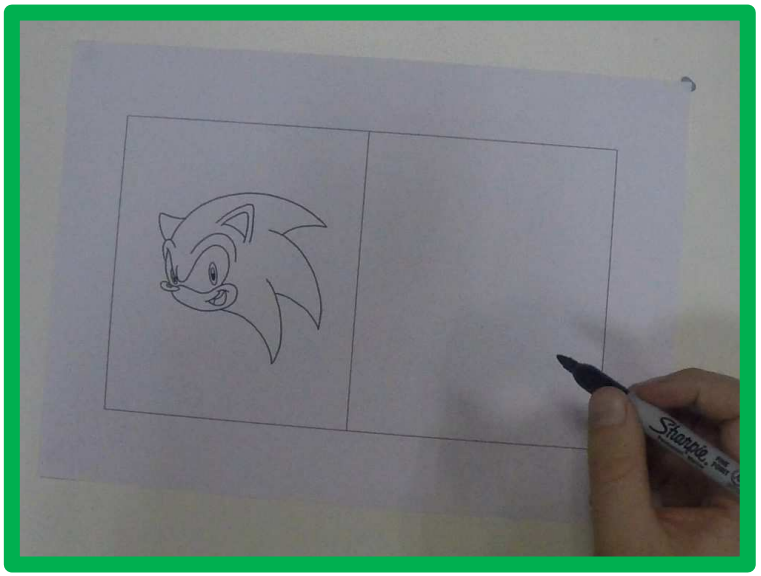
Assess relative skill for a collection of video sequences, applicable to a variety of tasks.

with: Hazel Doughty
Walterio Mayol-Cuevas

Skill Determination from Video

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



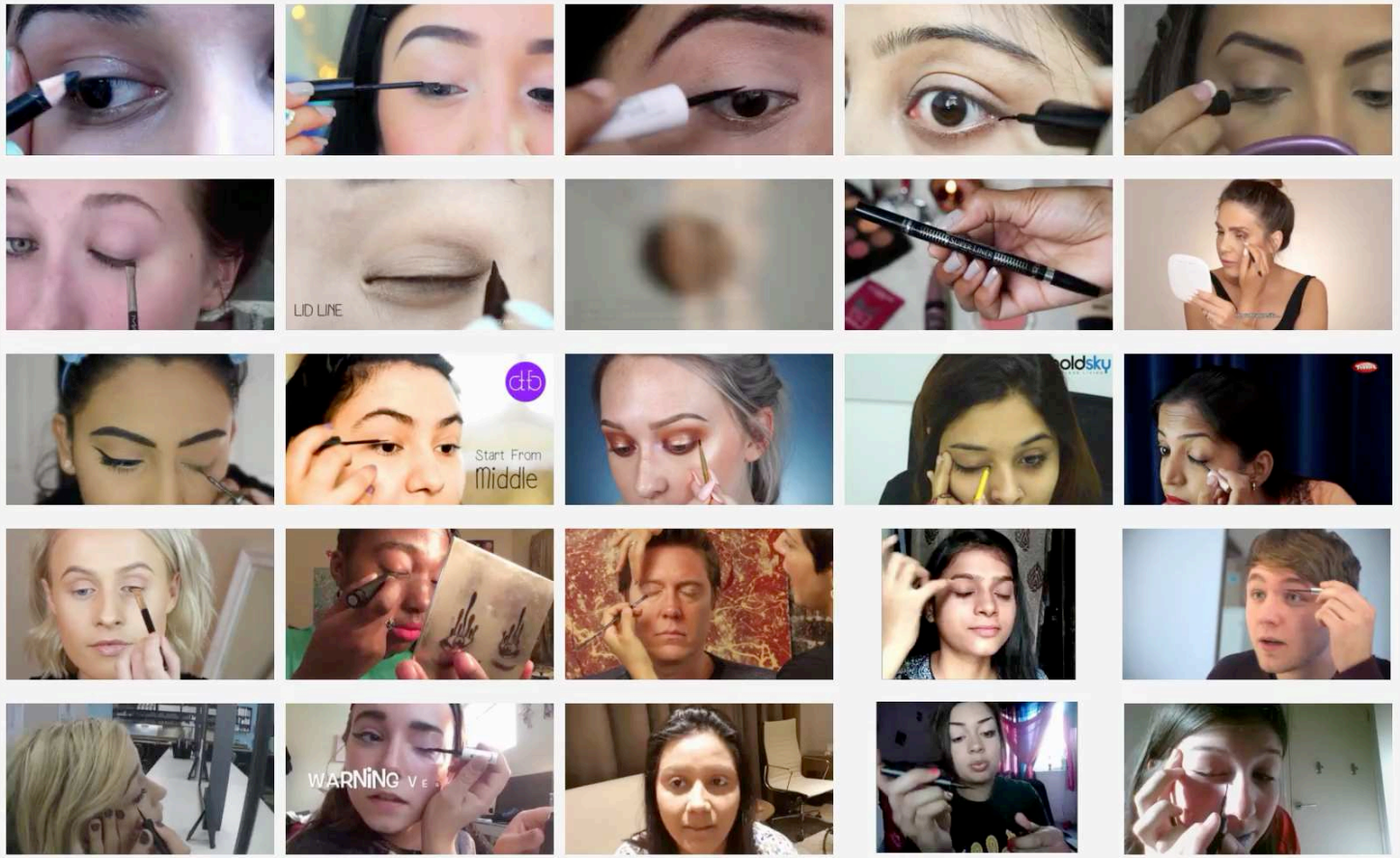


H Doughty, W Mayol-Cuevas, D Damen (2019). The Pros and Cons: Rank-aware Temporal Attention for Skill Determination in Long Videos. *Computer Vision and Pattern Recognition (CVPR)*

Skill Determination in Video

with: Hazel Doughty
Walterio Mayol-Cuevas

Best

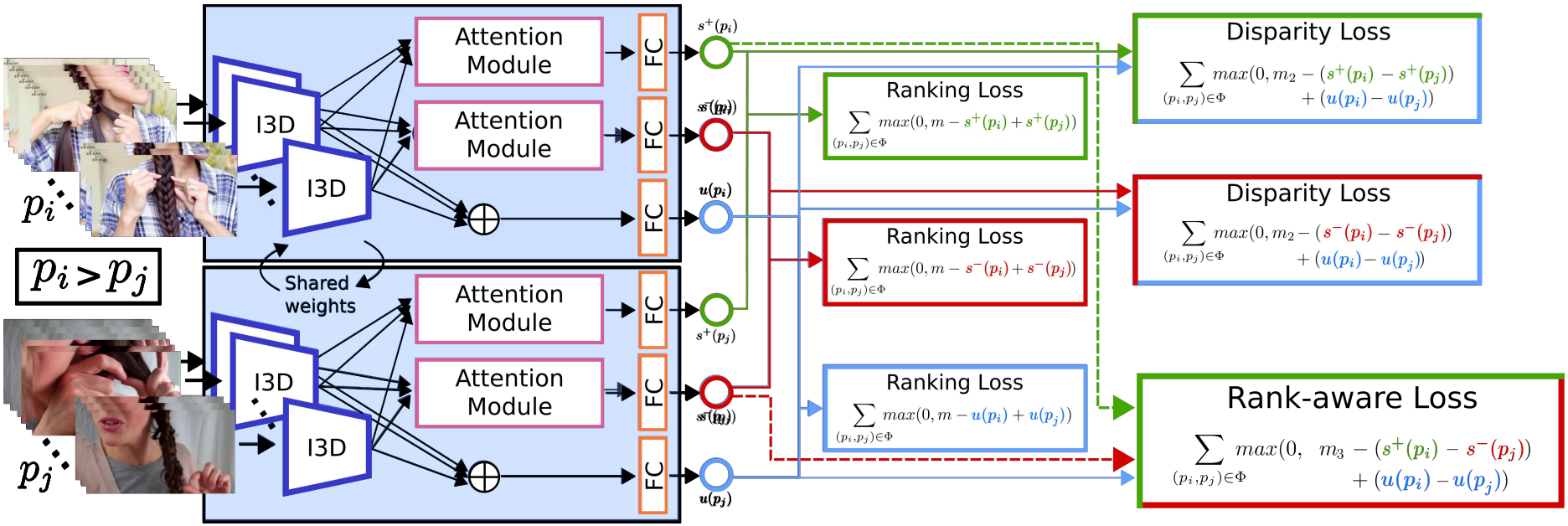


Worst

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)*

The Pros and Cons: Rank-Aware Temporal Attention

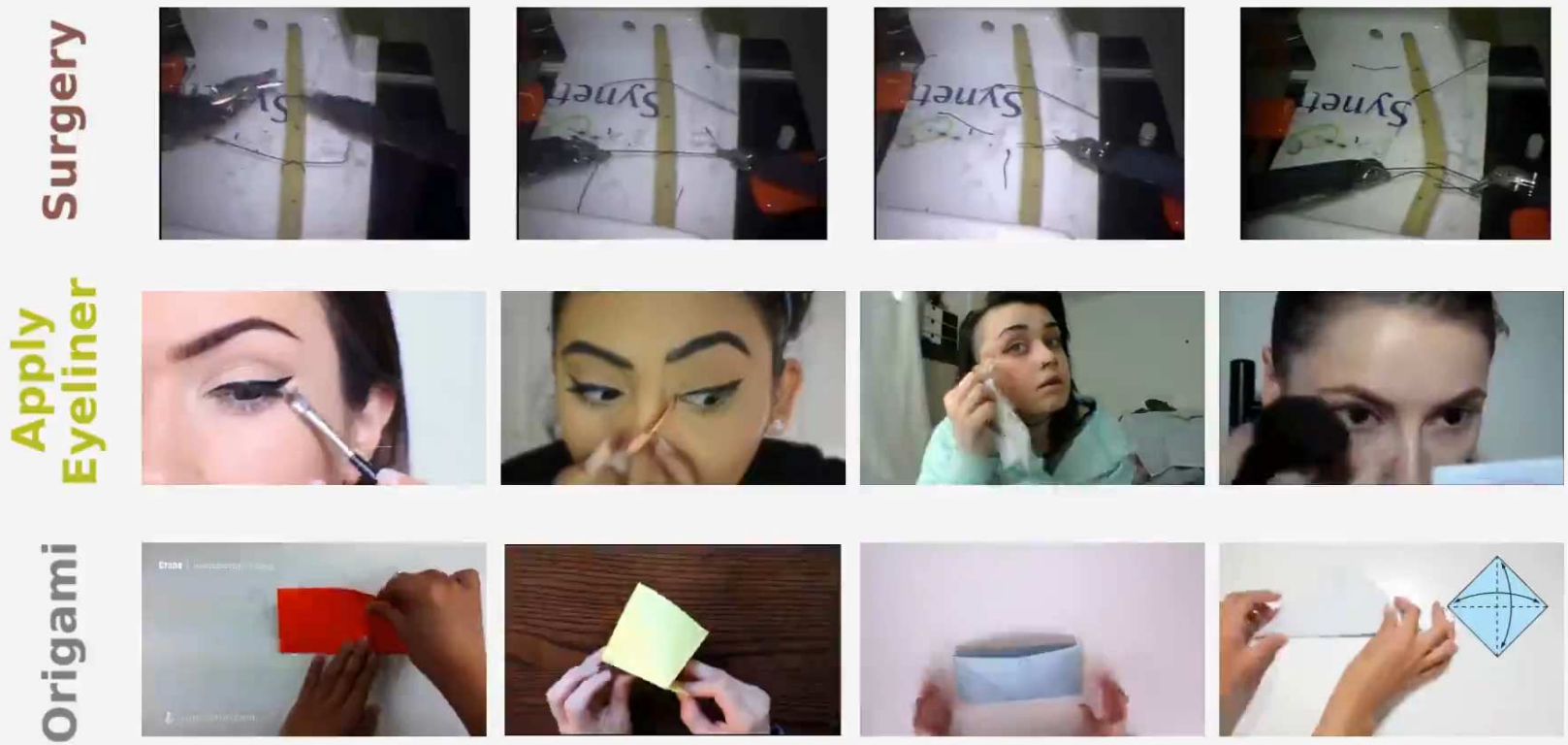
with: Hazel Doughty
Walterio Mayol-Cuevas



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)*

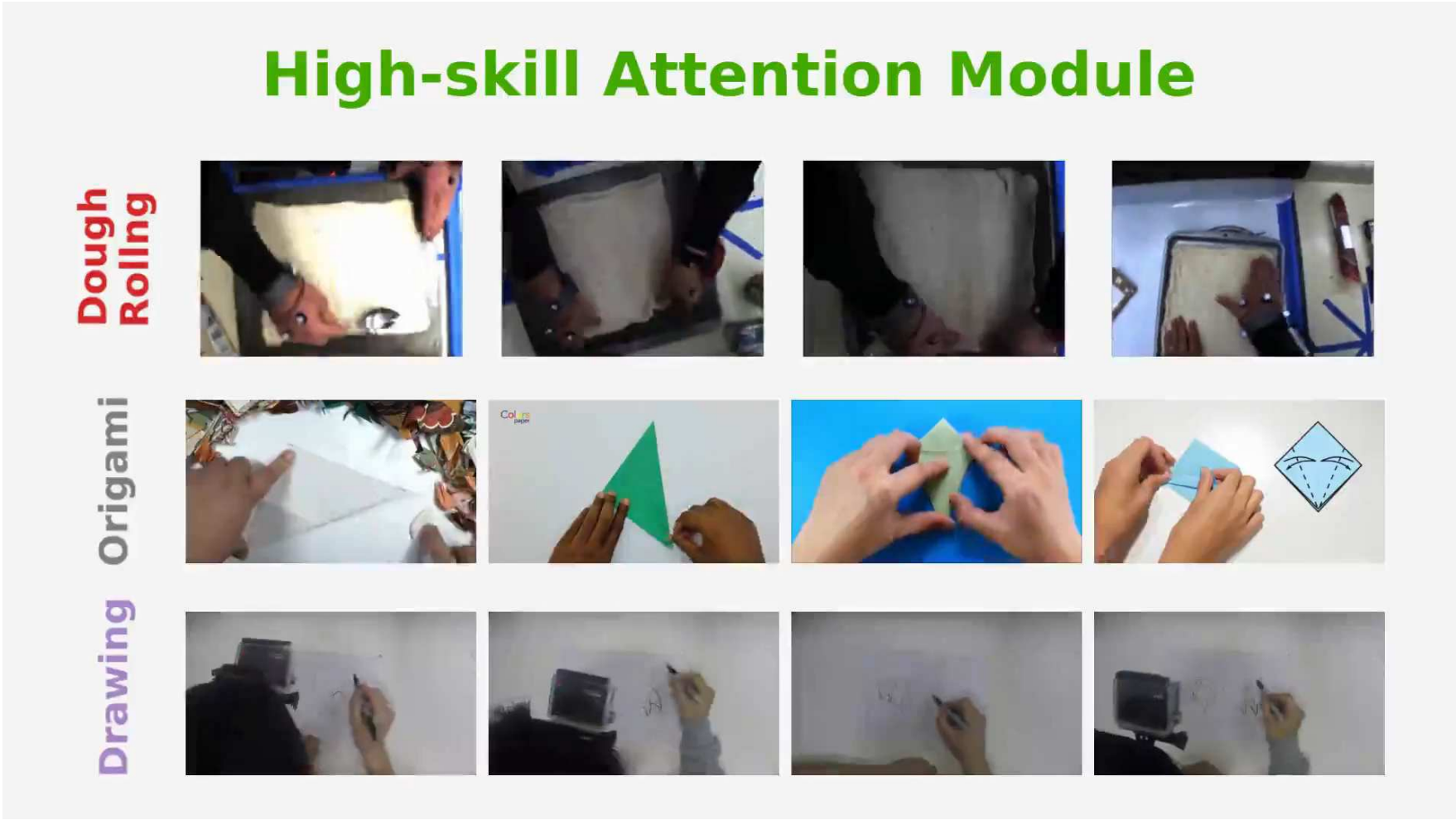
The Pros and Cons: Rank-Aware Temporal Attention

with: Hazel Doughty
Walterio Mayol-Cuevas



The Pros and Cons: Rank-Aware Temporal Attention

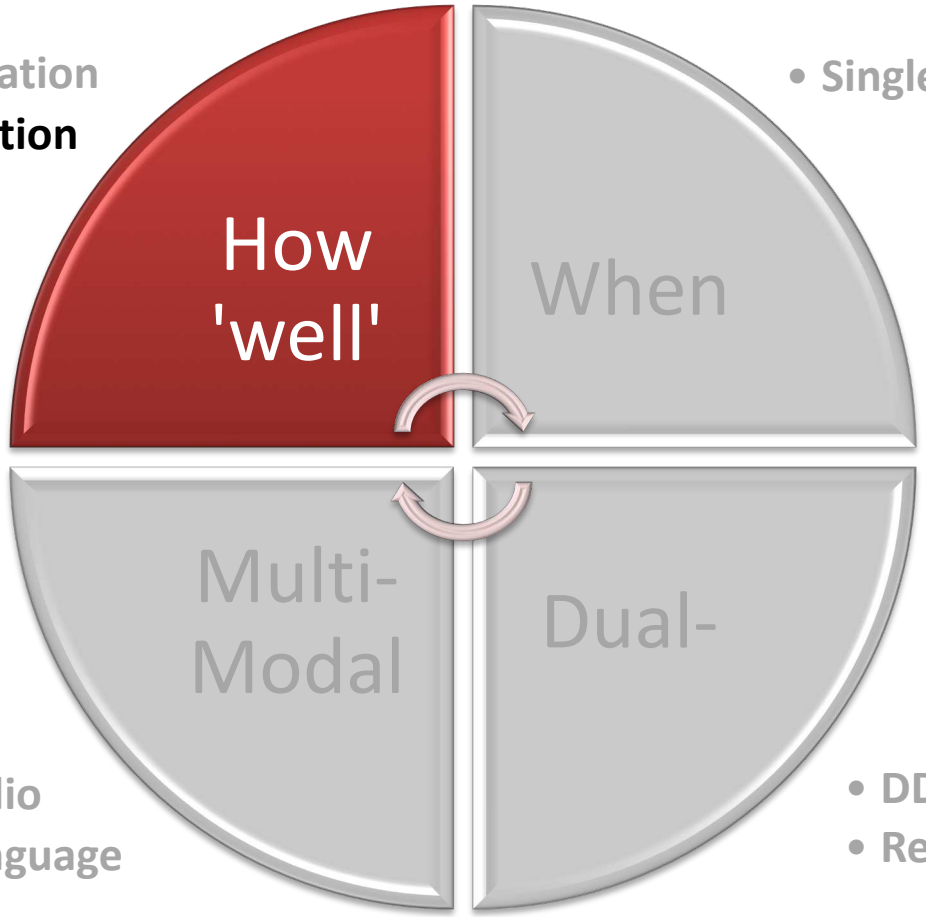
with: Hazel Doughty
Walterio Mayol-Cuevas



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)*

Fine-Grained Object Interactions

- Skill Determination
- **Action Completion**



- Single-timestamp

- Vision+Audio
- Vision+Language

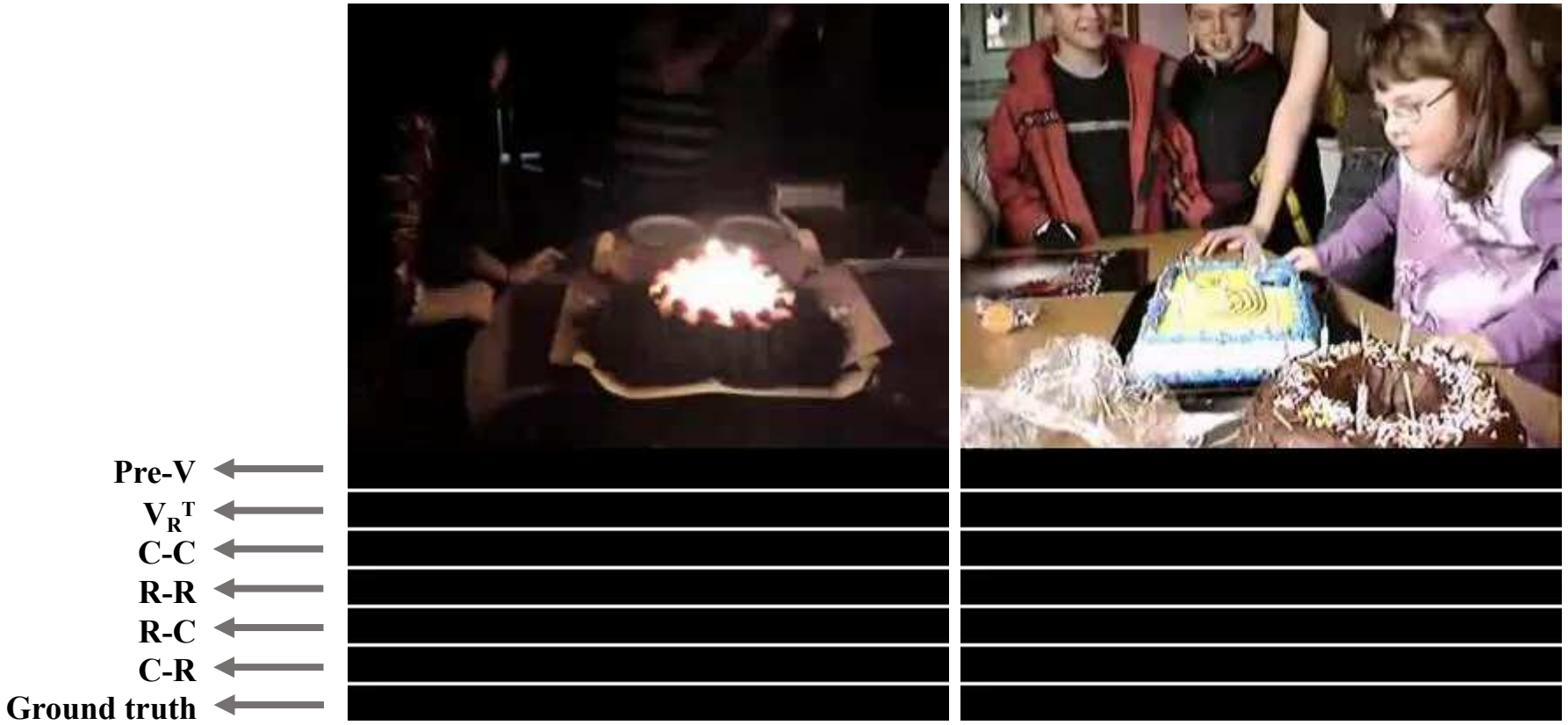
- DDLSTM
- Retro-Actions

with: Farnoosh Heidarivincheh
Majid Mirmehdi

Action Completion Detection

with: Farnoosh Heidarivincheh
Majid Mirmehdi

Action Completion Detection

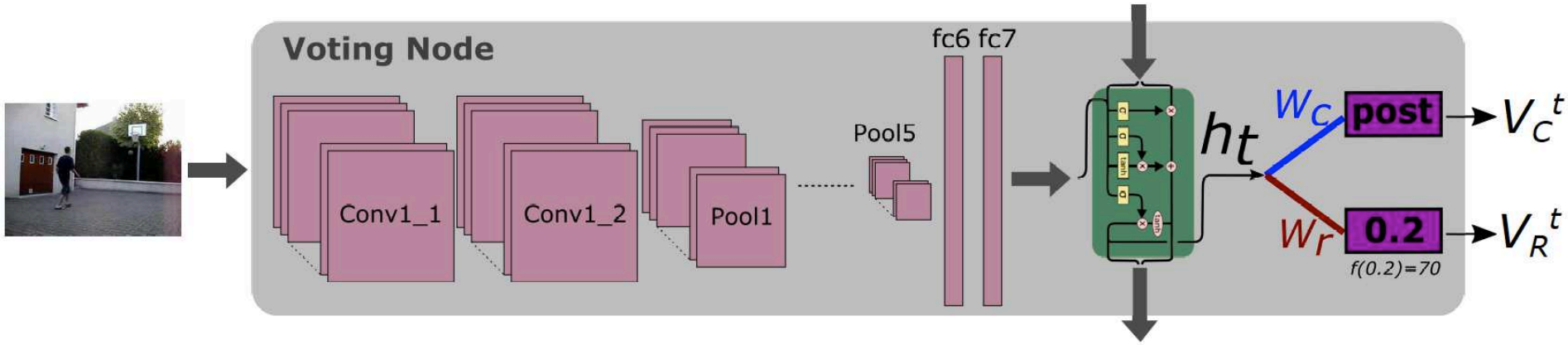


F Heidarivincheh, M Mirmehdi, D Damen (2018). Action Completion: A Temporal Model for Moment Detection. BMVC

with: Farnoosh Heidarivincheh
Majid Mirmehdi

Action Completion Detection

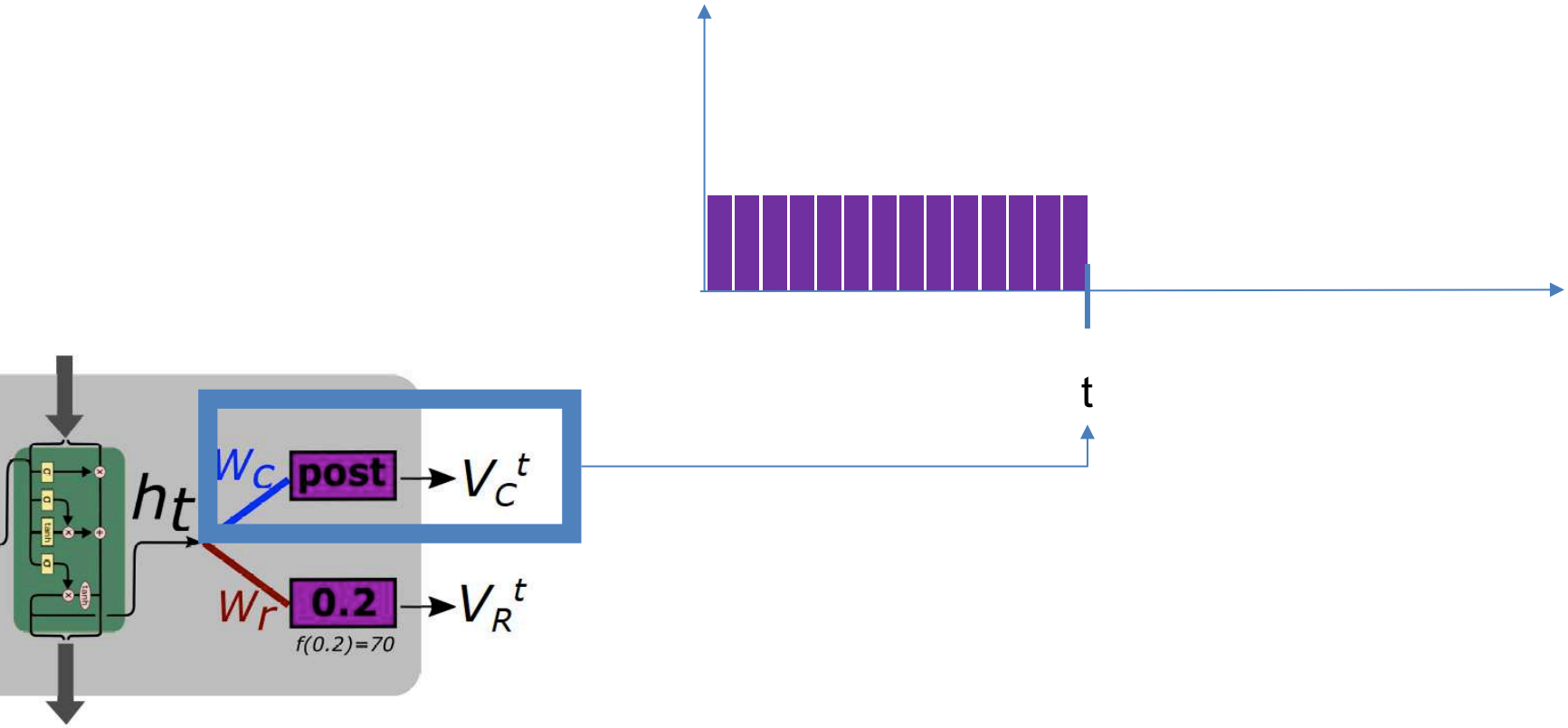
- Each frame in the sequence, contributes to the completion moment detection via ‘voting’



with: Farnoosh Heidarivincheh
Majid Mirmehdi

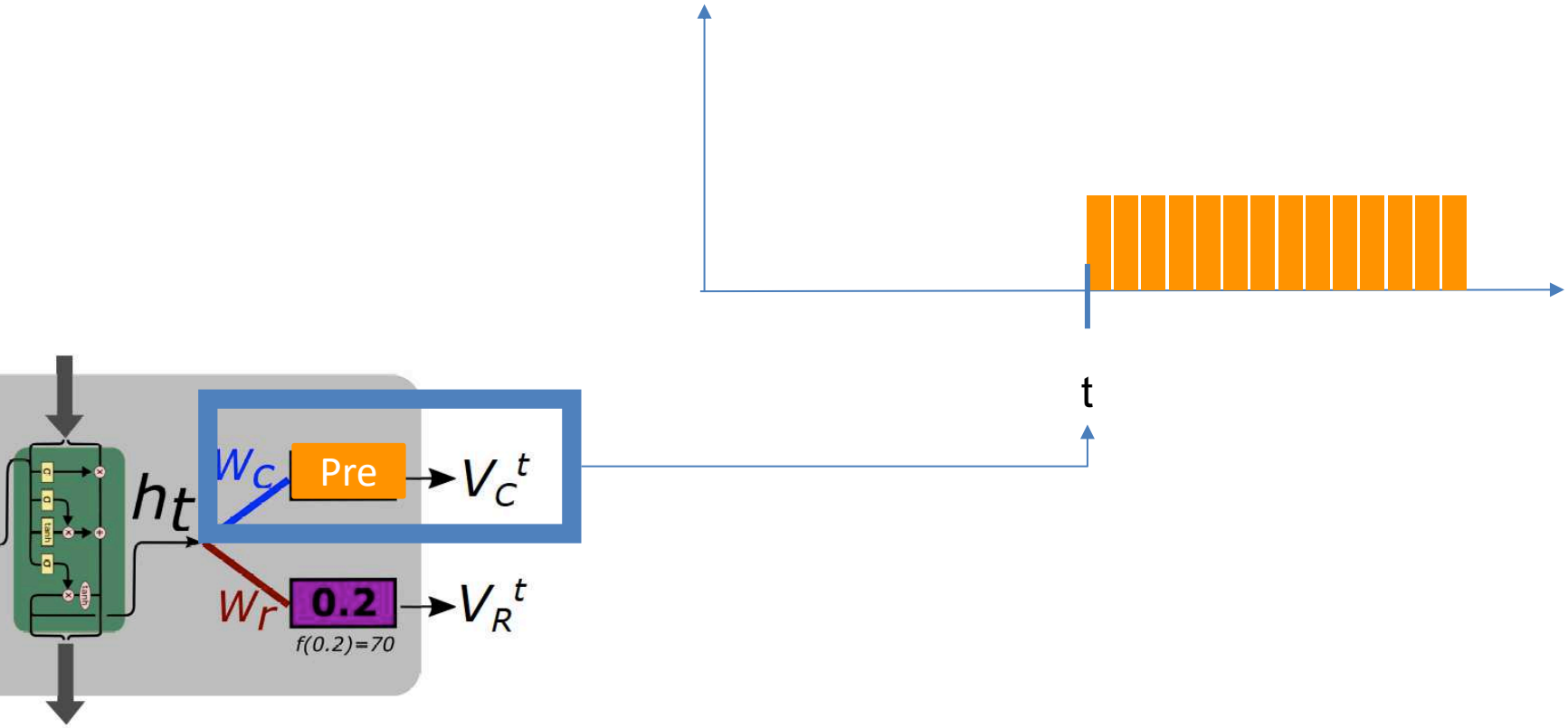
Action Completion Detection

1. Classification-Based Voting



Action Completion Detection

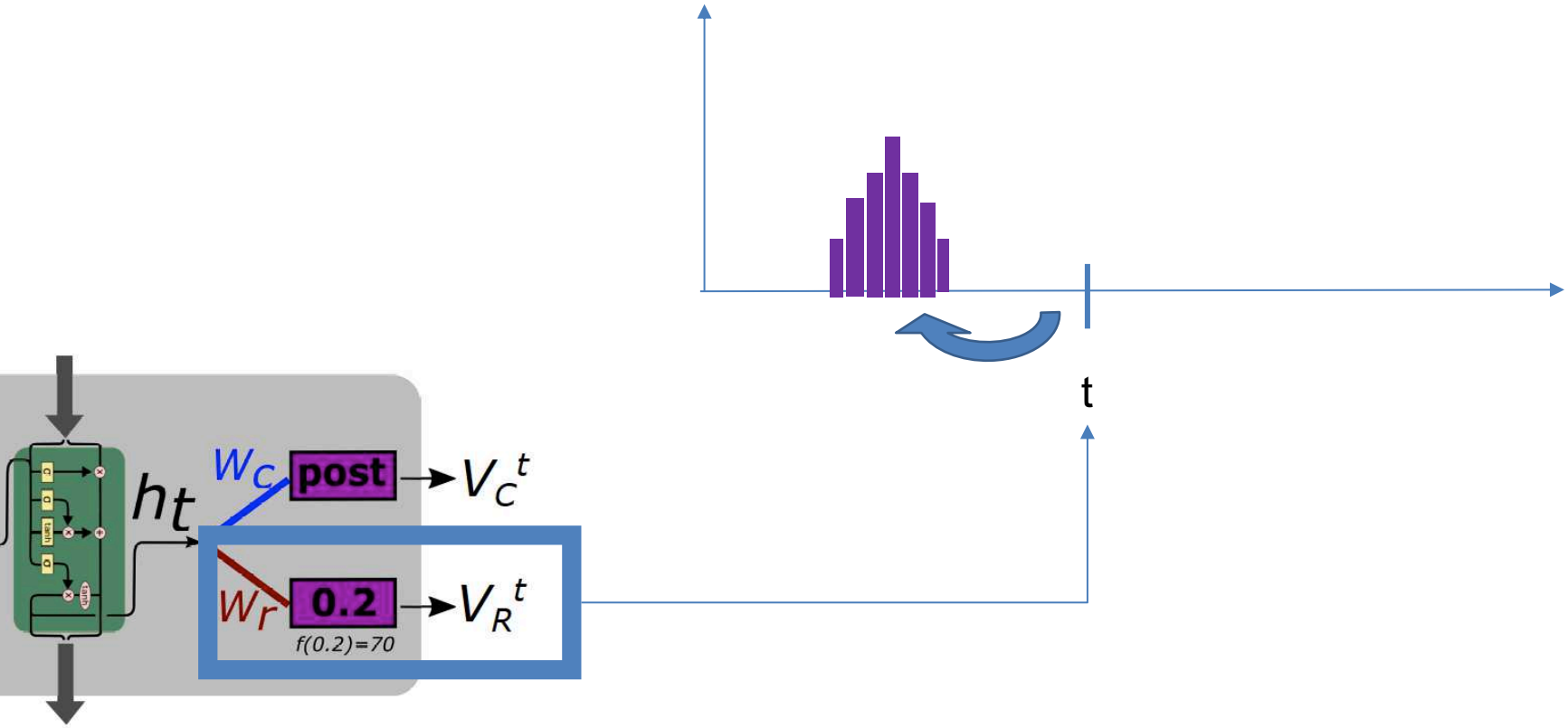
1. Classification-Based Voting



with: Farnoosh Heidarivincheh
Majid Mirmehdi

Action Completion Detection

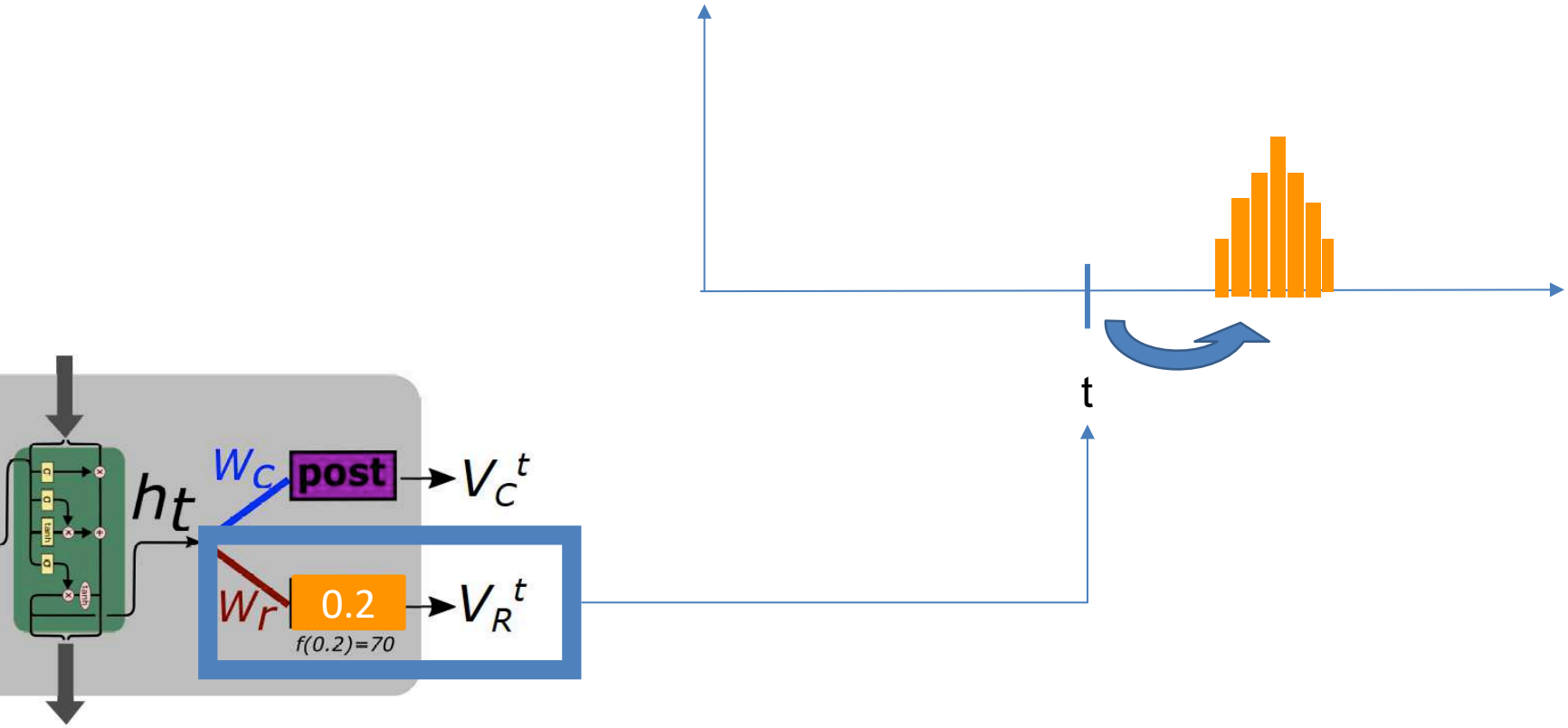
2. Regression-Based Voting



with: Farnoosh Heidarivincheh
Majid Mirmehdi

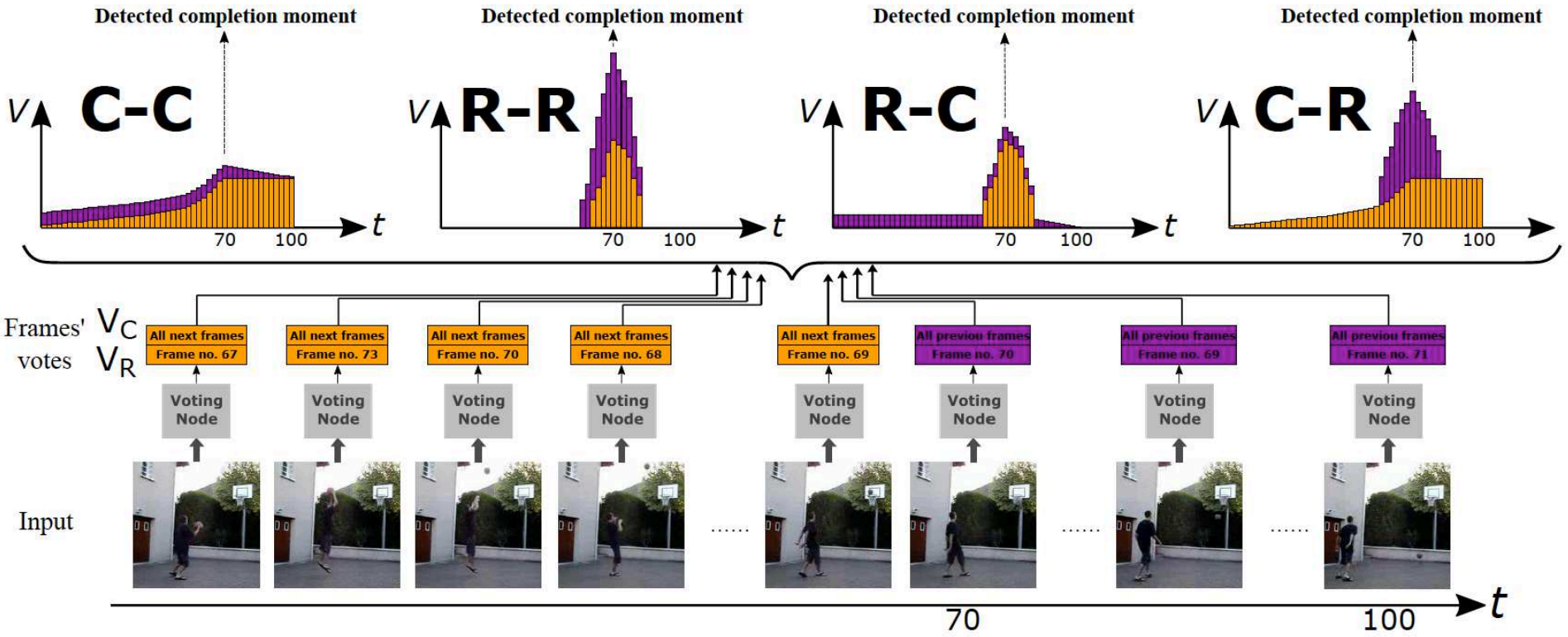
Action Completion Detection

2. Regression-Based Voting



with: Farnoosh Heidarivincheh
Majid Mirmehdi

Action Completion Detection



F Heidarivincheh, M Mirmehdi, D Damen (2018). Action Completion: A Temporal Model for Moment Detection. BMVC

with: Farnoosh Heidarivincheh
Majid Mirmehdi

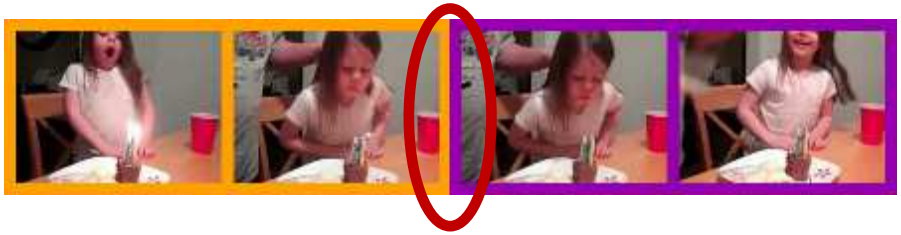
Action Completion Detection

F Heidarivincheh, M Mirmehdi, D Damen (2018). Action Completion: A Temporal Model for Moment Detection. BMVC

with: Farnoosh Heidarivincheh
Majid Mirmehdi

Action Completion Detection

Frame-level labels: annotations are expensive, subjective and noisy.



We detect completion using only **weak labels** during training.

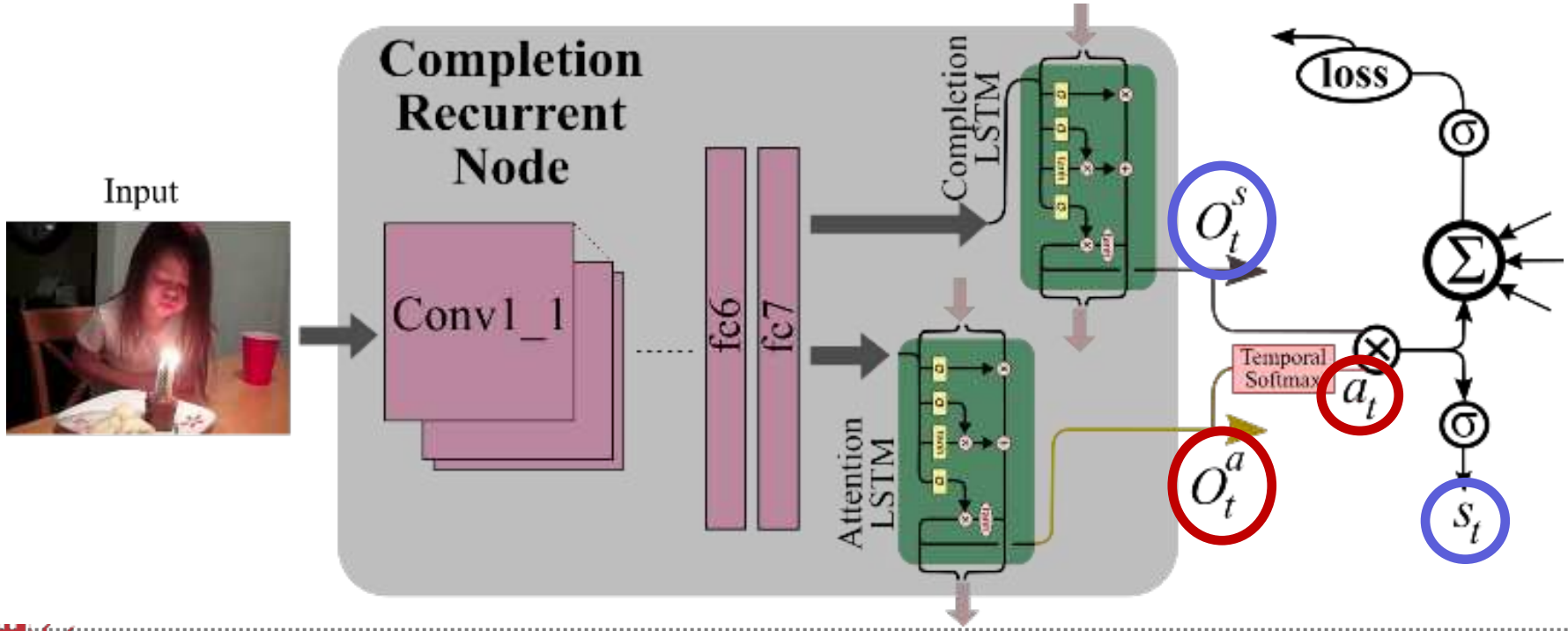
sequence-level *complete* and ***incomplete*** labels

with: Farnoosh Heidarivincheh
Majid Mirmehdi

Action Completion Detection

Attention LSTM → o_t^a → Softmax → a_t

Completion LSTM → o_t^s → $\sigma(o_t^s \times a_t)$ → s_t



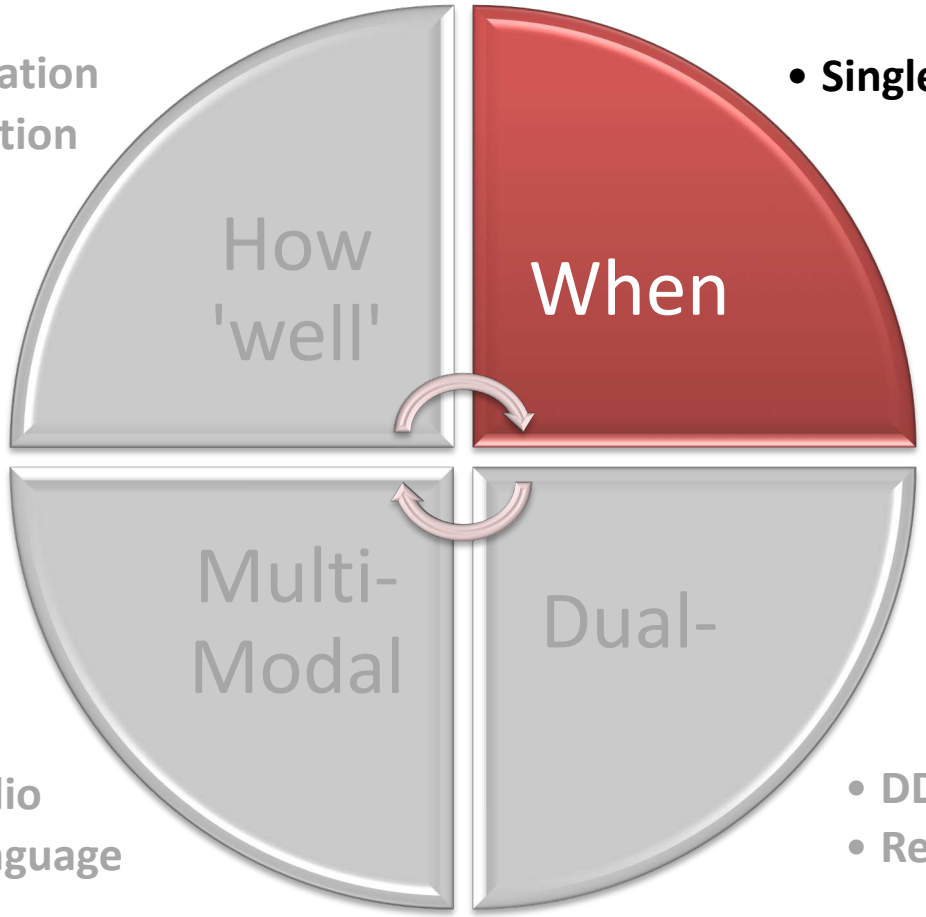
F Heidarivincheh, M Mirmehdi, D Damen (2019). Weakly-Supervised Completion Moment Detection using Temporal Attention. ICCV Workshop on Human Behaviour Understanding

with: Farnoosh Heidarivincheh
Majid Mirmehdi

Action Completion Detection

Fine-Grained Object Interactions

- Skill Determination
- Action Completion

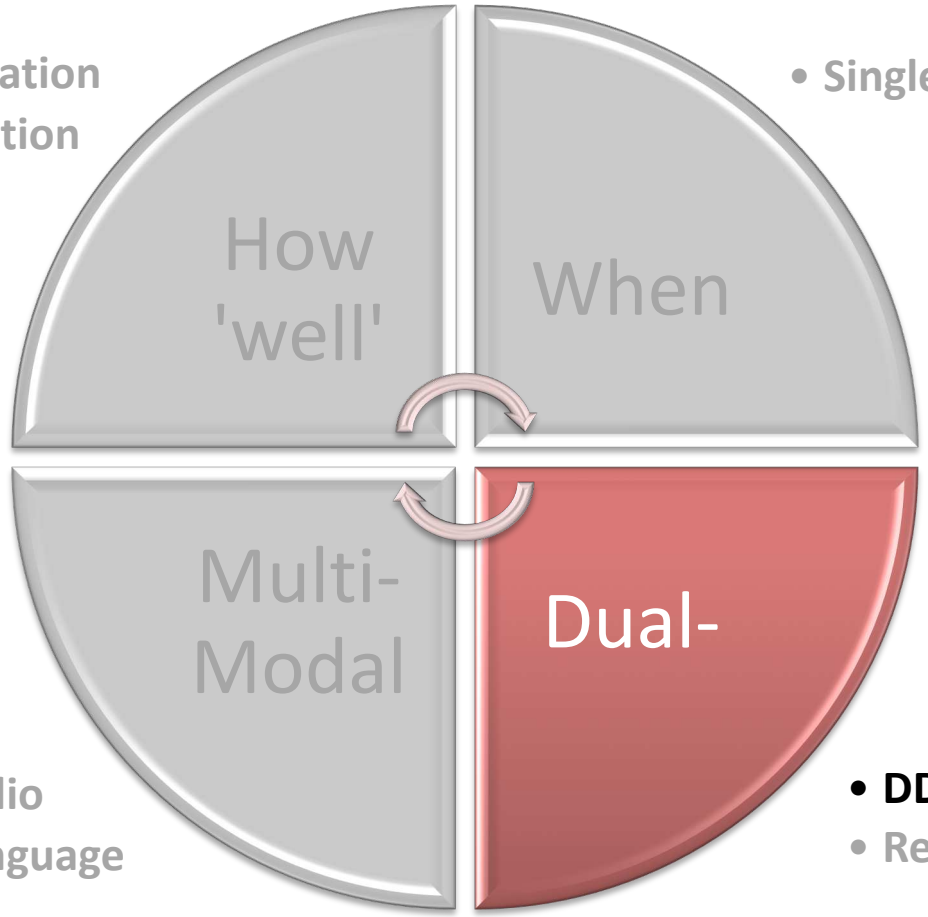


- **Single-timestamp**

- **Vision+Audio**
- **Vision+Language**

- **DDLSTM**
- **Retro-Actions**

Fine-Grained Object Interactions



- Skill Determination
- Action Completion

- Single-timestamp

- Vision+Audio
- Vision+Language

- **DDLSTM**
- Retro-Actions

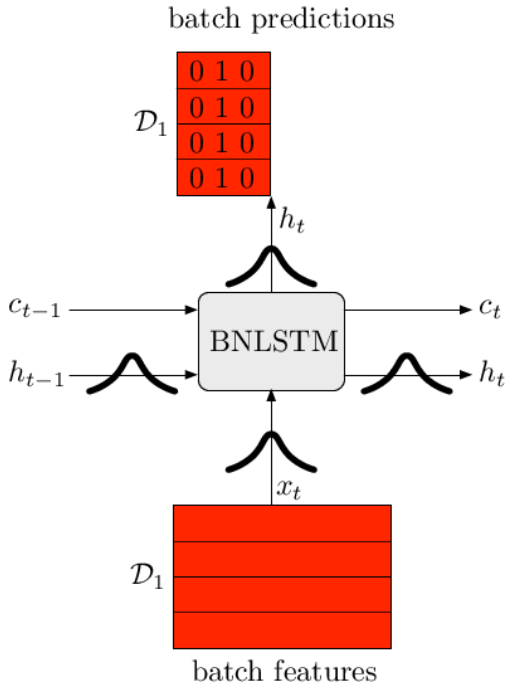
Dual-Domain LSTM for Cross-Dataset Action Recognition

with: Toby Perrett

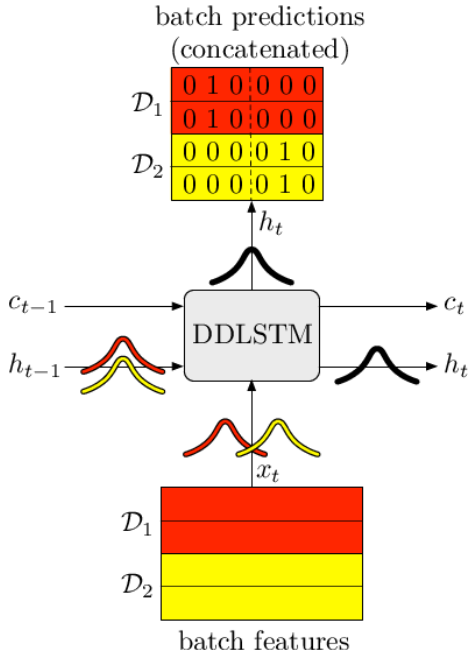
Dual-Domain LSTM for Cross-Dataset Action Recognition

with: Toby Perrett

BNLSTM
1 dataset

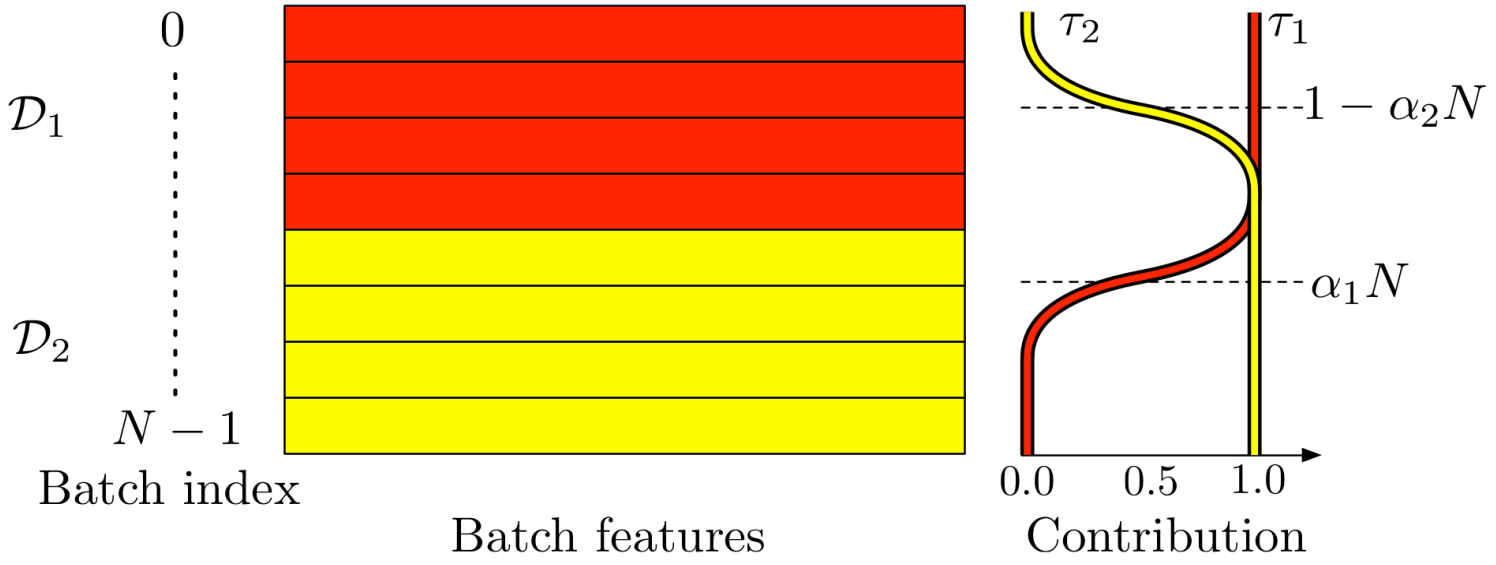


DDLSTM
2 datasets



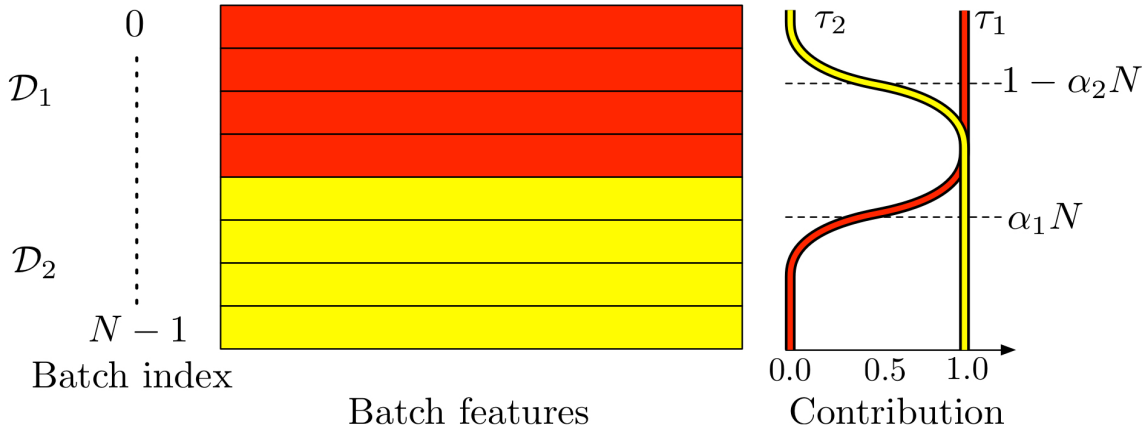
Dual-Domain LSTM for Cross-Dataset Action Recognition

with: Toby Perrett



Dual-Domain LSTM for Cross-Dataset Action Recognition

with: Toby Perrett

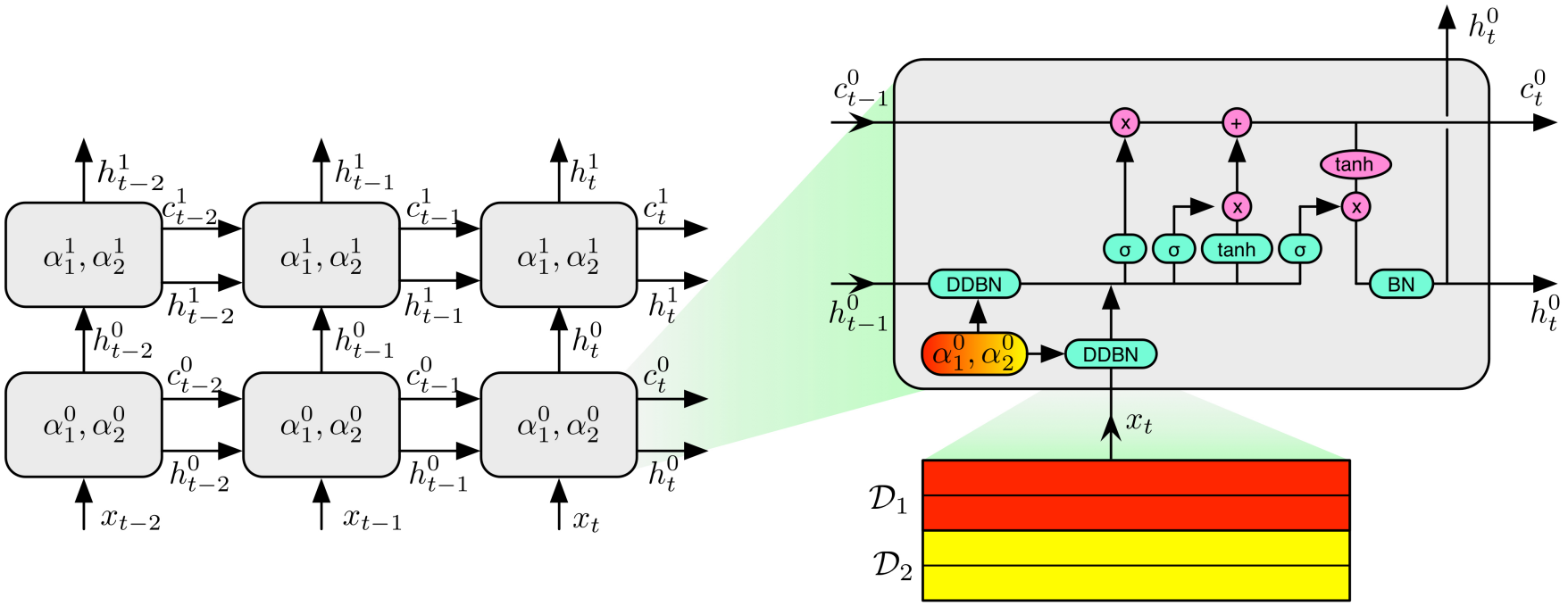


$$\tau_1(\alpha_1, j) = \frac{1 - \tanh(j - \alpha_1 N)}{2}$$

$$\tau_2(\alpha_2, j) = \frac{1 + \tanh(j - \alpha_2 N)}{2}$$

Dual-Domain LSTM for Cross-Dataset Action Recognition

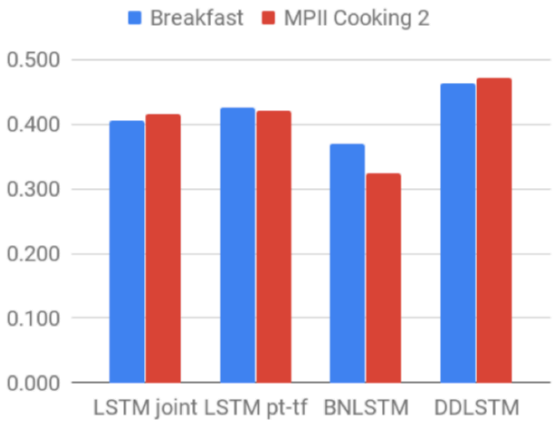
with: Toby Perrett



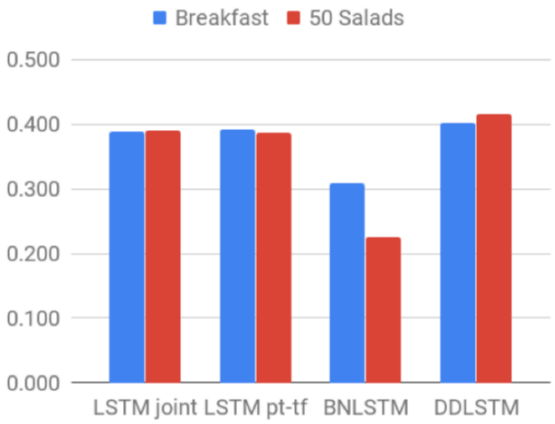
Dual-Domain LSTM for Cross-Dataset Action Recognition

with: Toby Perrett

(a) Breakfast



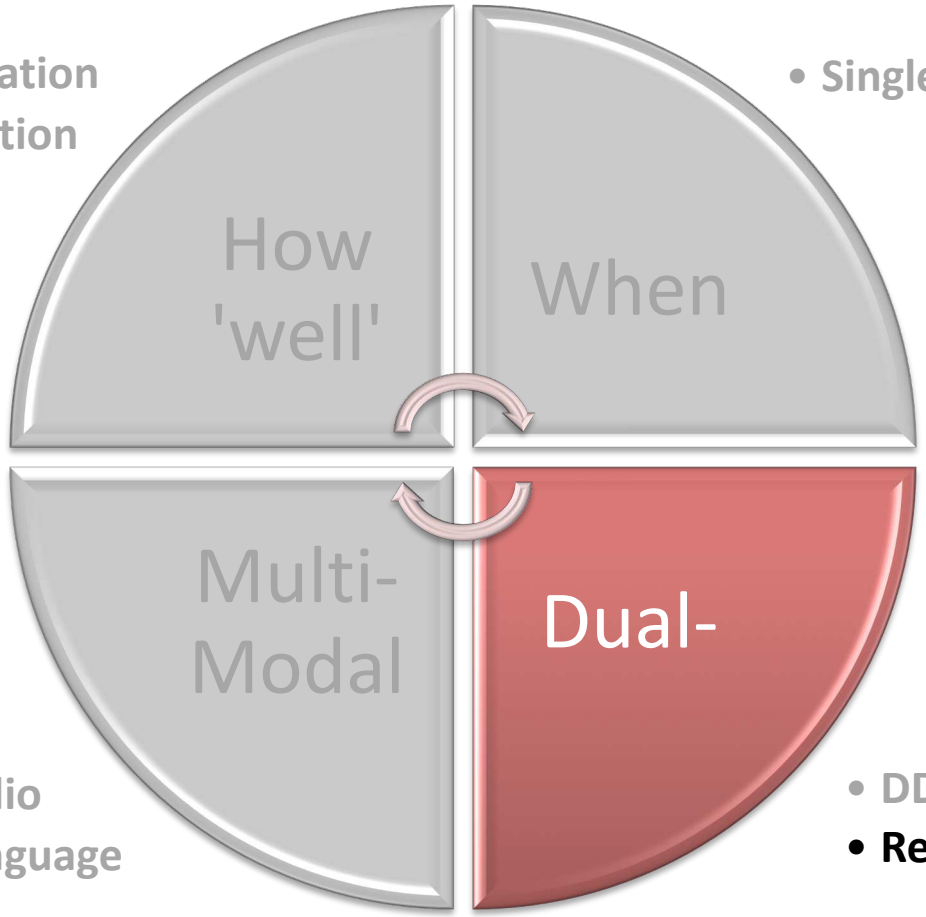
(b) 50 Salads



(c) MPII Cooking 2

Fine-Grained Object Interactions

- Skill Determination
- Action Completion

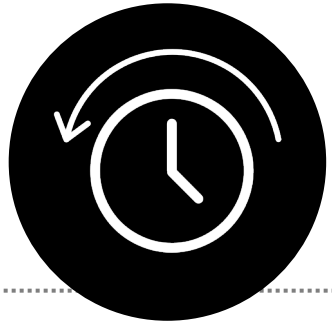


- Single-timestamp

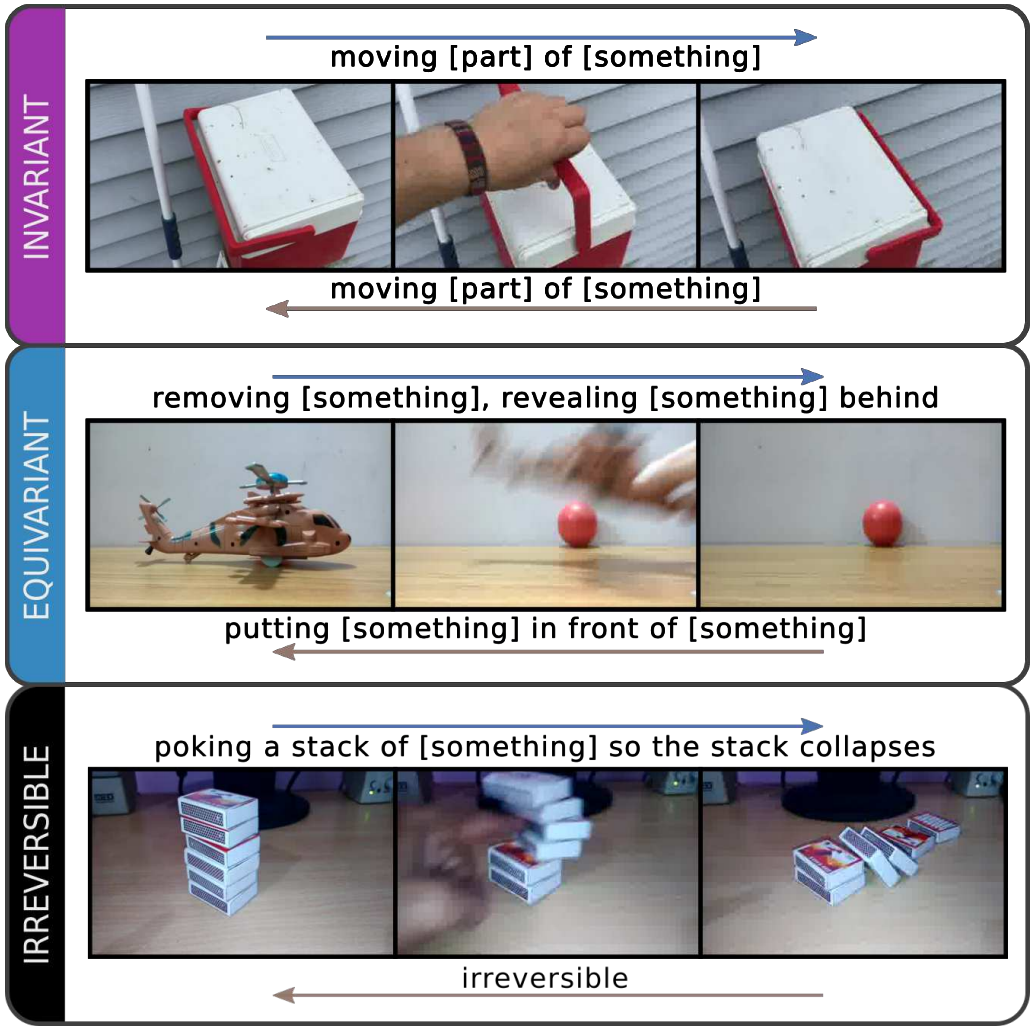
- Vision+Audio
- Vision+Language

- DDLSTM
- **Retro-Actions**

Retro-actions

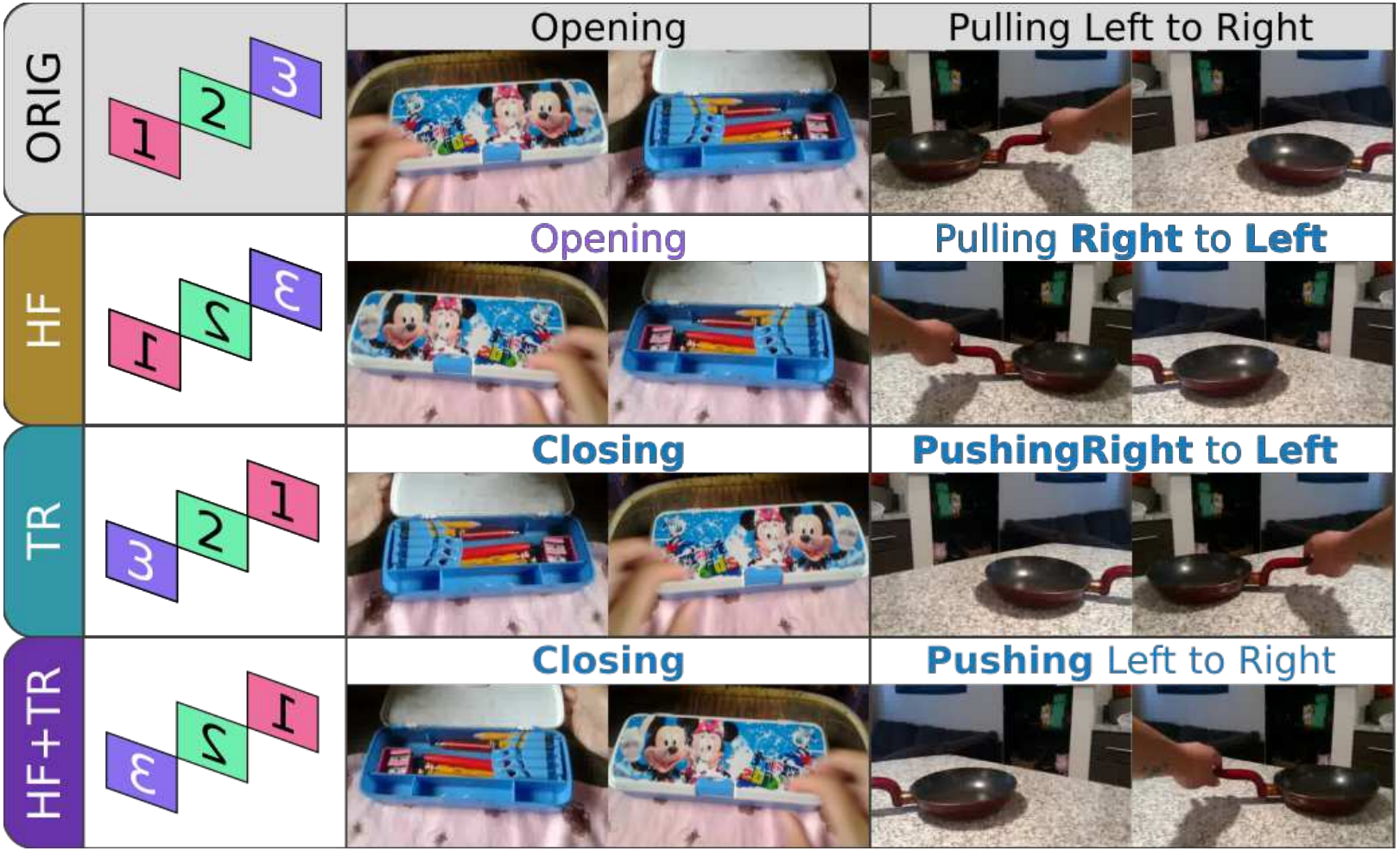


Retro-actions



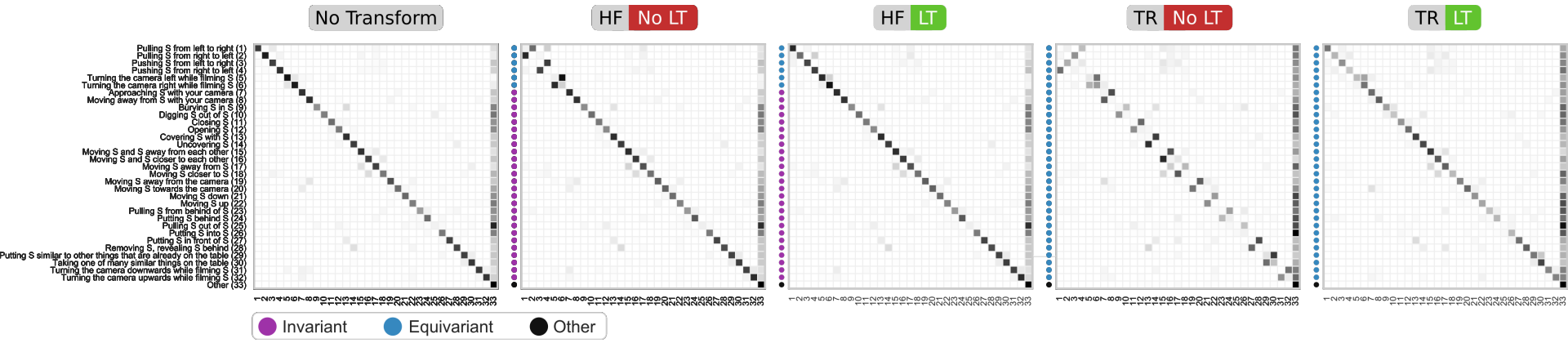
W Price, D Damen (2019). Retro-Actions: Learning 'Close' by Time-Reversing 'Open' Videos. ICCV MDALC Workshop

Retro-actions

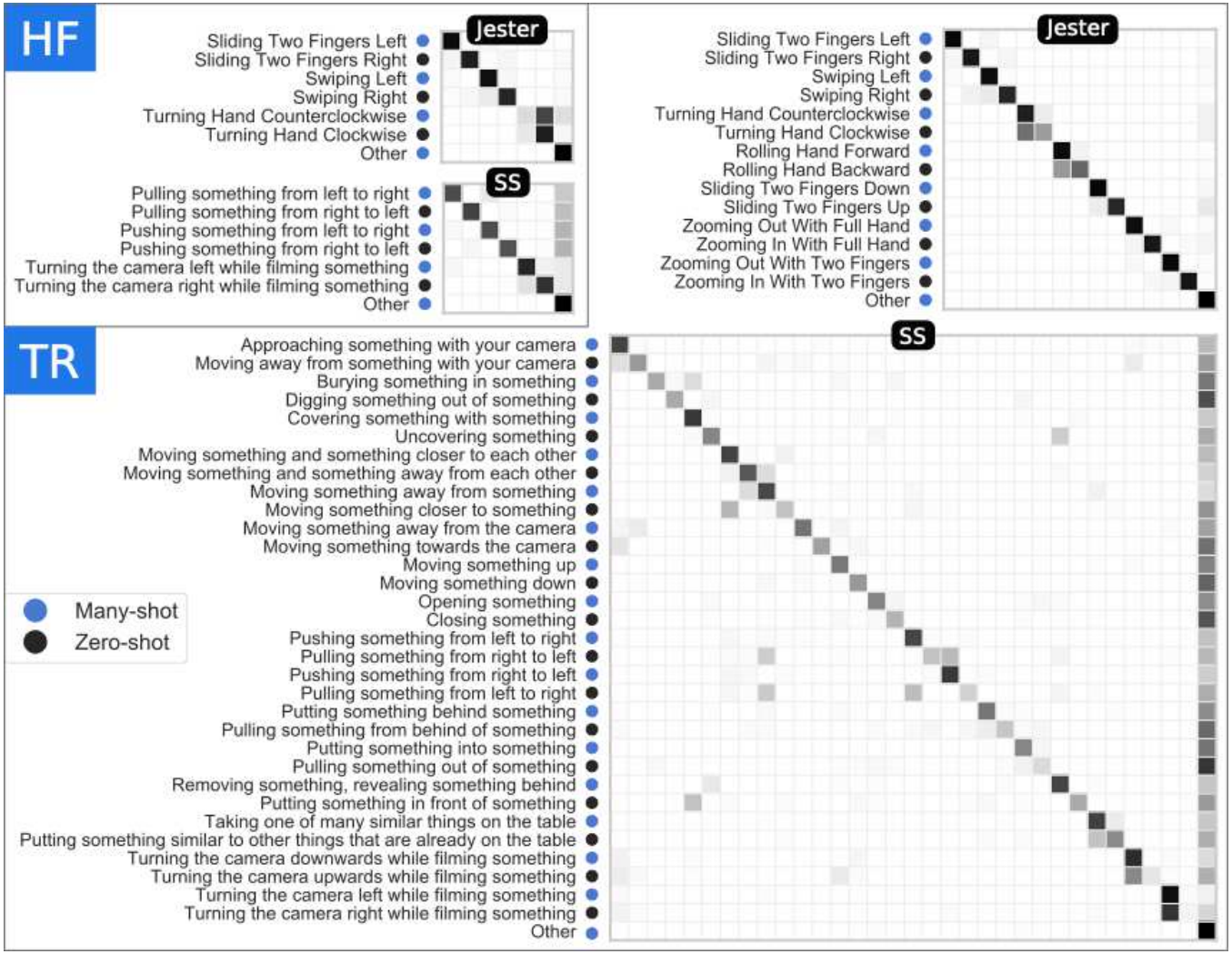


W Price, D Damen (2019). Retro-Actions: Learning 'Close' by Time-Reversing 'Open' Videos. ICCV MDALC Workshop

Retro-actions

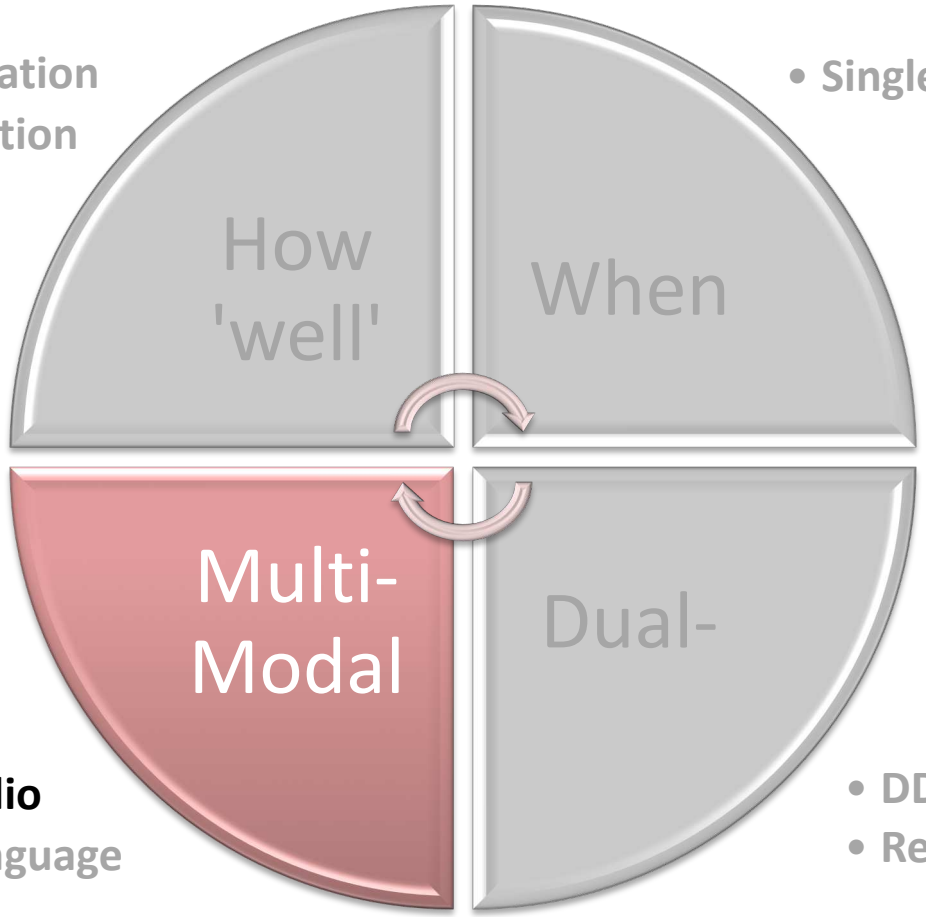


Retro-actions – Zero-Shot Learning



Fine-Grained Object Interactions

- Skill Determination
- Action Completion



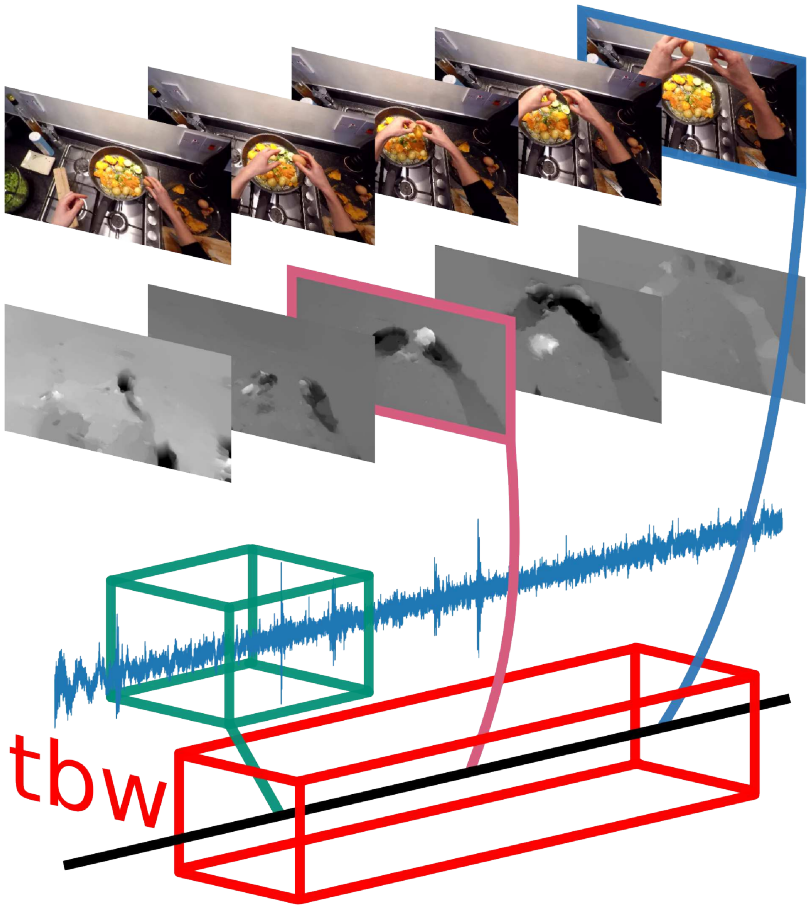
- Single-timestamp

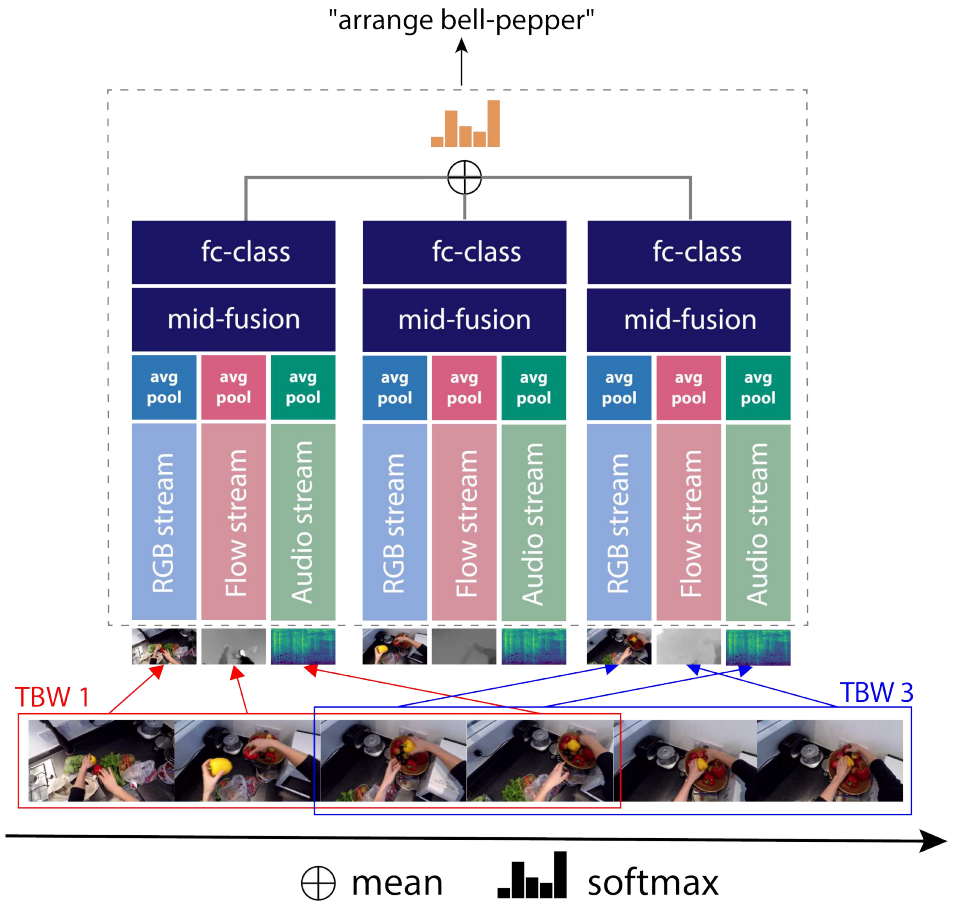
- **Vision+Audio**
- Vision+Language

- DDLSTM
- Retro-Actions

Audio-Visual Temporal Binding for Egocentric Action Recognition

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman

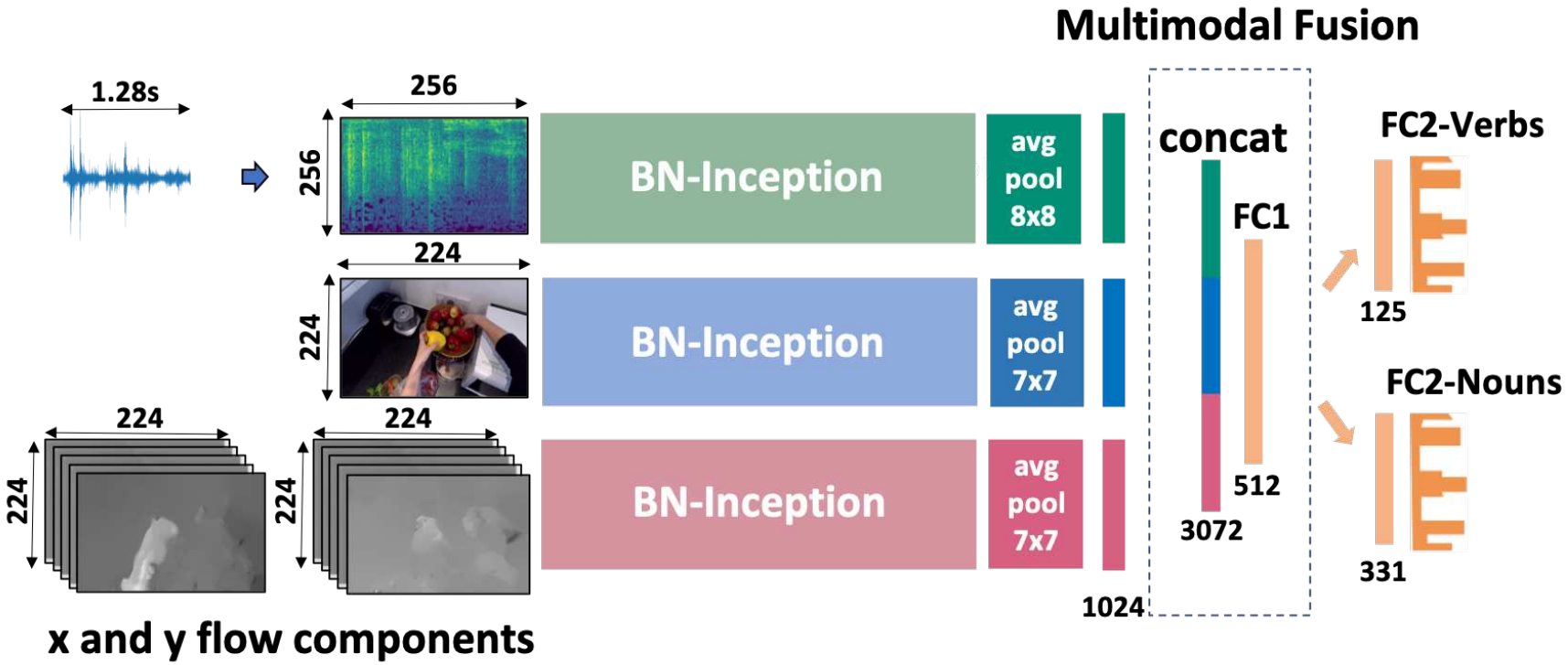




E Kazakos, A Nagrani, A Zisserman, D Damen (2019). EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition. ICCV

Audio-Visual Temporal Binding for Egocentric Action Recognition

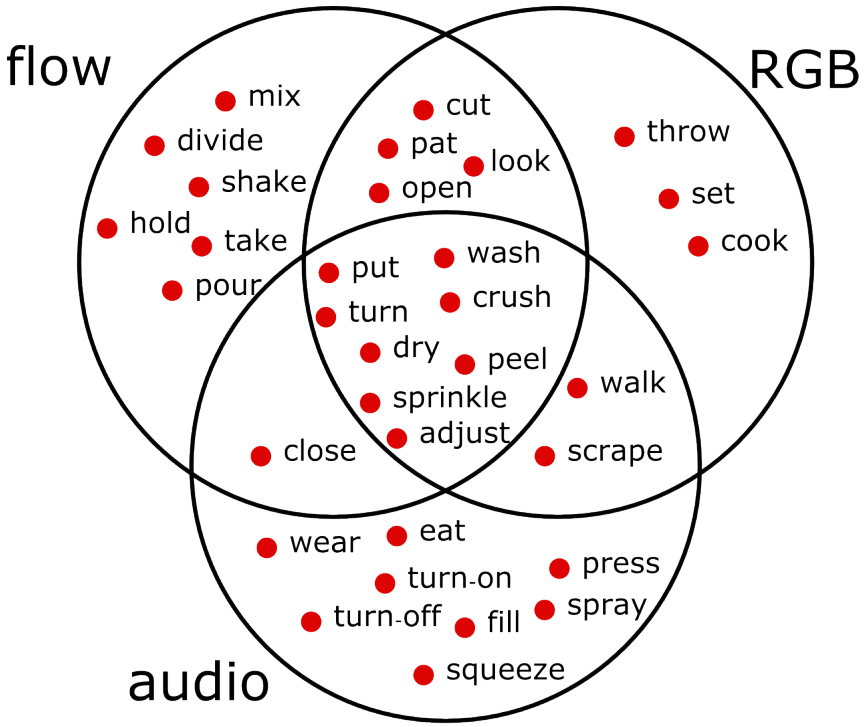
with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman



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

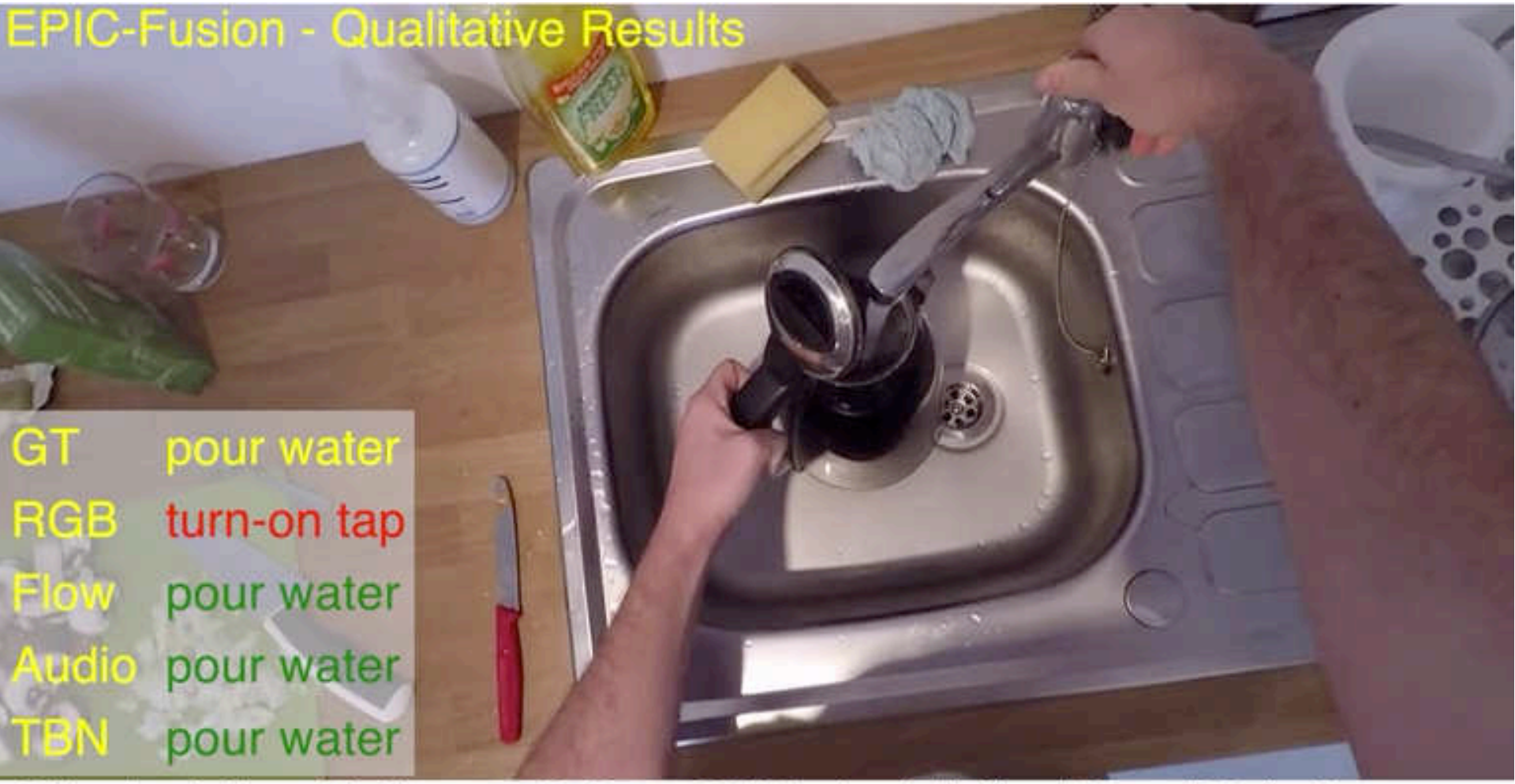
Audio-Visual Temporal Binding for Egocentric Action Recognition

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman



Audio-Visual Temporal Binding for Egocentric Action Recognition

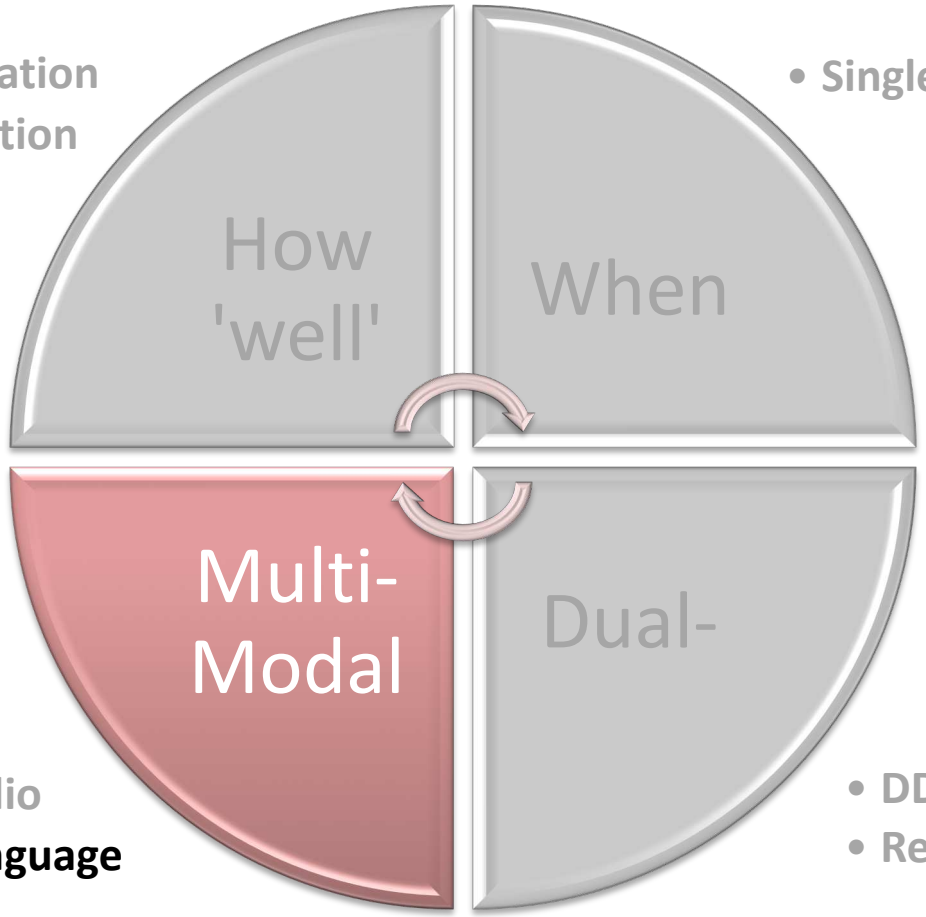
with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman



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

Fine-Grained Object Interactions

- Skill Determination
- Action Completion



- Single-timestamp

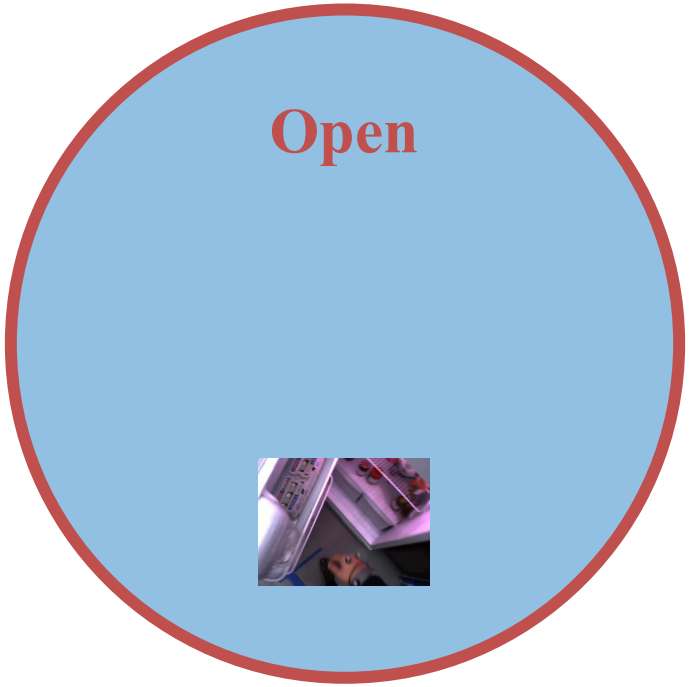
- Vision+Audio
- **Vision+Language**

- DDLSTM
- Retro-Actions

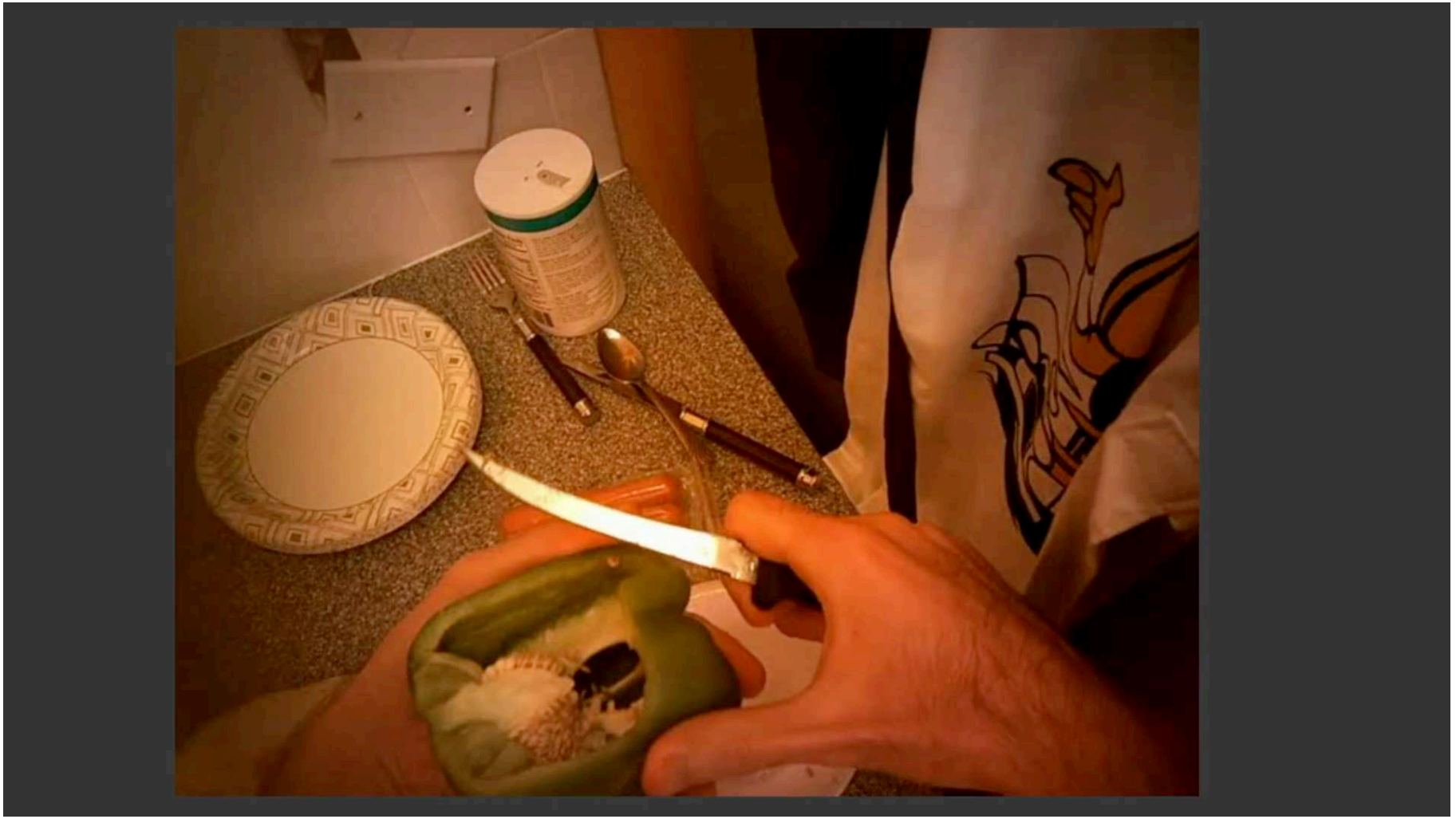
The *Verbs* Dilemma

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

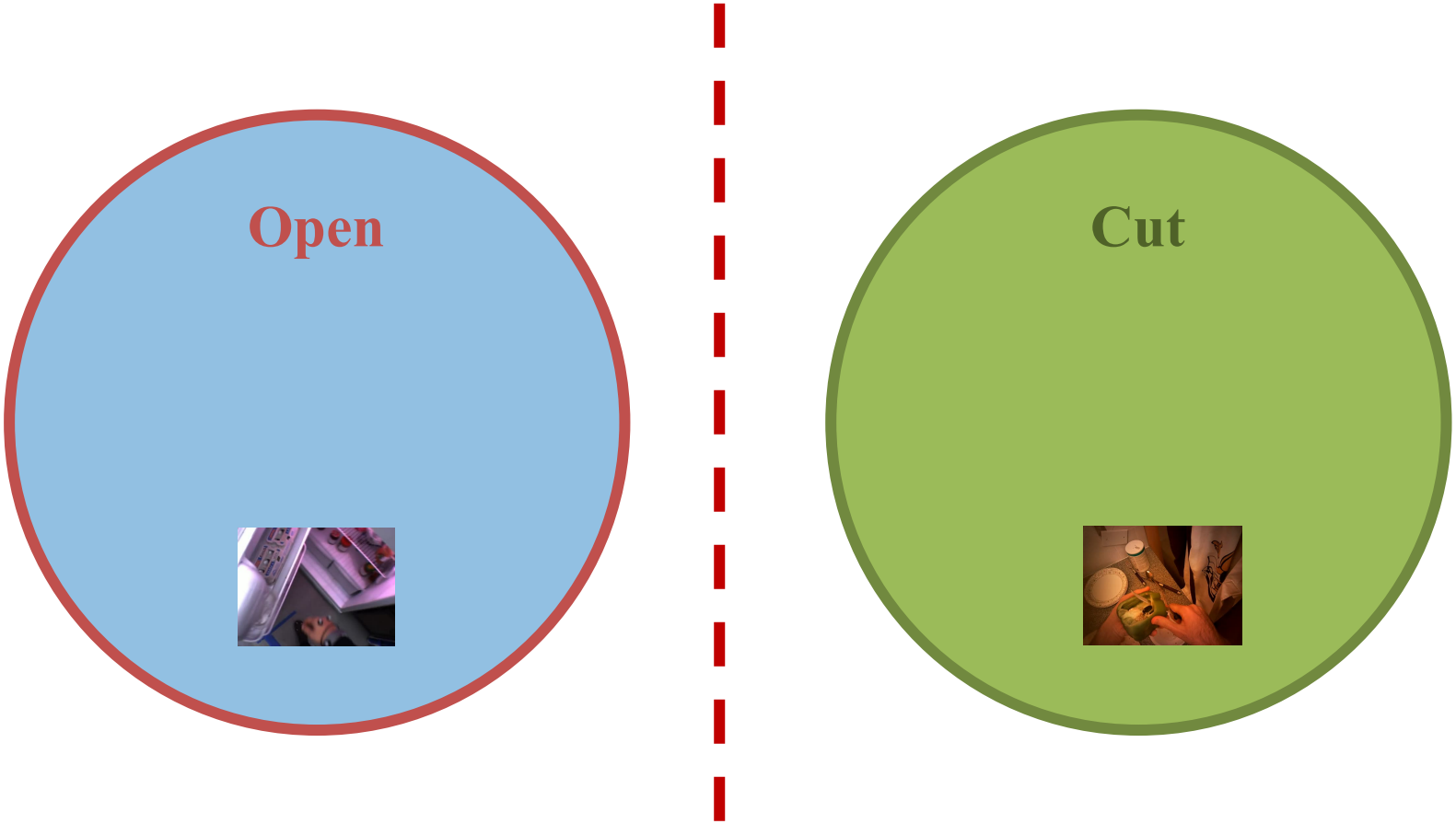
The *Verbs* Dilemma



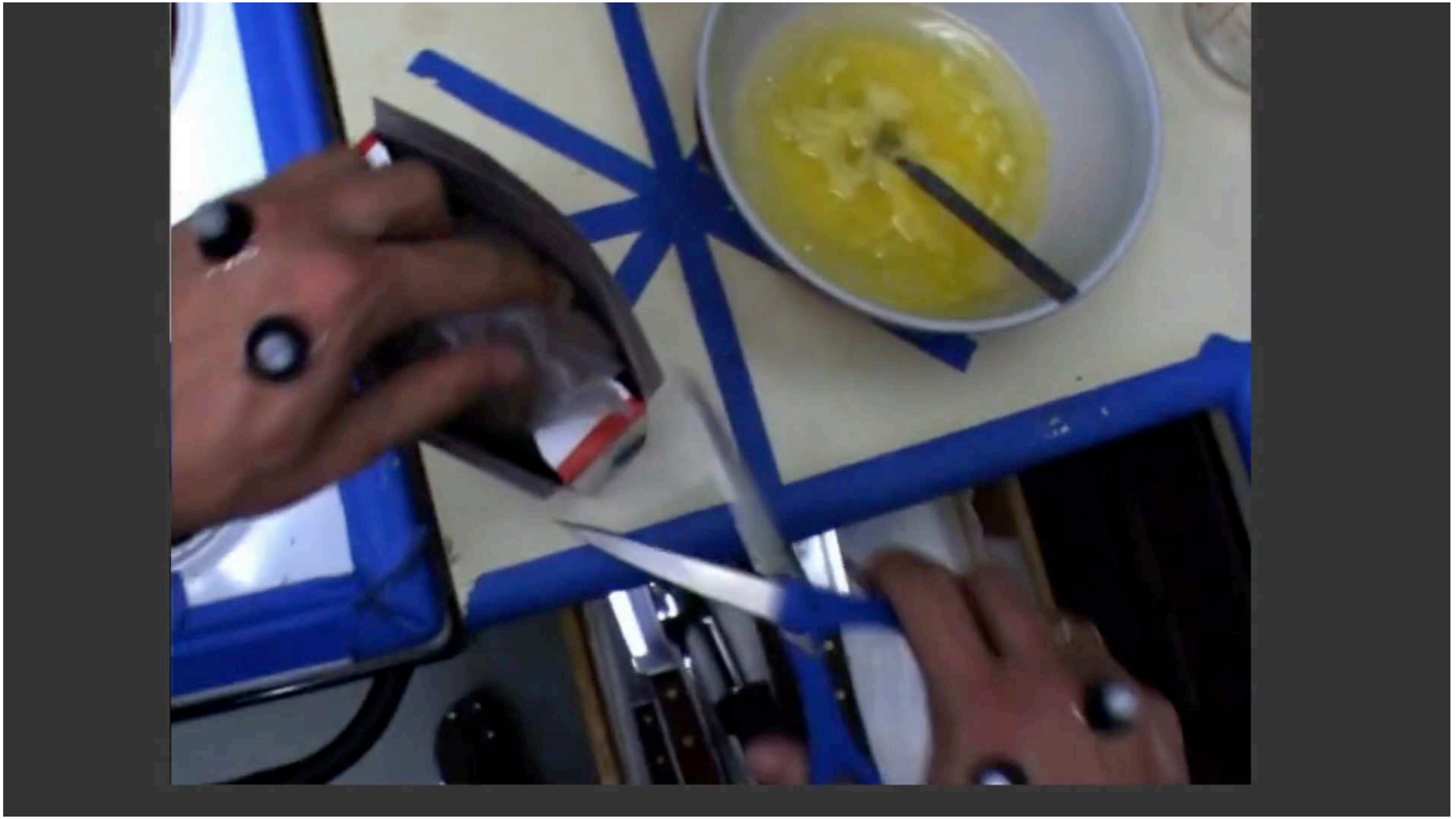
The *Verbs* Dilemma



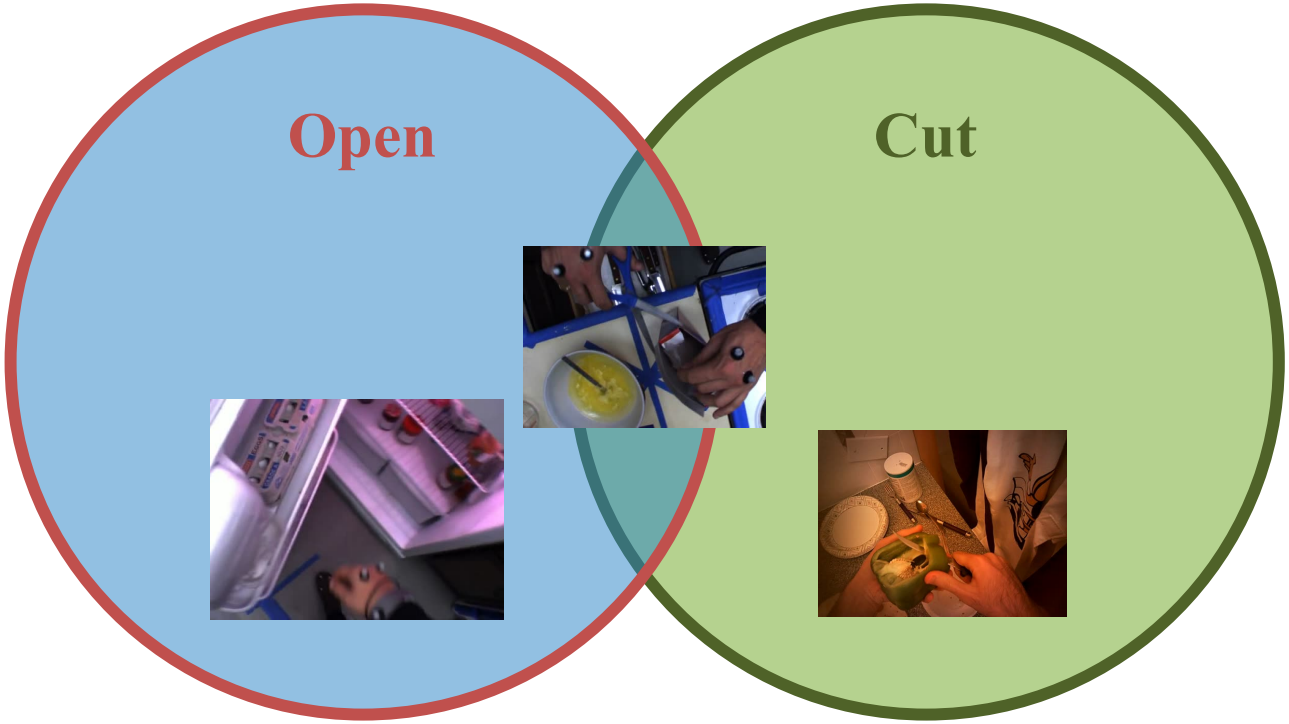
The *Verbs* Dilemma



The *Verbs* Dilemma



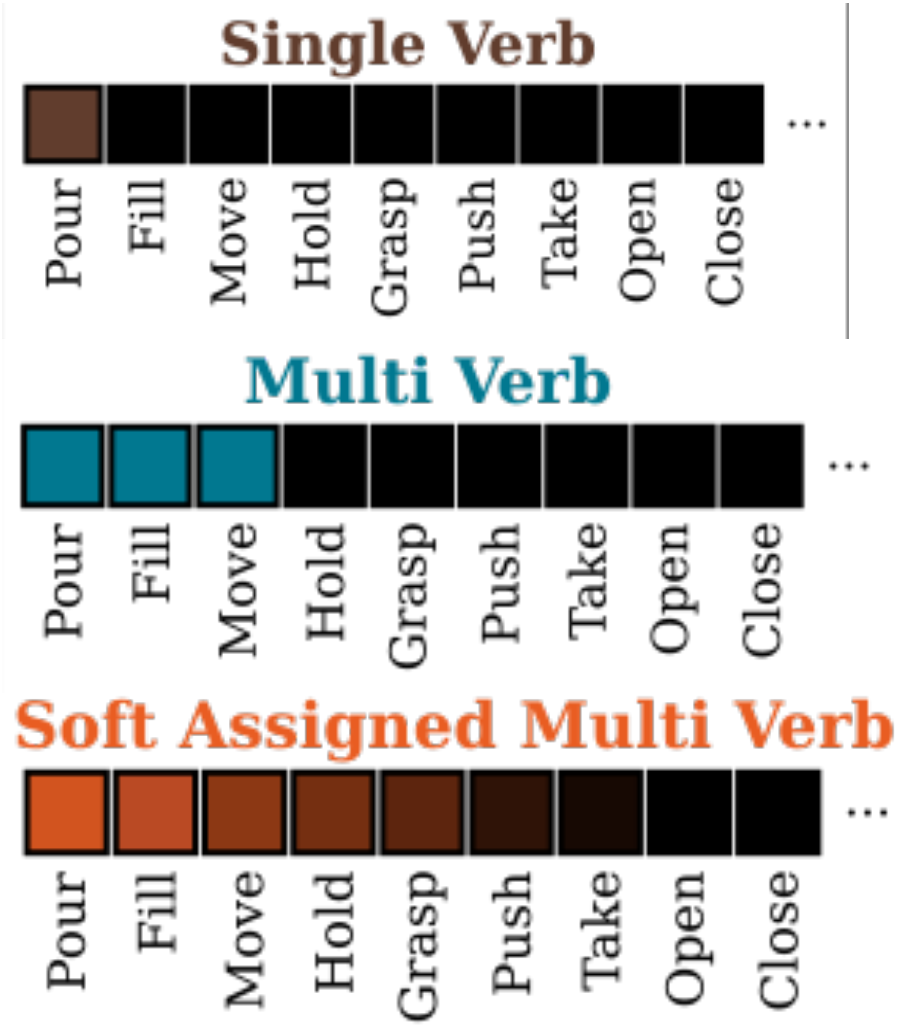
The *Verbs* Dilemma



The *Verbs* Dilemma

- 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

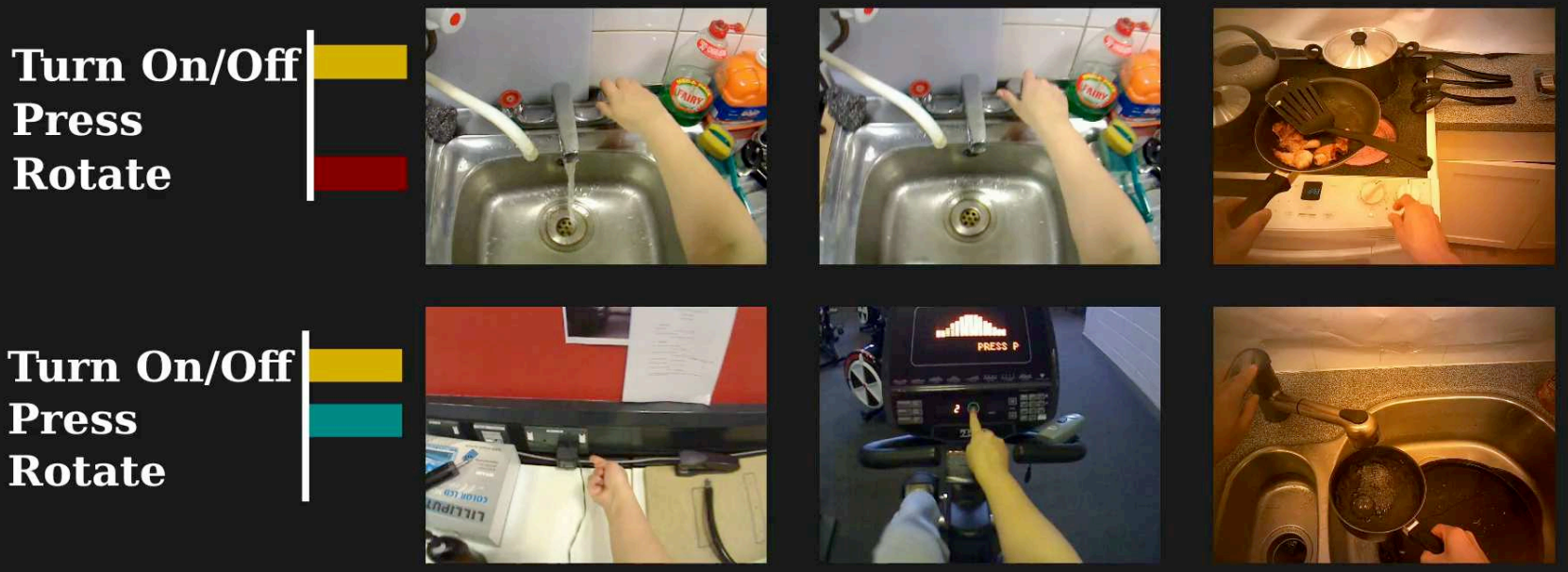
The *Verbs* Dilemma



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

The *Verbs* Dilemma

Top 3 retrieved classes across all datasets.



Labelling Method can differentiate turn On/Off tap by pressing and by rotating.

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

Fine-Grained Action Retrieval

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

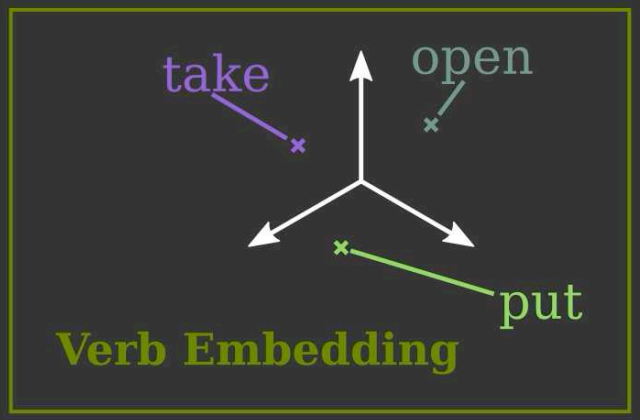
Fine-Grained Action Retrieval

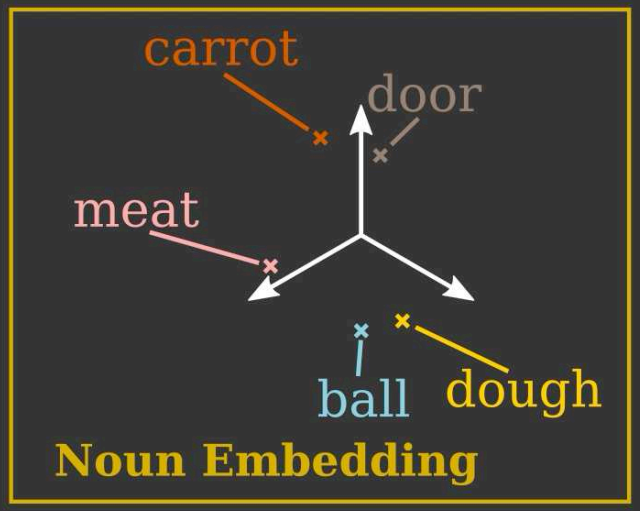
with: Michael Wray
Gabriela Csurka
Diane Larlus

We embed the video and representations

[put]

[meat, ball, dough]

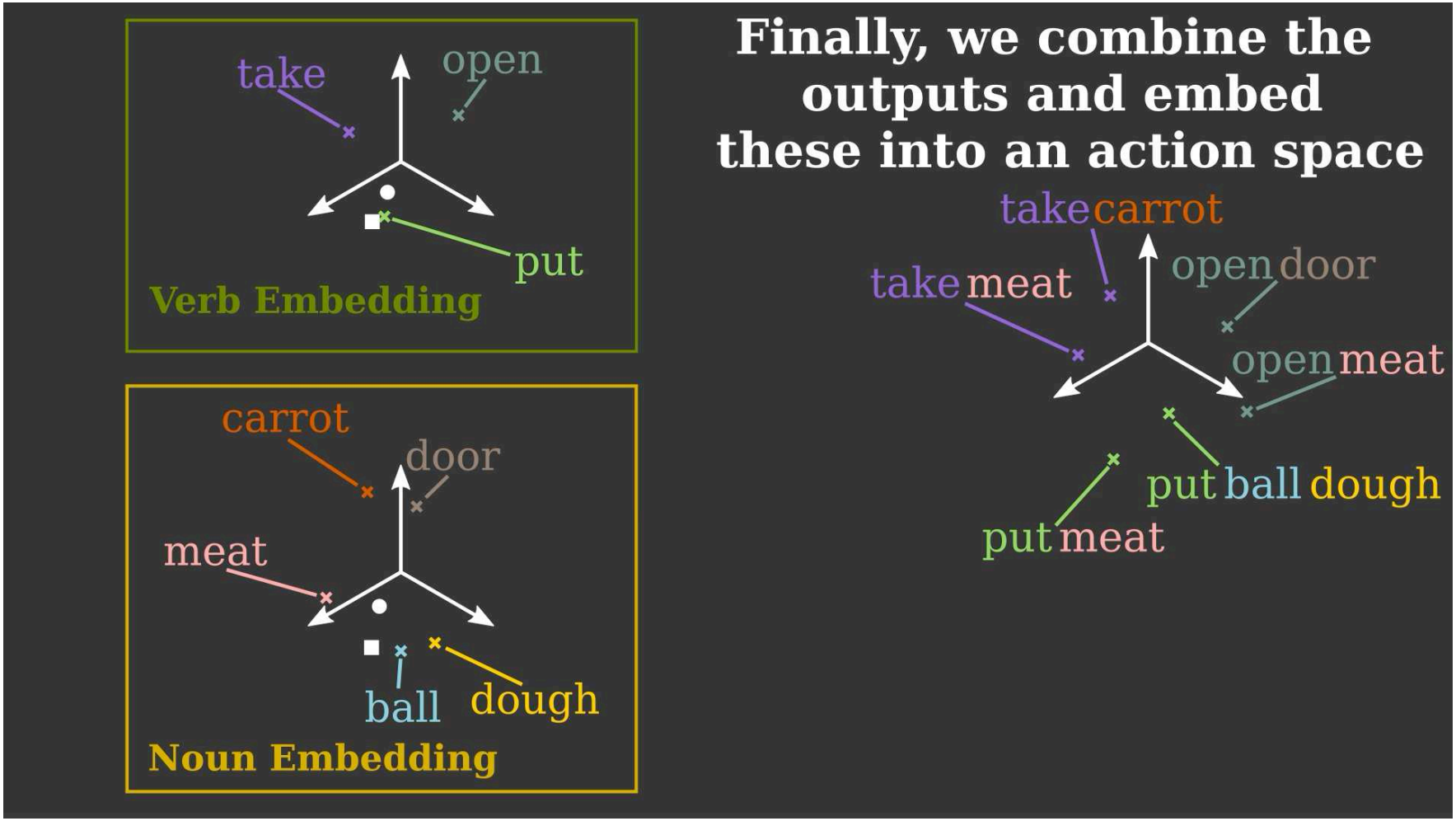




M Wray, D Larlus, G Csurka, D Damen (2019). Fine-Grained Action Retrieval through Multiple Parts-of-Speech Embeddings. ICCV

Fine-Grained Action Retrieval

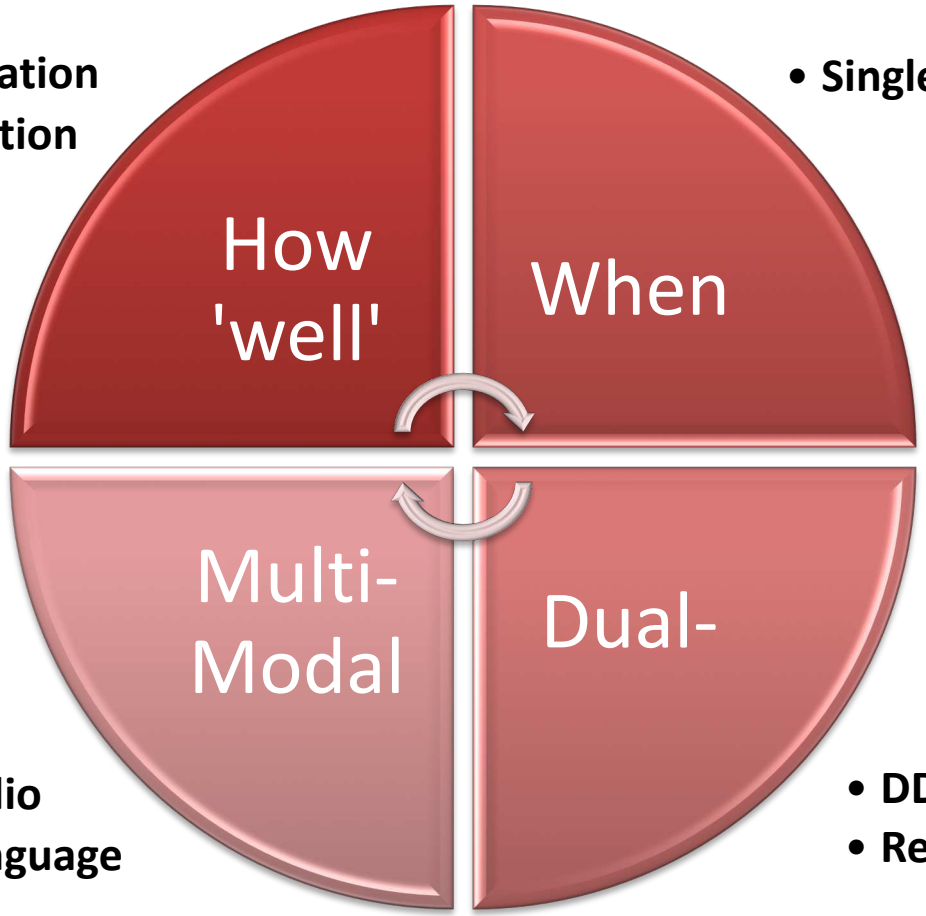
with: Michael Wray
Gabriela Csurka
Diane Larlus



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

Fine-Grained Object Interactions

- Skill Determination
- Action Completion



- Single-timestamp

- Vision+Audio
- Vision+Language

- DDLSTM
- Retro-Actions

Bristol and University of Bristol

Thank you…

For further info, datasets, code, publications…

http://dimadamen.github.io

@dimadamen

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

EPIC KITCHENS

University of BRISTOL | UNIVERSITY OF TORONTO | UNIVERSITÀ degli STUDI di CATANIA

Scaling Egocentric Vision: The EPIC-KITCHENS Dataset


Dima Damen


Hazel Doughty


Giovanni M. Farinella


Sanja Fidler


Antonino Furnari


Evangelos Kazakos


Davide Moltisanti


Jonathan Munro


Toby Perrett


Will Price

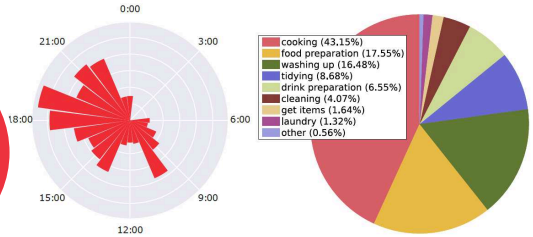

Michael Wray

Scaling Egocentric Vision

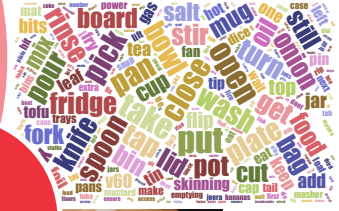
EPIC KITCHENS

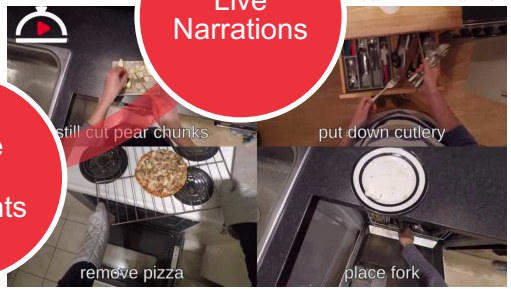
Data Collection

Native Environment, Natural Interactions



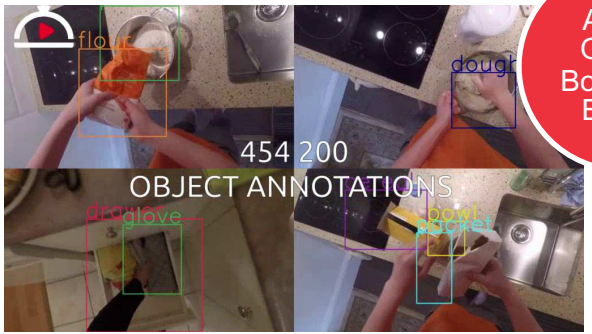
Live Narrations





Dense Action Segments

Active Object Bounding Boxes



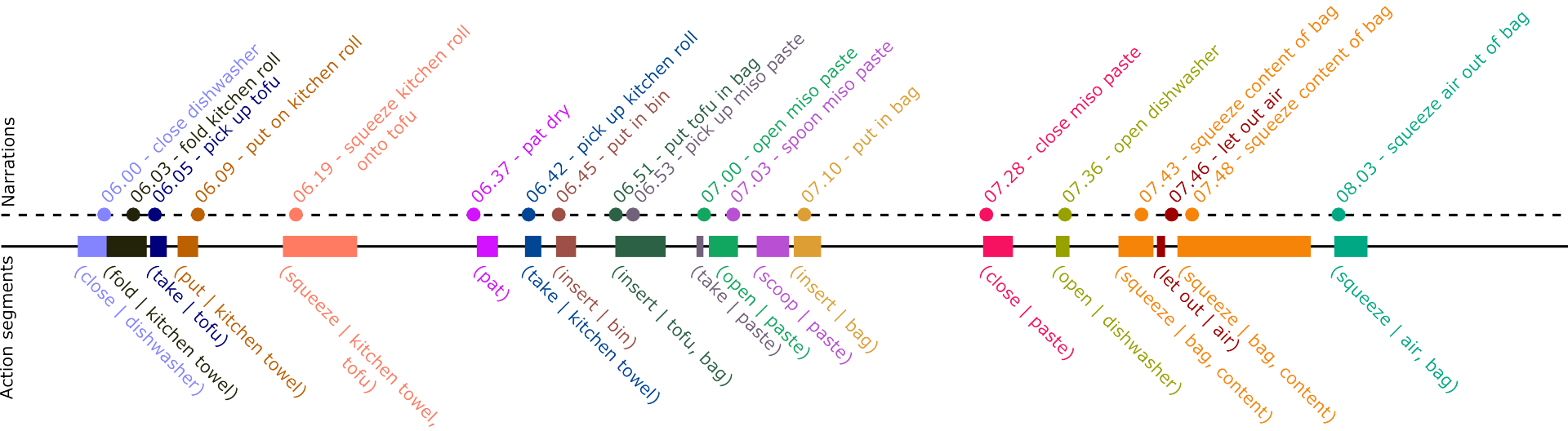
Benchmark and Challenges

UNIVERSITY OF TORONTO

University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA

Narrations to Action Segments

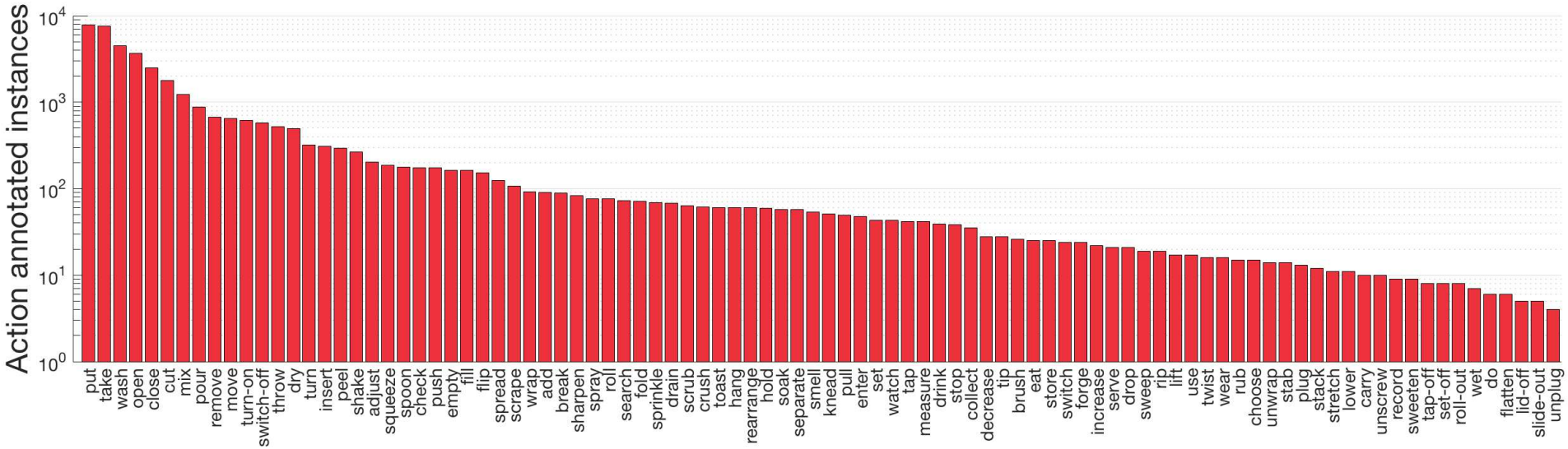


UNIVERSITY OF TORONTO

University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA

Annotations Statistics

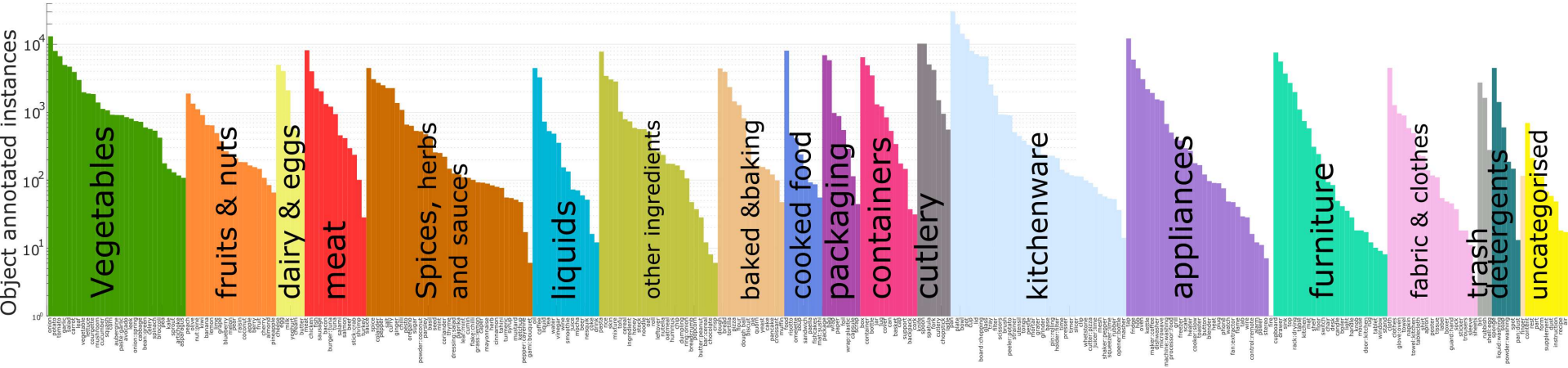


UNIVERSITY OF TORONTO

University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA

Annotations Statistics

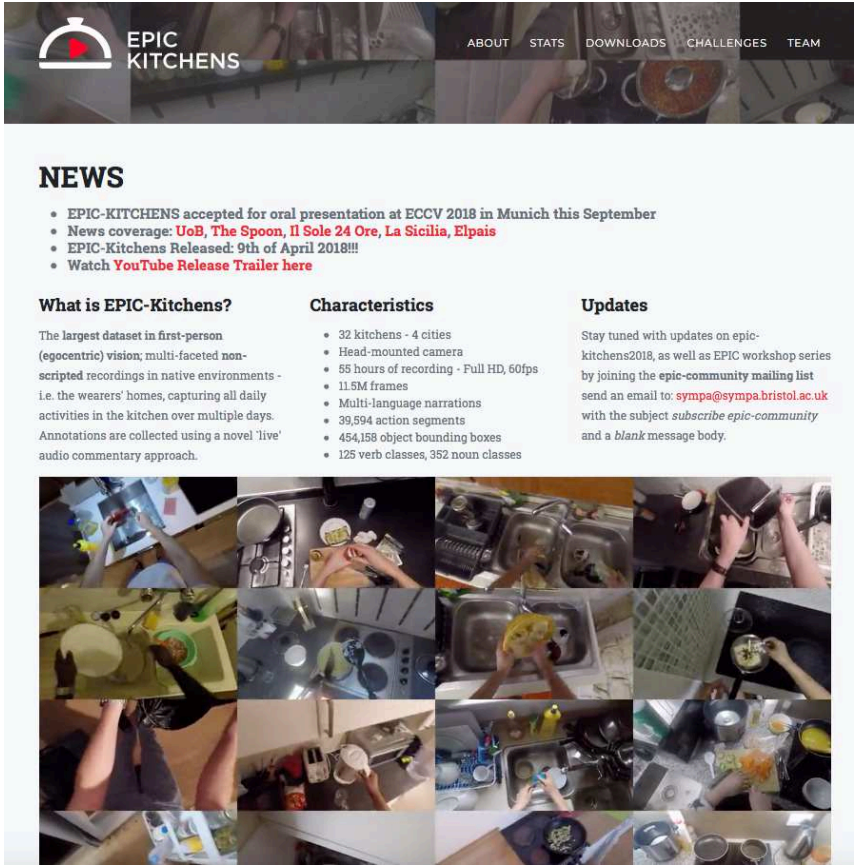


39 000
ACTION SEGMENTS

454 200
OBJECT ANNOTATIONS

More?

http://epic-kitchens.github.io



UNIVERSITY OF TORONTO

University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA