Katha AI

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

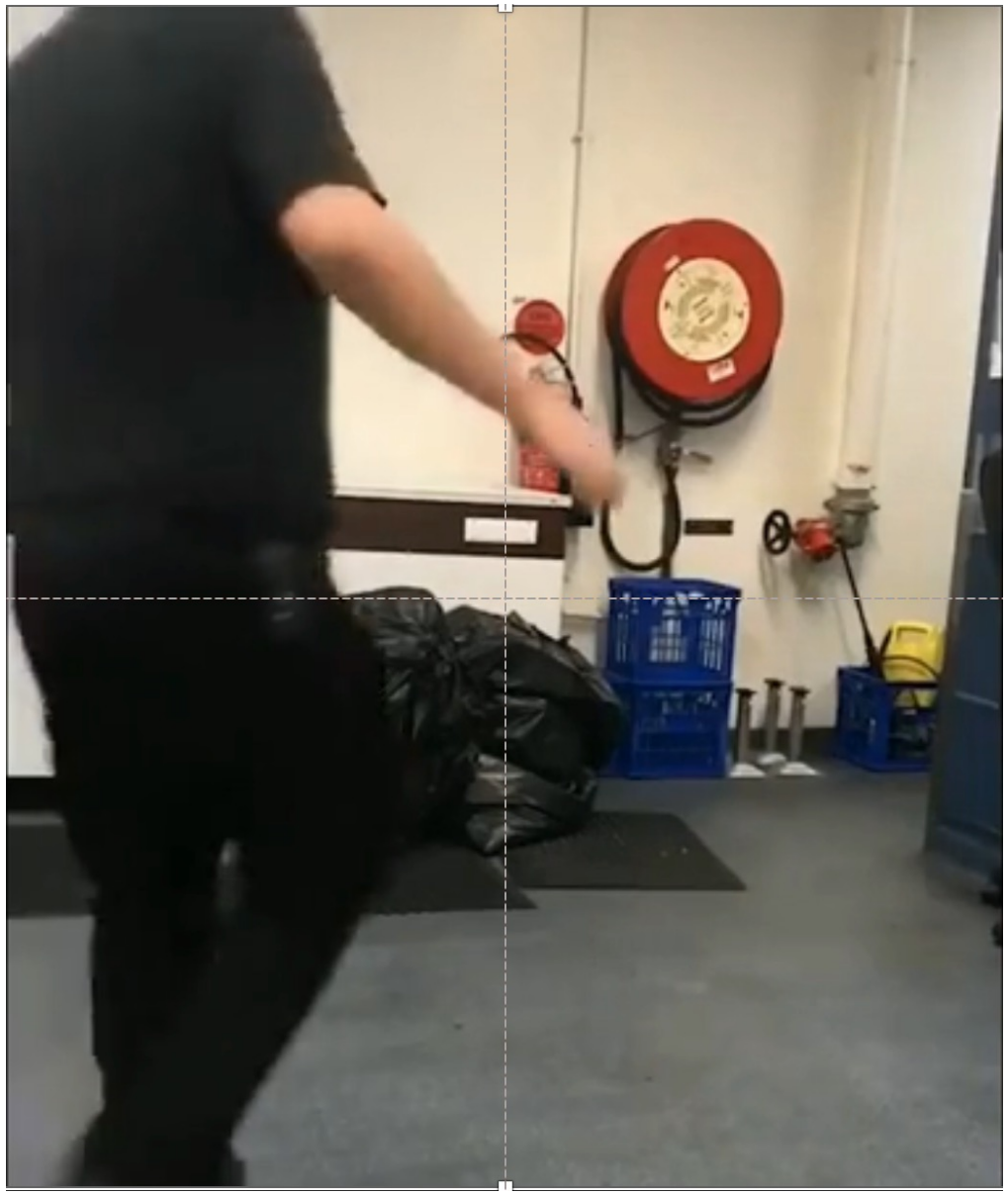
IIIT Hyderabad

Video Understanding ❤️

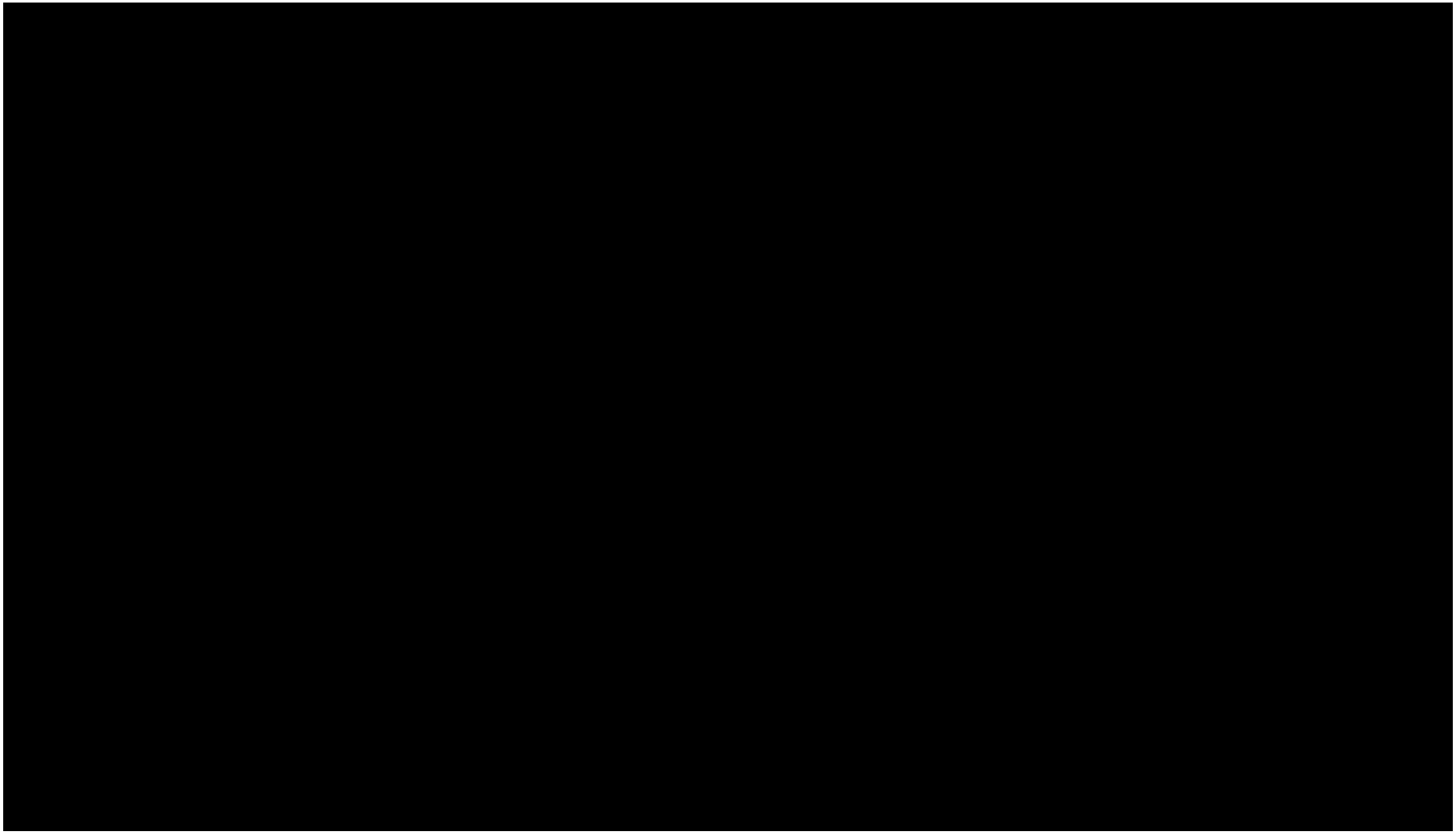
Makarand Tapaswi

CS7.505 Spring 2024

13th April 2024

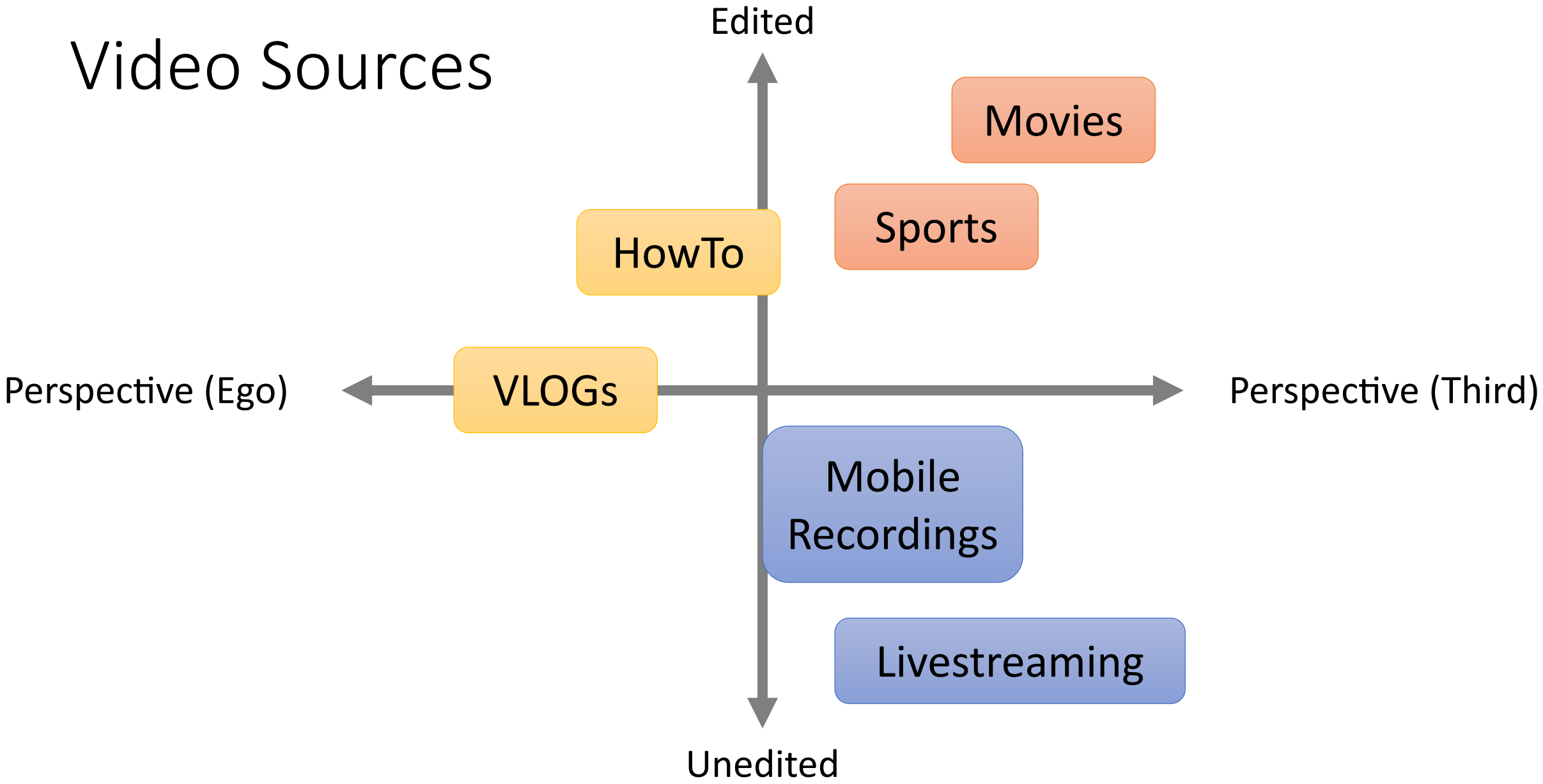


Epstein, et al. Oops! Predicting Unintentional Action in Video. CVPR 2020



Gupta, et al. Photorealistic Video Generation with Diffusion Models. arXiv: 2312.06662.
Tweet Credit: https://twitter.com/agrimgupta92/status/1734253883076063426

Video Sources



From Dima Damen's talks (U Bristol)

Outline

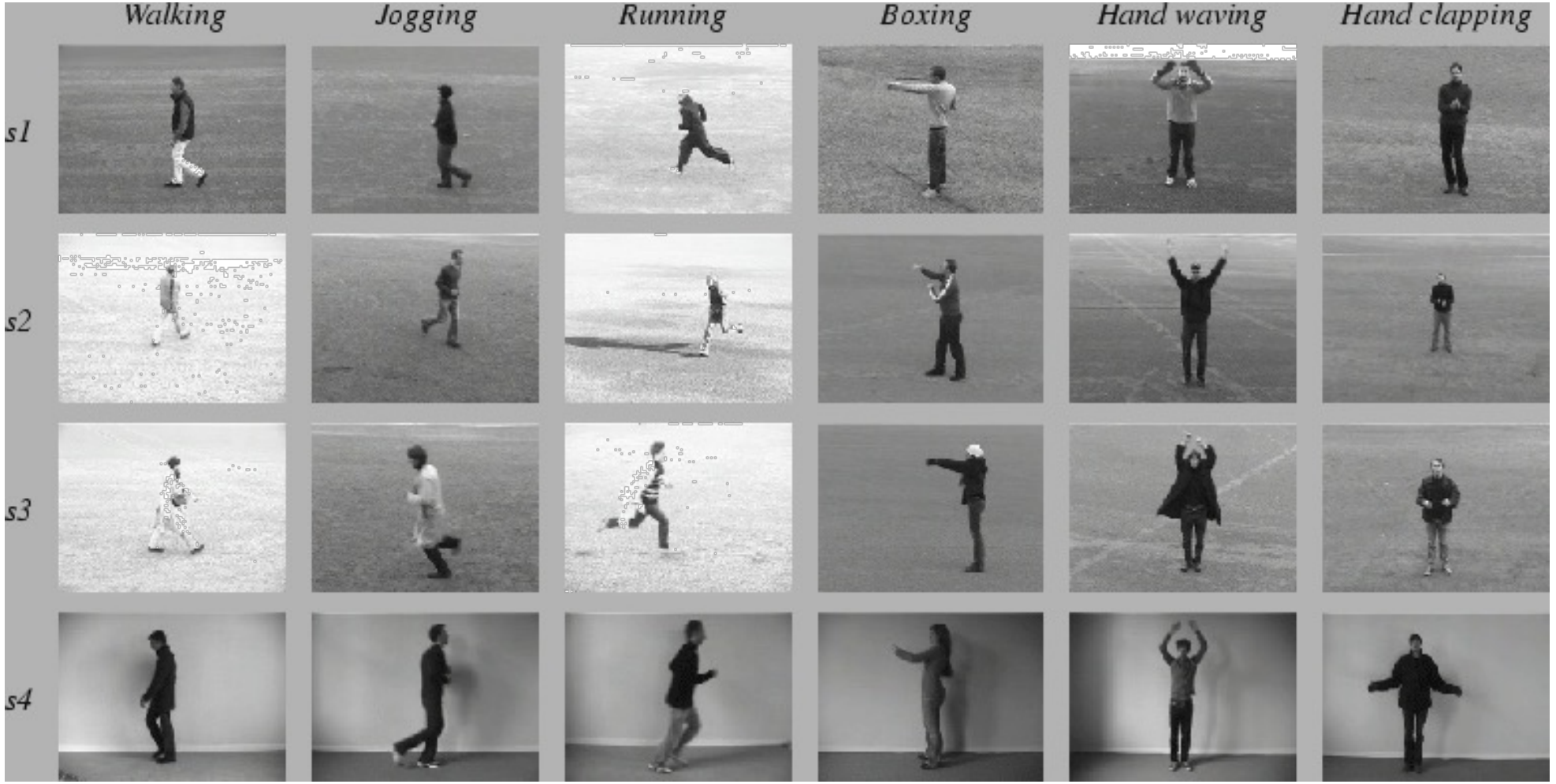
- Part 1: Popular video understanding tasks
- Part 2: Video backbones

(break)

- Part 3: Video-language models
- Part 4: Open challenges

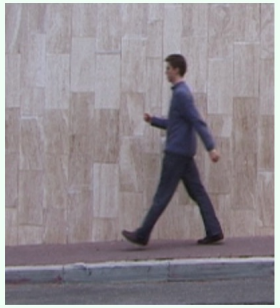
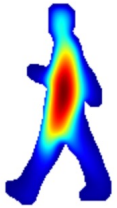
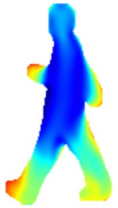
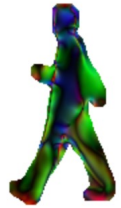
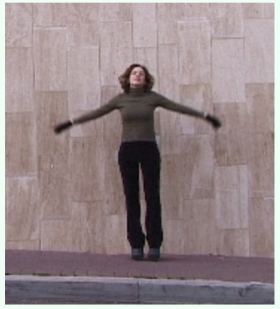
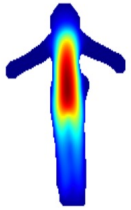
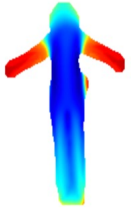
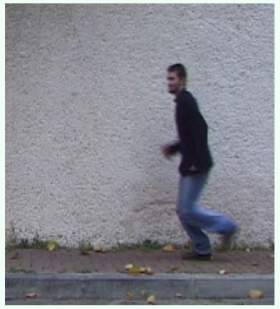
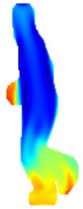
Part 1
Video Understanding

Golden years of actions: KTH Dataset



Laptev and Lindeberg. Space-Time Interest Points. ICCV 2003

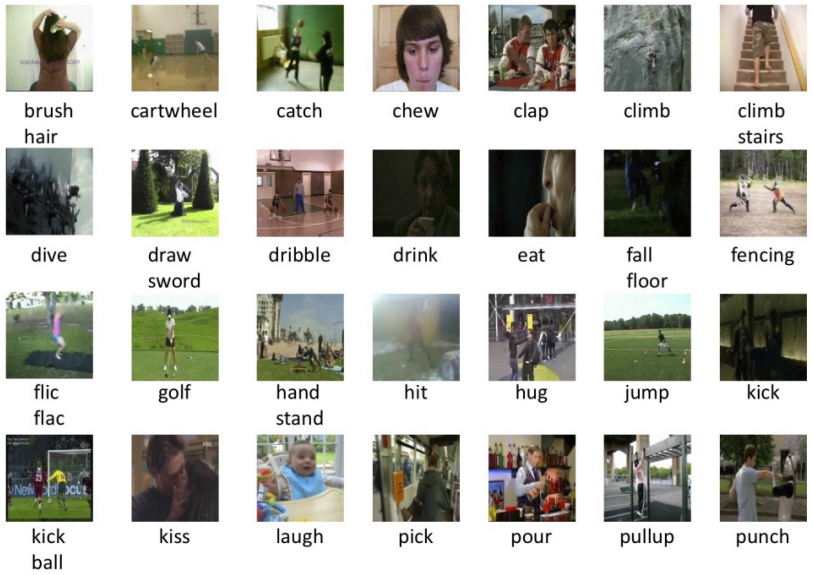
Golden years of actions: Weizmann dataset

Input Sequence	Foreground mask	Solution of Poisson eq.	Space-Time "Saliency"	Measure of "Plateness"	Measure of "Stickness"
					
					
					

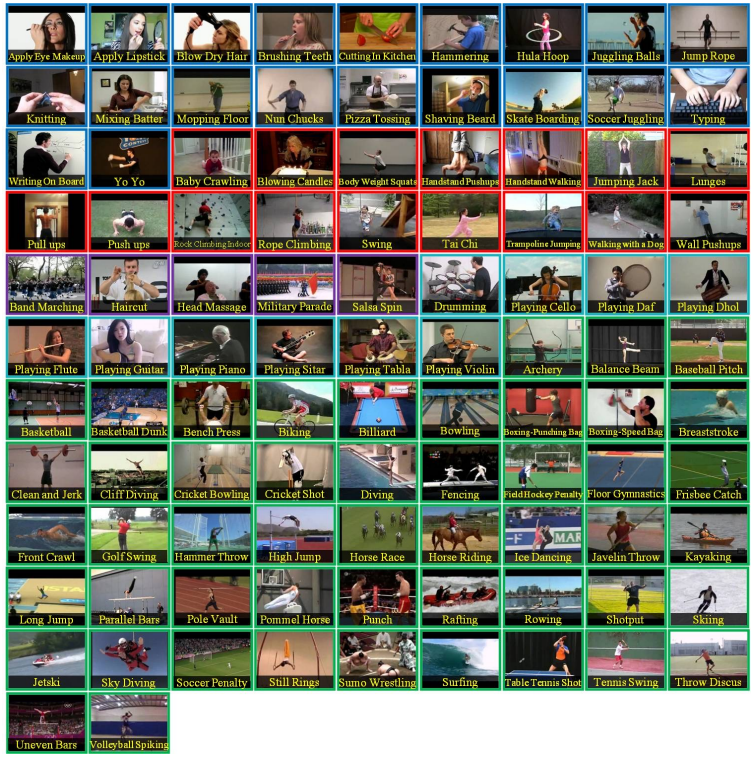
Gorelick, et al. Actions as Space-Time Shapes. ICCV 2005

Actions datasets: sub-100 classes

HMDB (51)



UCF (101)

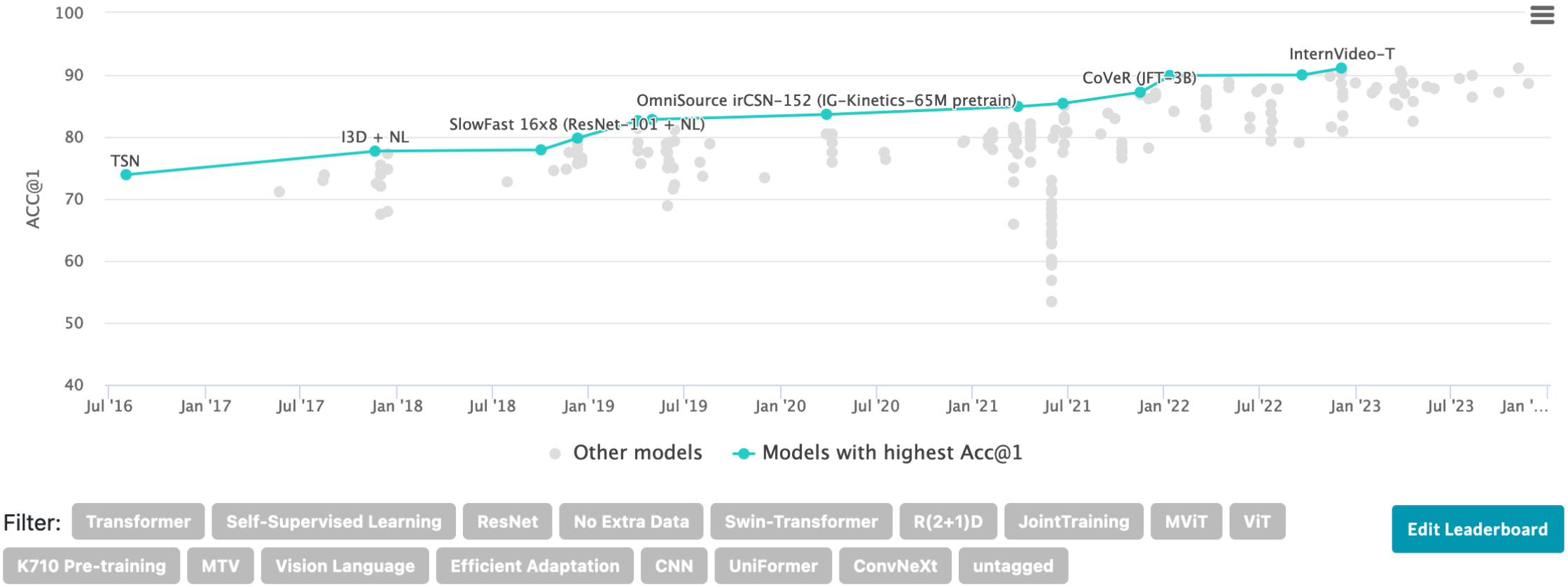


Something-Sth. (174)

Kuehne, et al. **HMDB**: A Large Video Database for Human Motion Recognition. ICCV 2011
Soomro, et al. **UCF101**: A Dataset of 101 Human Action Classes from Videos in the Wild. CRCV-TR 2012
Goyal, et al. The **something something** video database for learning and evaluating visual common sense. ICCV 2017

Is there an ImageNet for videos?

Kinetics dataset

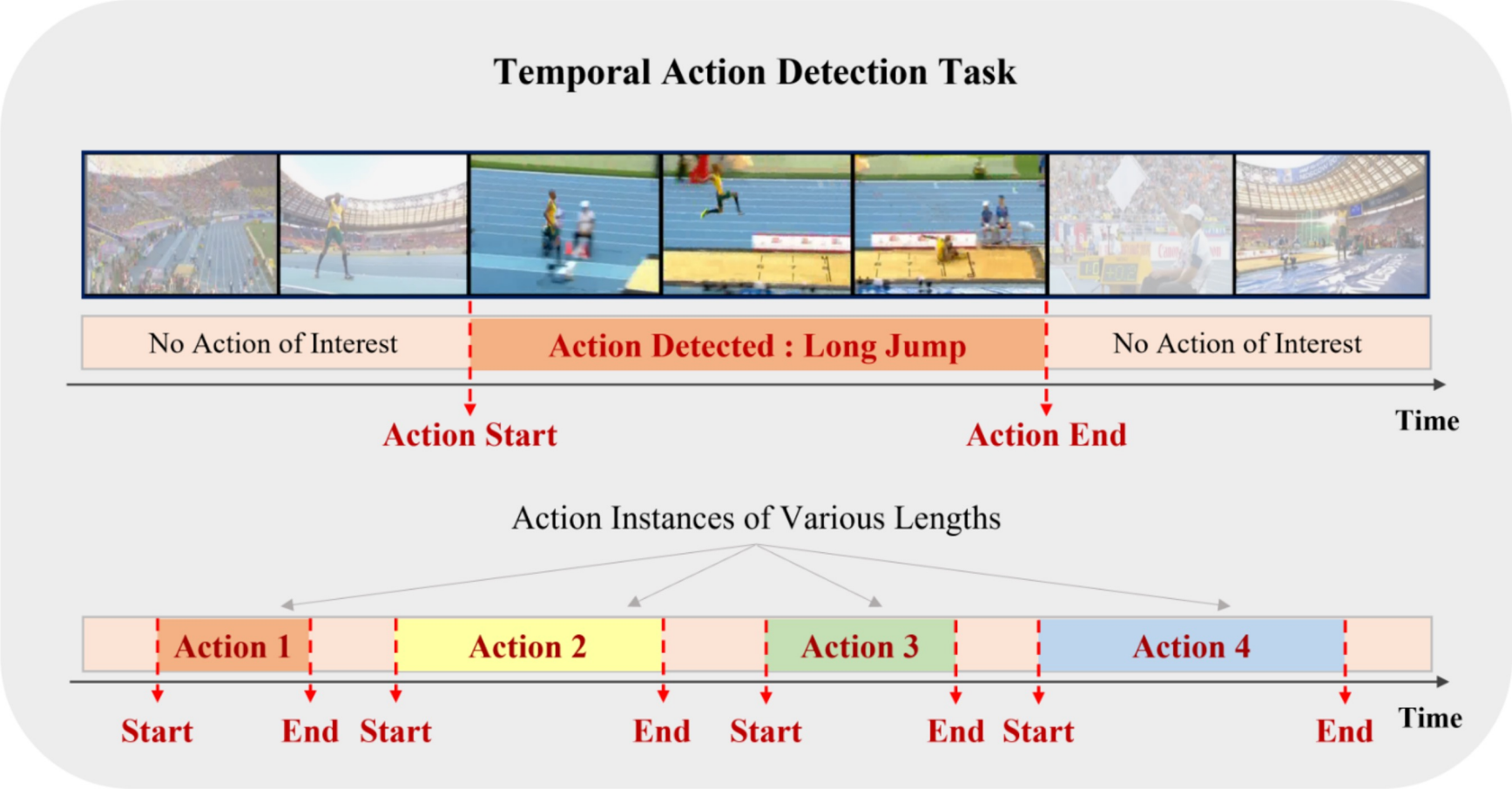


Carreira and Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. CVPR 2017
Image credit: https://paperswithcode.com/sota/action-classification-on-kinetics-400

Challenges of action understanding

- who is doing the action?
- when does an action start?
- how long is the action?
- how to deal with actions vs. interactions?
- what are essential components of any action?
- what is the importance of the background scene?

Temporal Action Localization

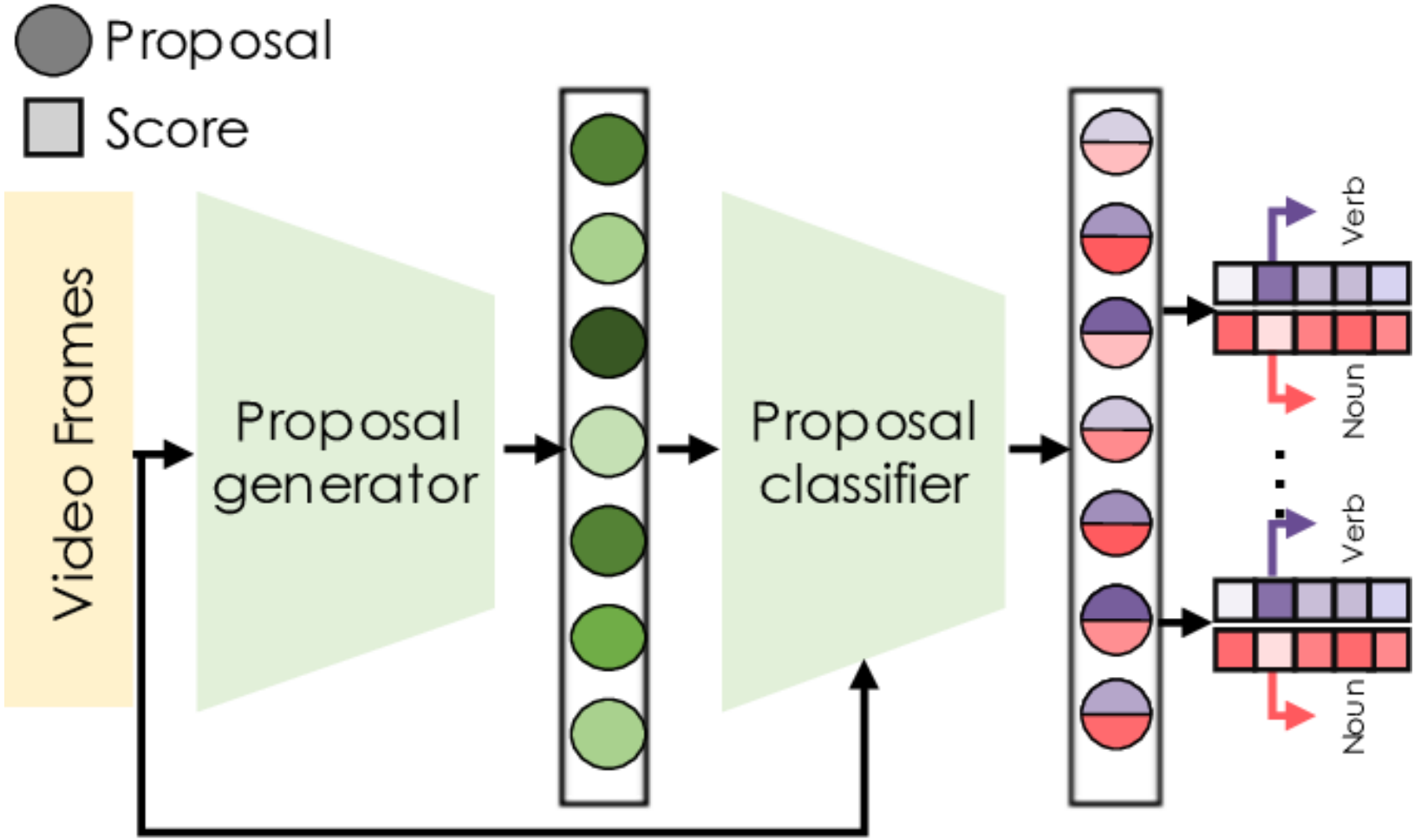


Vahdani and Tian. Deep Learning-based Action Detection in Untrimmed Videos: A Survey. PAMI 2022.

(Detour) Audio-visual Action Localization

Ramazanova, et al. OWL (Observe, Watch, Listen): Audiovisual Temporal Context for Localizing Actions in Egocentric Videos. CVPR 2023W

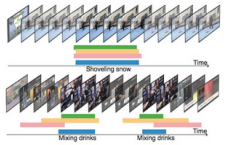
Temporal Action Localization



Ramazanova, et al. OWL (Observe, Watch, Listen): Audiovisual Temporal Context for Localizing Actions in Egocentric Videos. CVPR 2023W

ACTIVITYNET

Large Scale Activity Recognition Challenge



ActivityNet Temporal Action Localization

This task is intended to evaluate the ability of algorithms to **temporally localize activities in untrimmed video sequences**. Here, videos can contain more than one activity instance, and mutiple activity categories can appear in the video.

ActivityNet Event Dense-Captioning

This task involves both **detecting and describing events** in a video. For this task, participants will use the ActivityNet Captions dataset, a new large-scale benchmark for dense-captioning events.

AVA-Kinetics & Active Speakers

This challenge addresses two fundamental problems for spatio-temporal video understanding: (i) localize actions extents in space and time, and (ii) densely detect active speakers in video sequences.

DETAILS

Atomic Visual Actions

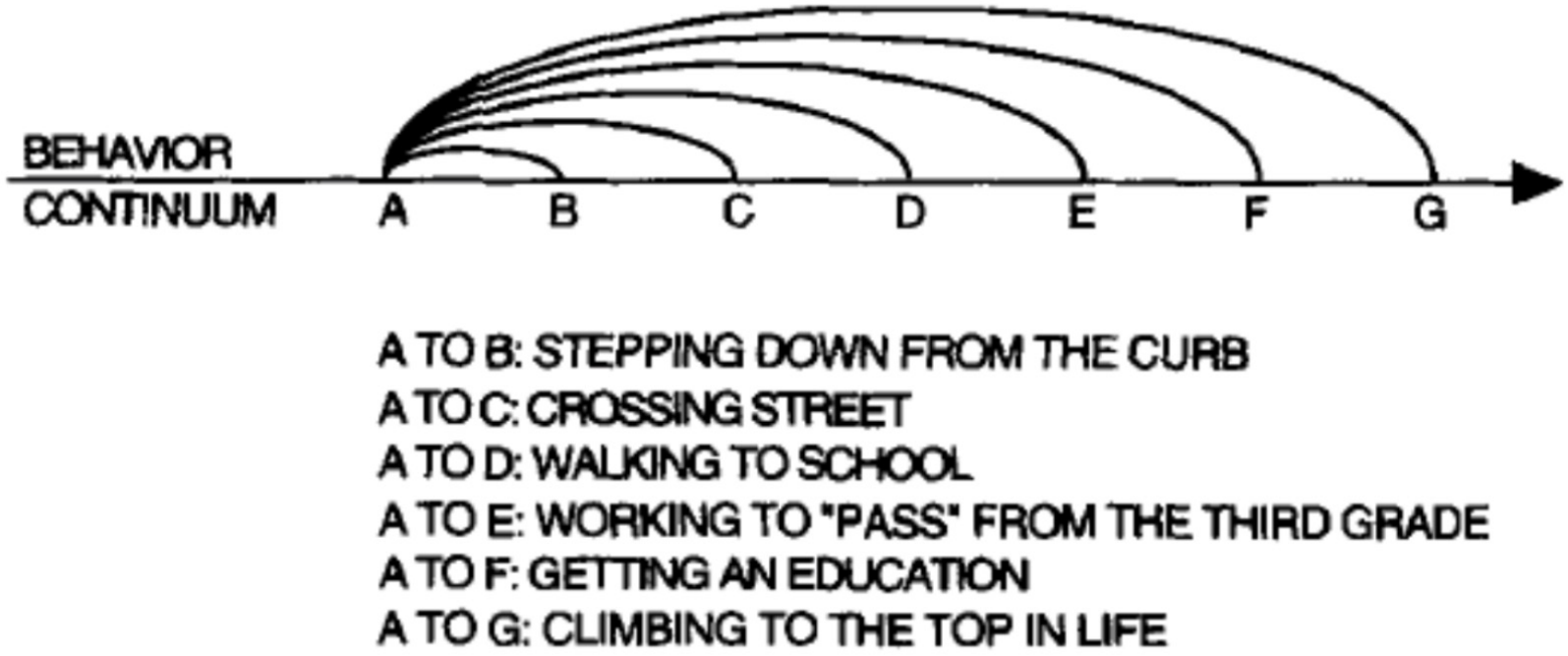
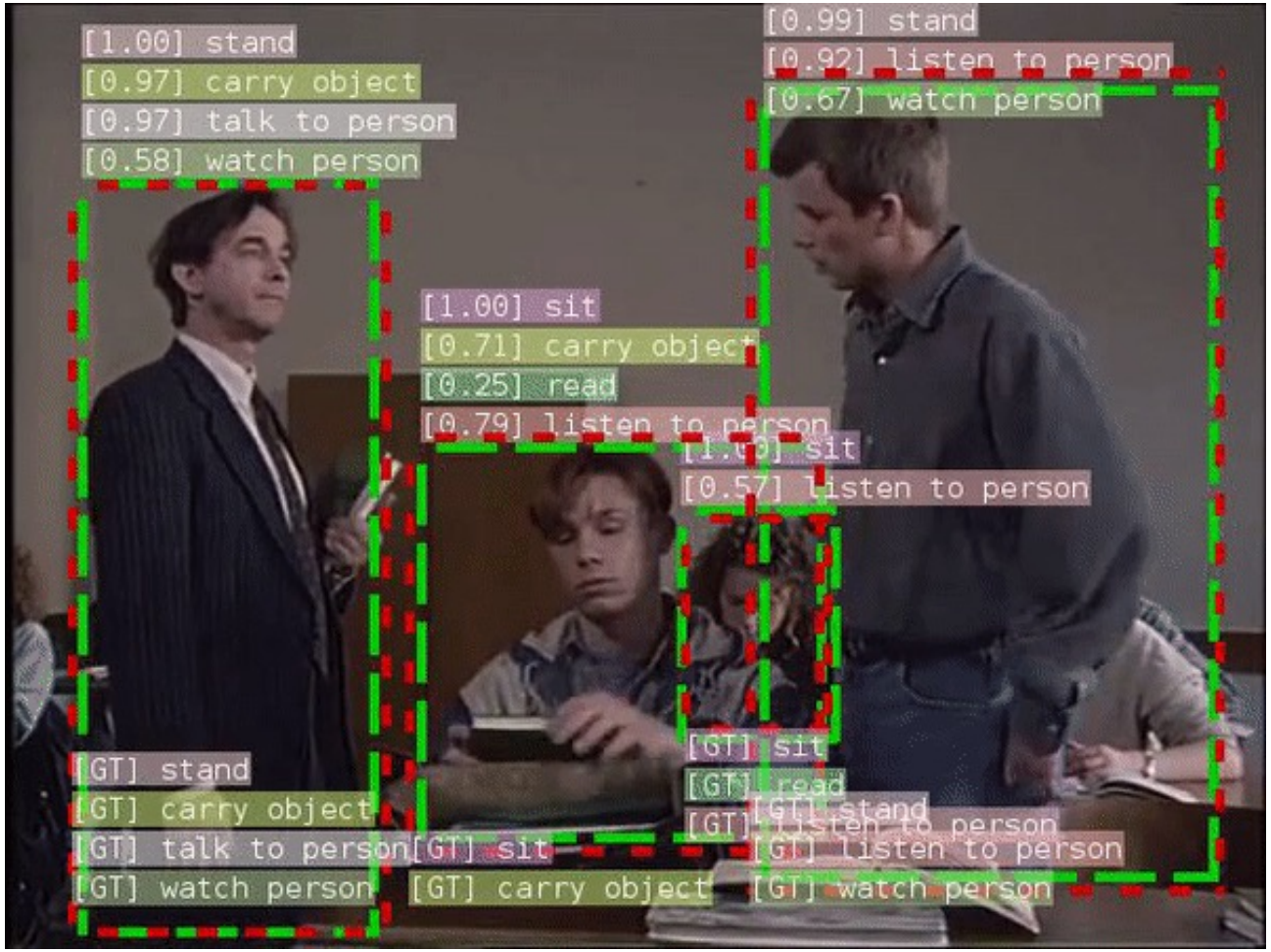


Figure 2. This figure illustrates the hierarchical nature of an activity. From Barker and Wright [3], pg. 247.

Gu, et al. AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions. CVPR 2018

Atomic Visual Actions

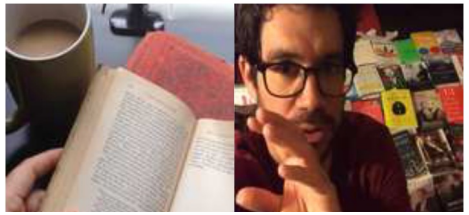


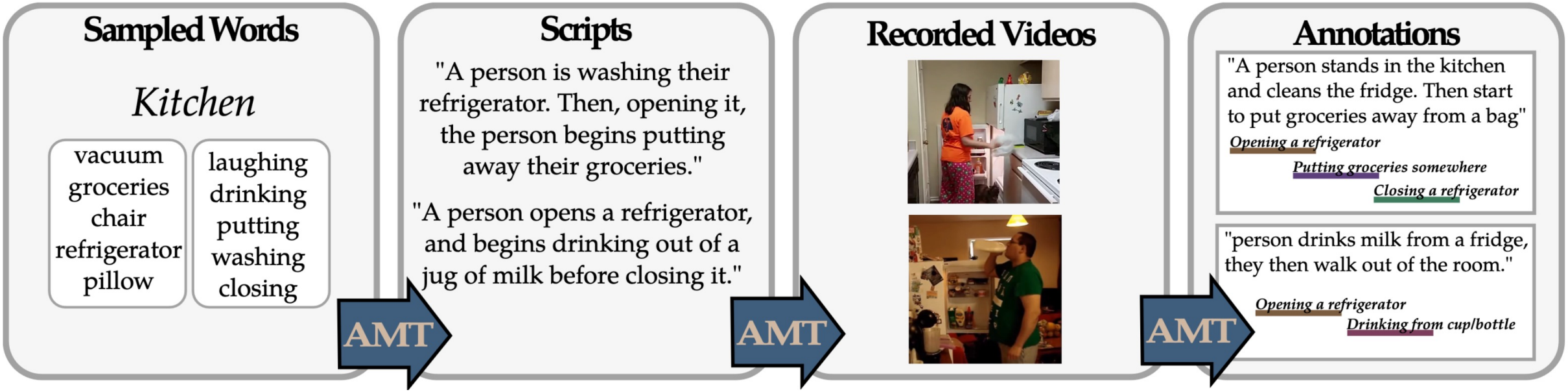
- Pose
- Person-Object
- Person-Person

Gu, et al. AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions. CVPR 2018

Charades dataset

Hollywood in Homes: Crowdsourcing Data Collection





Sigurdsson, et al. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. ECCV 2016

Why use a single label?

Let’s be descriptive …

Image captioning

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

A horse carrying a large load of hay and two people sitting on it.

Bunk bed with a narrow shelf sitting underneath it.

Fang, et al. From Captions to Visual Concepts and Back. CVPR 2015

Video captioning

MSR-VTT

1. A black and white horse runs around.
2. A horse galloping through an open field.
3. A horse is running around in green lush grass.
4. There is a horse running on the grassland.
5. A horse is riding in the grass.

Audio Descriptions (LSMDC)

Mike leans over and sees how high they are.

Abby clasps her hands around his face and kisses him passionately.

Xu, et al. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. CVPR 2016
Rohrbach, et al. Movie Description. IJCV 2017

Text-to-video retrieval

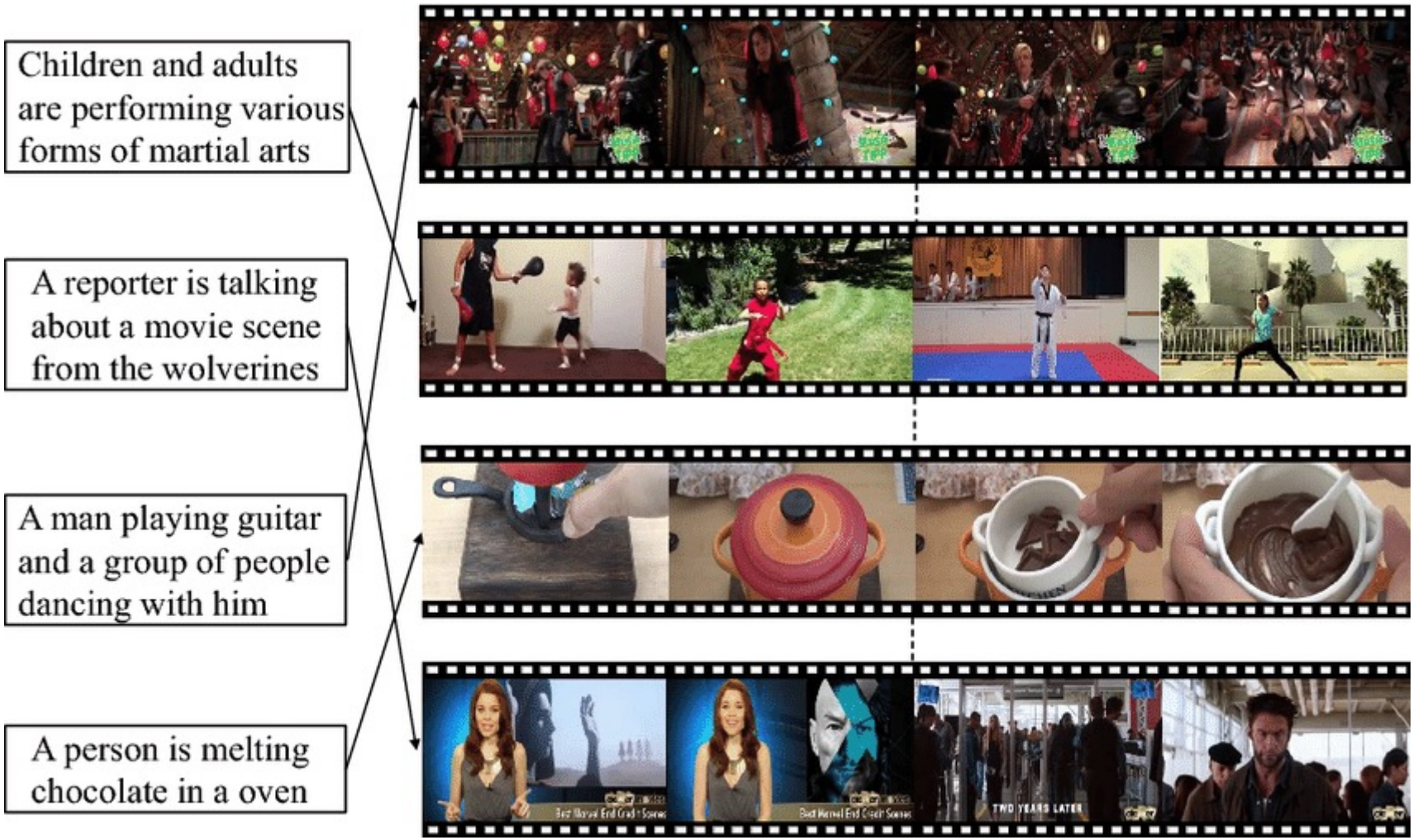
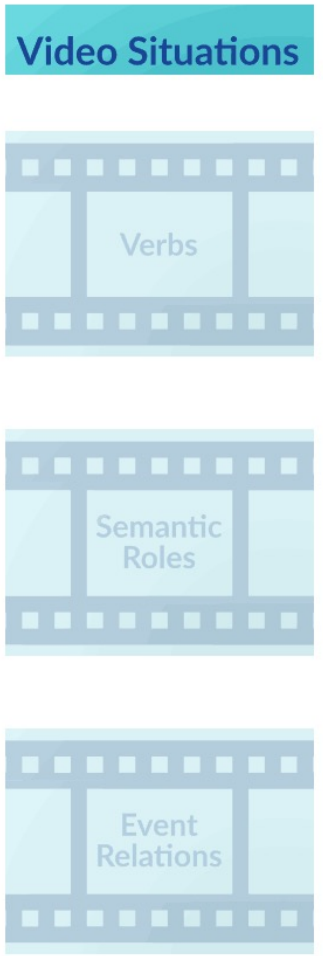


Image Credit: Mithun, et al. Joint embeddings with multimodal cues for video-text retrieval. IJMIR 2019

Video Situation Recognition

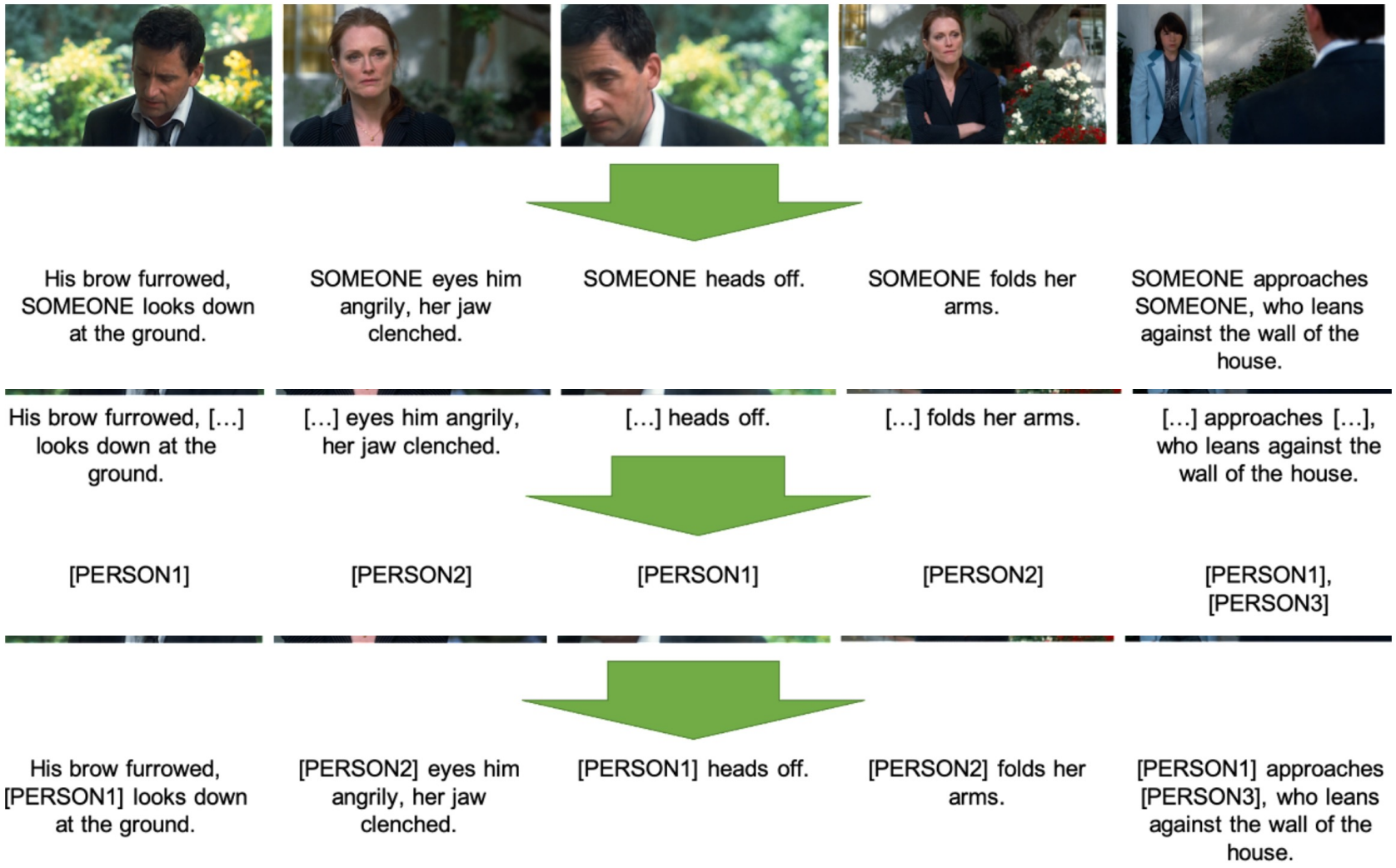


Sadhu, et al. Visual Semantic Role Labeling for Video Understanding. CVPR 2021

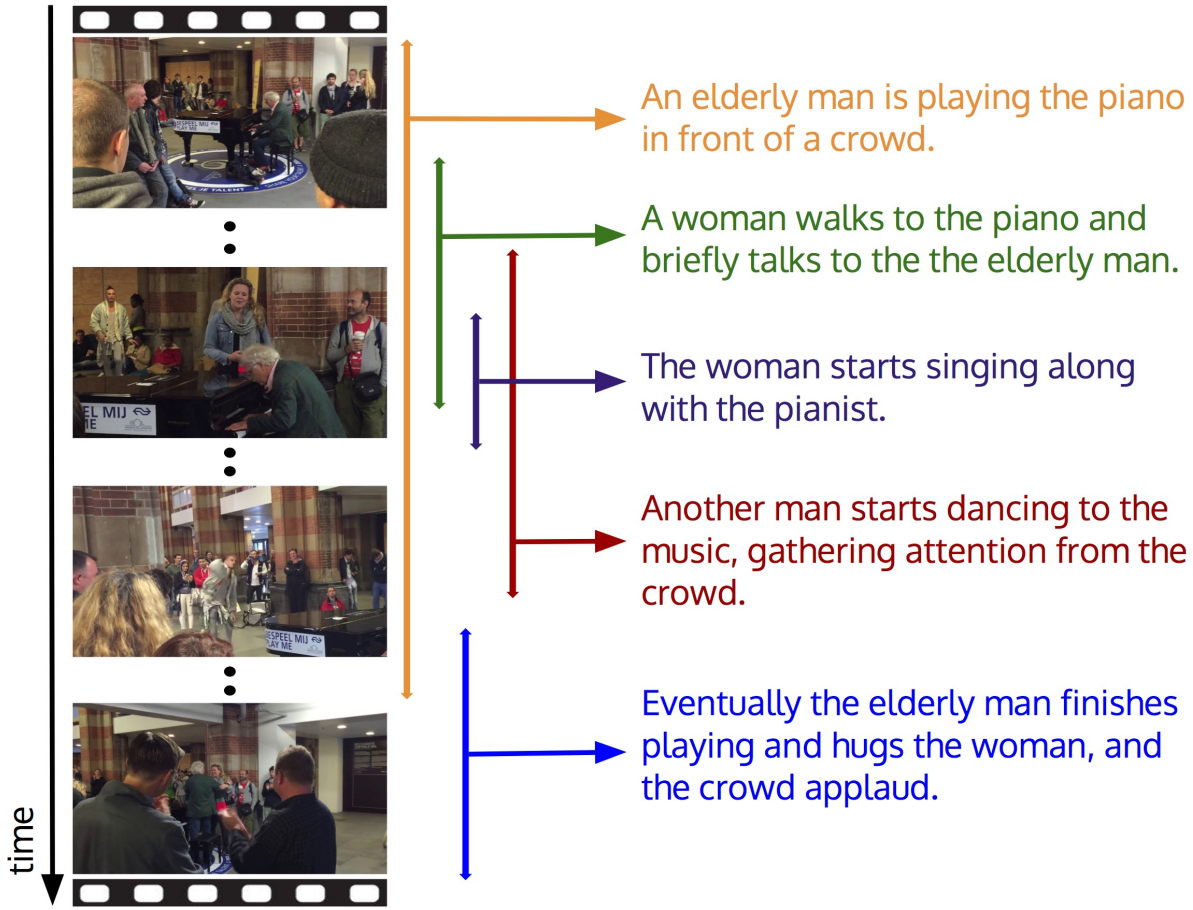
Long-form videos

what is long?!

Identity-aware Captioning (LSMDC)

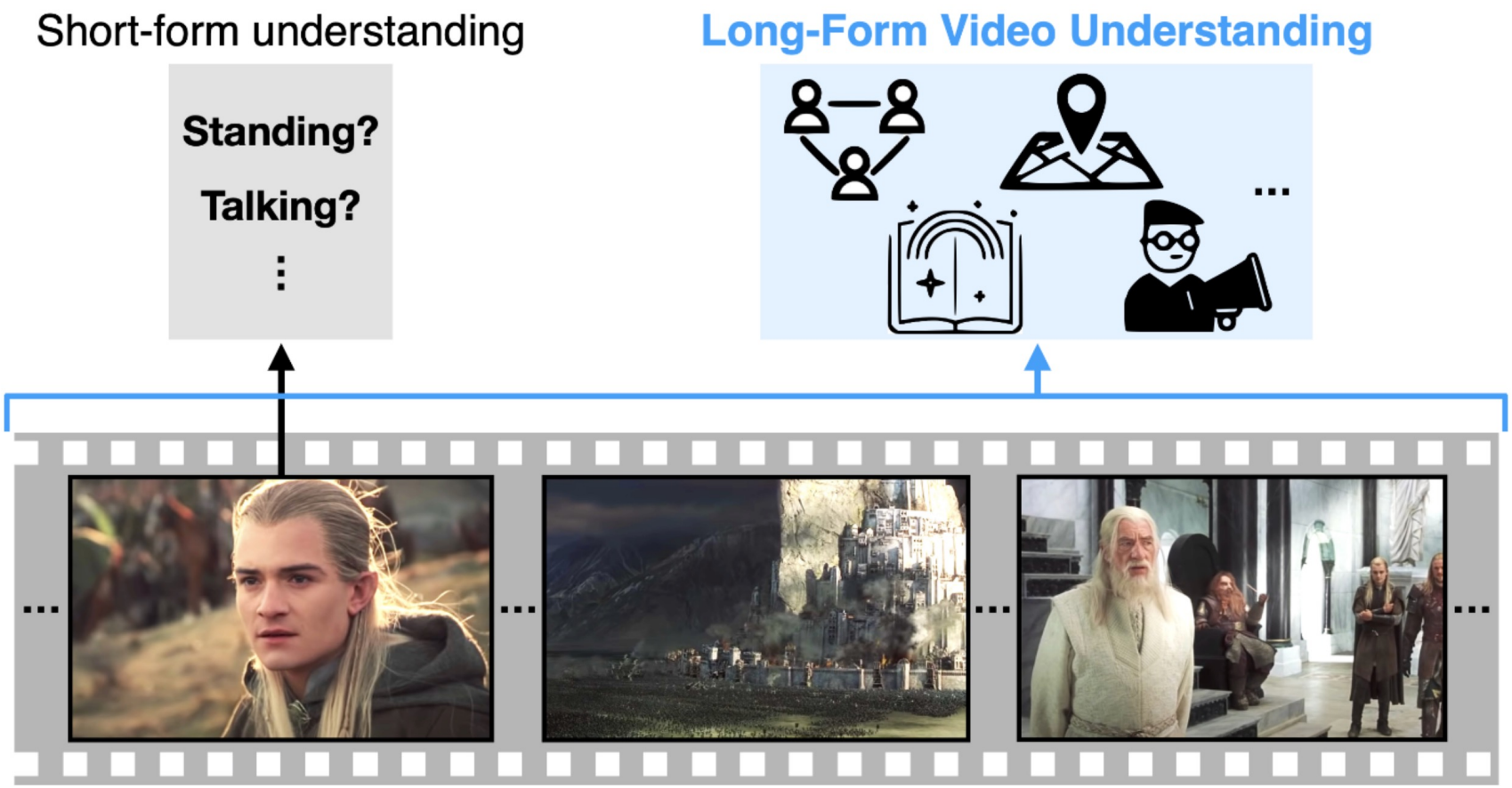


Dense event captioning



Krishna, et al. Dense-Captioning Events in Videos. ICCV 2017

Long-form Video Understanding



Wu and Krähenbühl. Towards Long-form Video Understanding. CVPR 2021