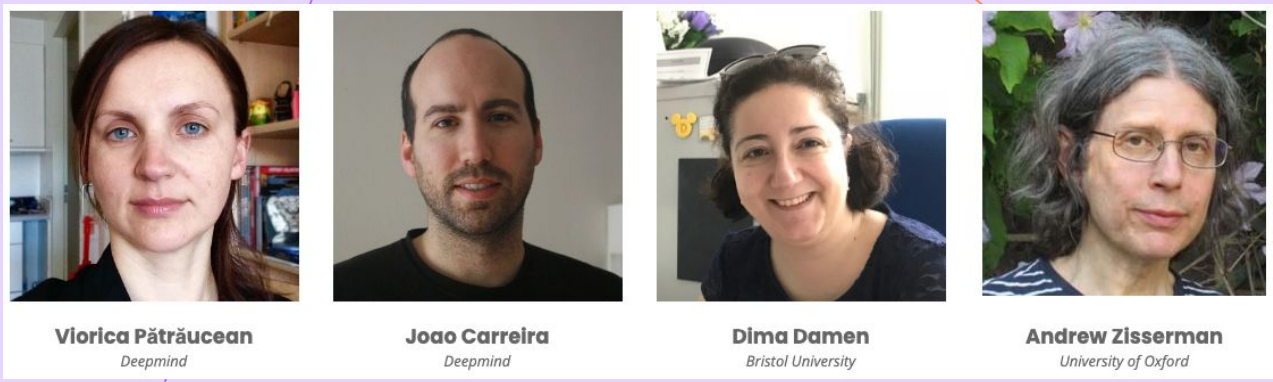
DeepMind

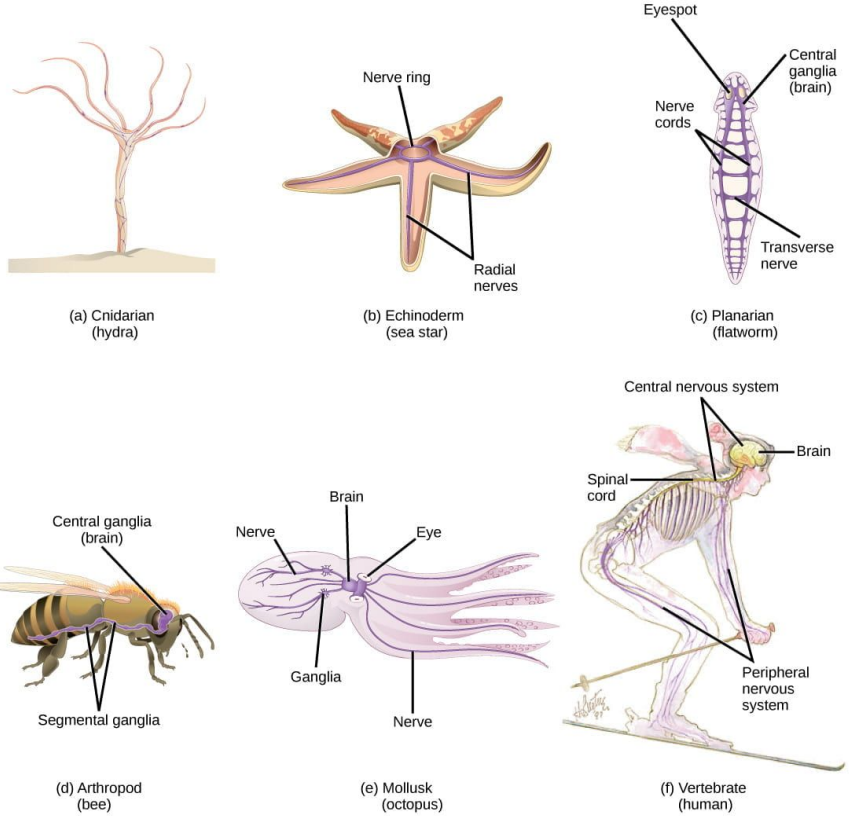
Computer Perception Workshop

Design and evaluation of general perception models



General Perception

General Perception



Credits: *https://organismalbio.biosci.gatech.edu/chemical-and-electrical-signals/nervous-systems/*

Ambitious early days

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group — July 7, 1966

Vision Memo. No. 100.

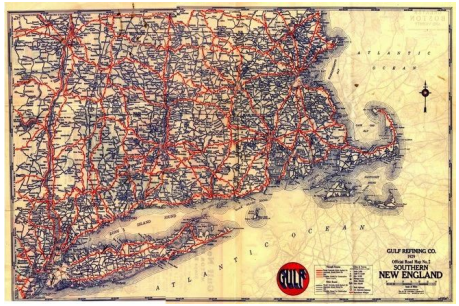
THE SUMMER VISION PROJECT

Seymour Papert

Today - tasks

hand gesture recognition, facial landmark detection, attribute detection, visual question answering, optical flow estimation, semantic image segmentation, face attributes, image retrieval, pose estimation, image generation, image-to-image translation, image captioning, instance segmentation, 3D object detection, visual odometry, image retargeting, video object segmentation, 3D shape estimation, 2D object detection, action localization, 2D pose estimation, face alignment, video object segmentation, object detection in videos, activity detection, human body pose estimation, multi-person pose estimation, text localization, video frame generation, facial expression recognition, image super-resolution, object recognition in videos, person re-identification, panoptic segmentation, 3D reconstruction, structure-from-motion, multiview stereo, depth-from-defocus, normal prediction, single-image depth estimation, speaker detection, object tracking, emotion recognition, iris recognition, shadow detection, cut-shot detection, image super-resolution, video interpolation, arrow of time, object counting, event counting, ...

Overspecialization ?



or

Growing interest in consolidating back on the model side

Audio+Video

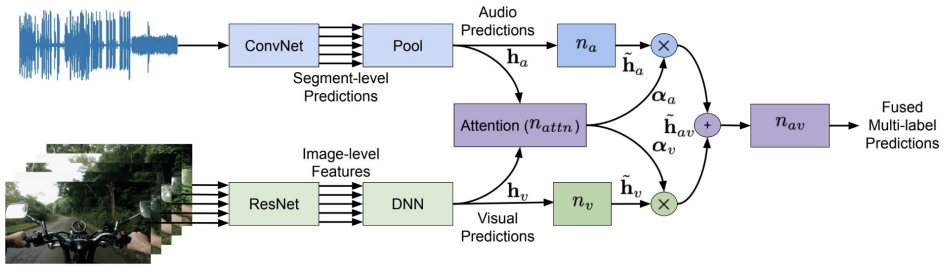
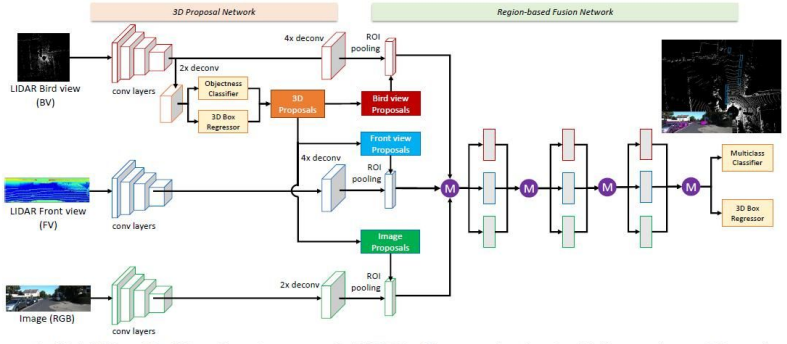
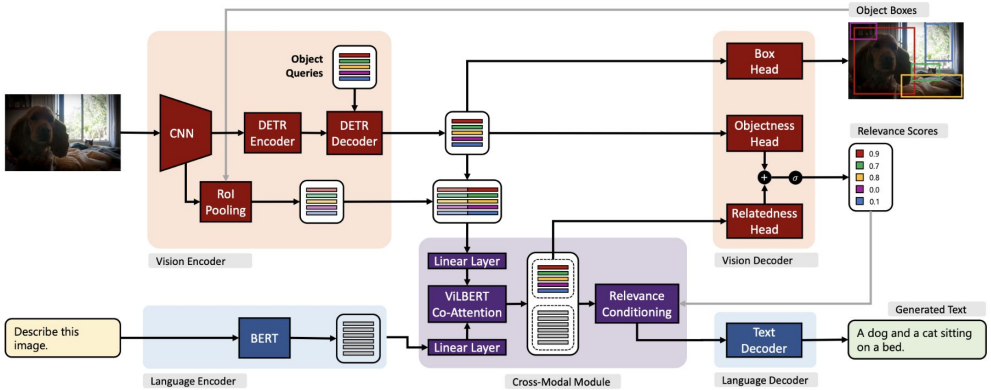


Image + Lidar



Image+Language

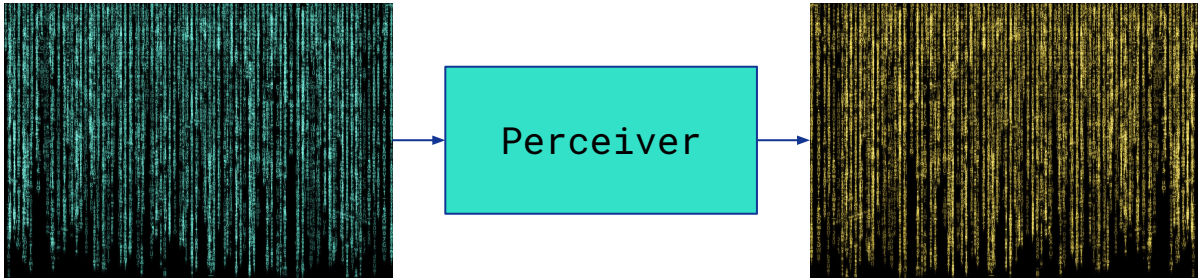


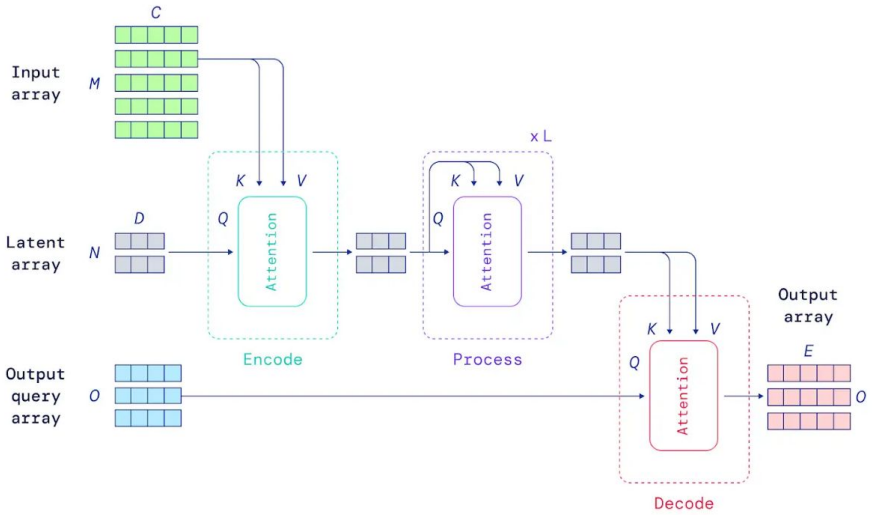
Fayek & Kumar IJCAI 2020 - Large Scale Audiovisual Learning of Sounds with Weakly Labeled Data (audio+video)

Chen et al CVPR 2017 - Multi-View 3D Object Detection Network for Autonomous Driving (image+Lidar)

Gupta et al arXiv 2021 - Towards General Purpose Vision Systems (image+text)

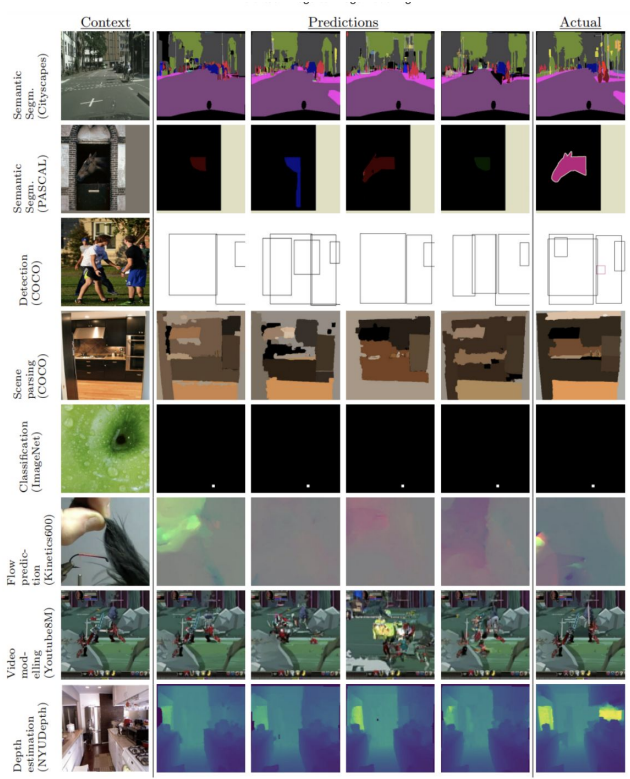
Growing interest in consolidating back on the model side



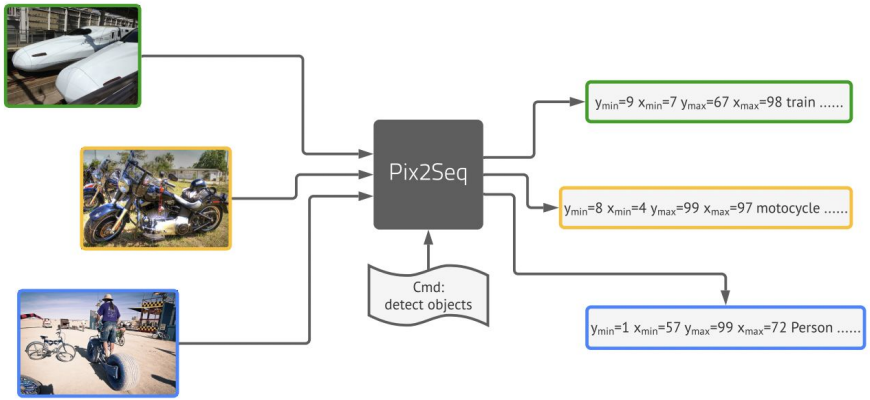


Growing interest in consolidating back on the model side

Transframer



Pix2Seq

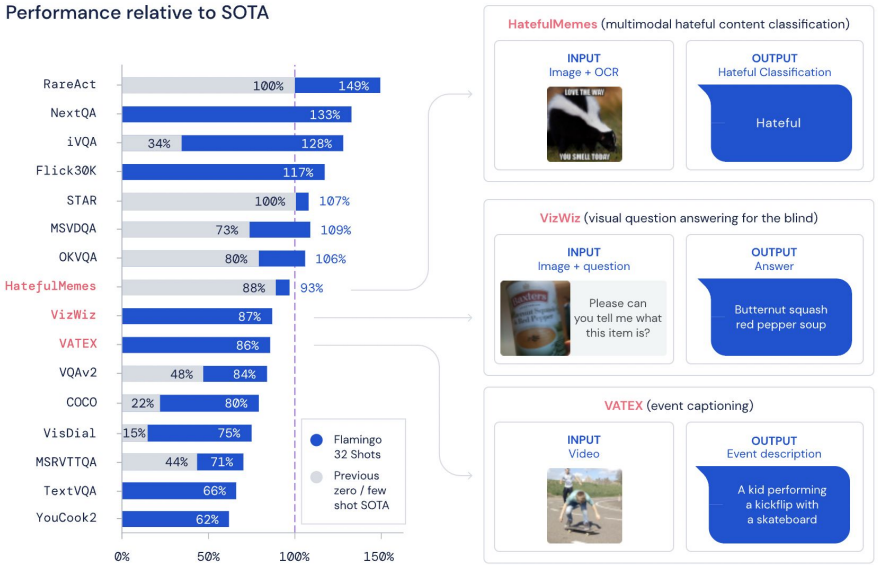


Chen et al 2021 - Pix2seq: A language modeling framework for object detection

Nash et al arXiv 2022 - Transframer: Arbitrary Frame Prediction with Generative Models

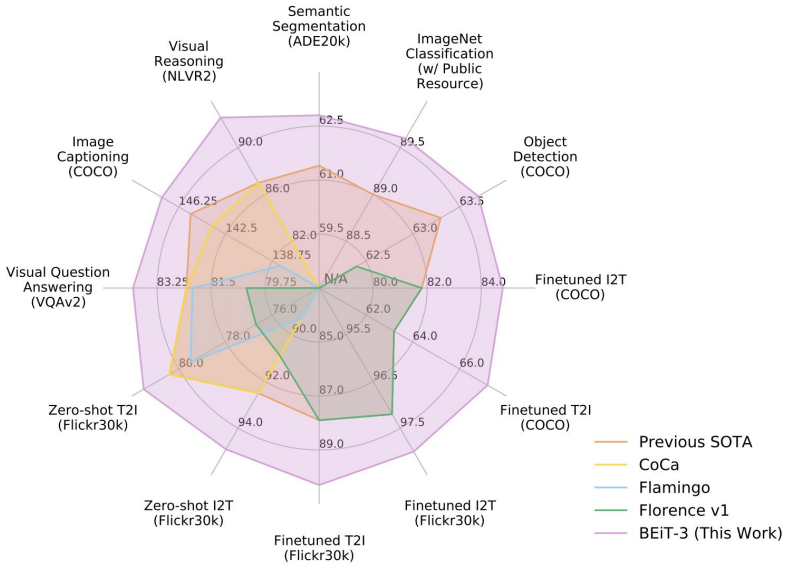
Growing interest in consolidating back on the model side

Flamingo



Alayrac et al 2022 - Flamingo: a visual language model for few shot learning

BeIT-3

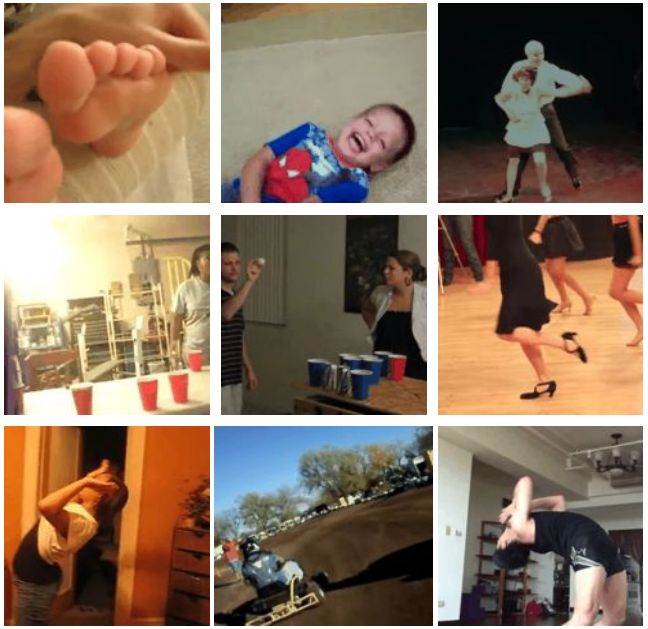


Wang et al 2022 - Image as a Foreign Language: BeiT pretraining for all Vision and Vision-Language tasks

Evaluating general perception

ImageNet

Kinetics

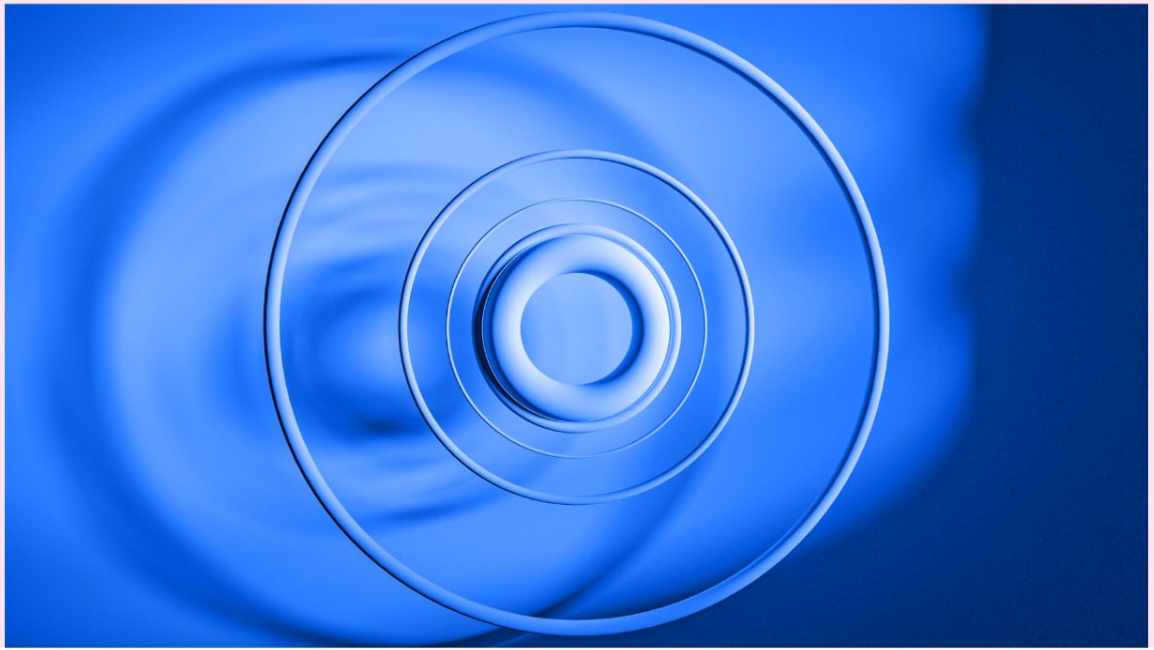


Evaluating general perception: Perception Test (new!)

Research

Measuring perception in AI models

October 12, 2022



DeepMind

Perception Test Team

23 October 2022

DeepMind

Invited Speakers

Aude Oliva
MIT

Daniel Yamins
Stanford University

Jitendra Malik
UC Berkeley

Matt Botvinick
DeepMind

Michael Auli
Facebook AI Research

Olga Russakovsky
Princeton University

DeepMind

Agenda

12:00 - 12:15	**Opening notes**
12:15 - 13:10	**Perception Test (part 1)**
13:10 - 14:00	**Lunch break**
14:00 - 14:35	**Keynote Olga Russakovsky**
14:35 - 15:30	**Perception Test (part 2)**
15:30 - 16:00	**Coffee break**
16:00 - 16:35	**Keynote Aude Oliva**
16:35 - 17:10	**Keynote Matt Botvinick**
17:10 - 17:45	**Keynote Michael Auli**
17:45 - 18:00	**Coffee break**
18:00 - 18:35	**Keynote Daniel Yamins**
18:35 - 19:10	**Keynote Jitendra Malik**
19:10 - 19:55	**Panel discussion**
19:55 - 20:00	**Closing notes**

DeepMind

Perception Test Team

Presenting next

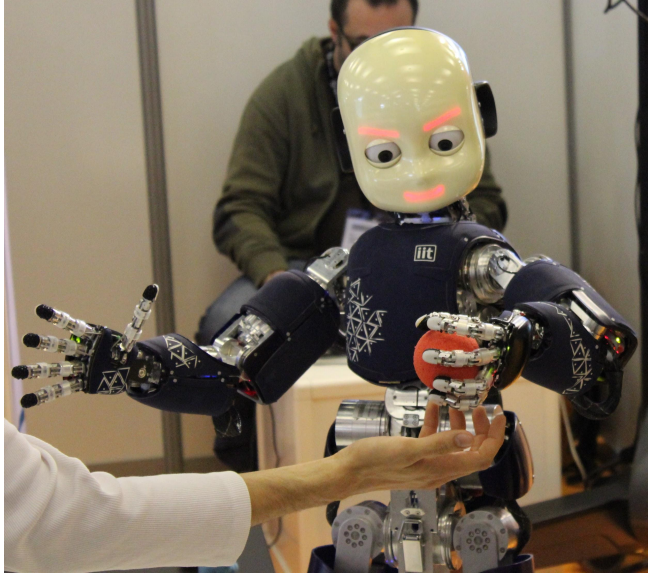
23 October 2022

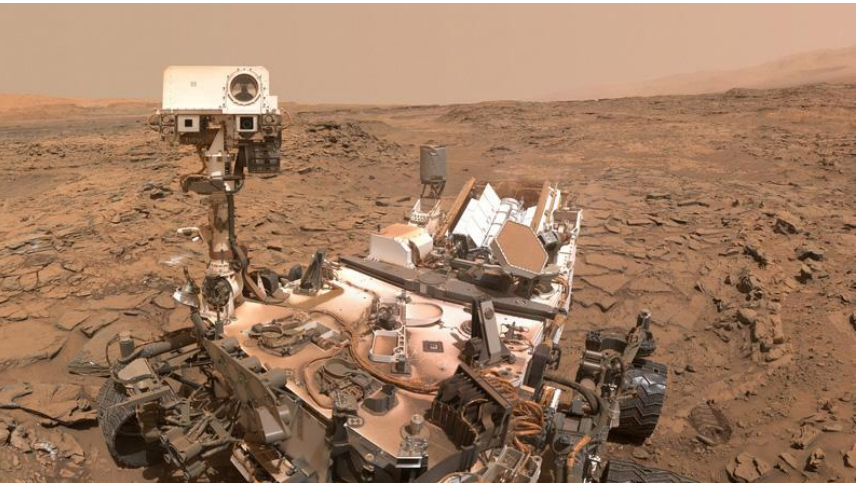
DeepMind

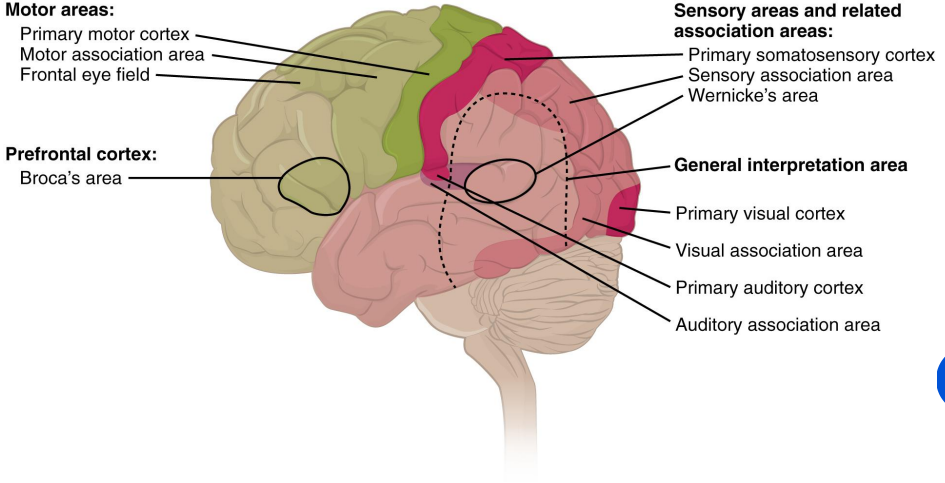
Evaluating general perception models

Viorica Pătrăucean, DeepMind


Holy grail
a model that achieves
human level
scene understanding






Motor areas:
Primary motor cortex
Motor association area
Frontal eye field
Prefrontal cortex:
Broca's area
Sensory areas and related association areas:
Primary somatosensory cortex
Sensory association area
Wernicke's area
General interpretation area
Primary visual cortex
Visual association area
Primary auditory cortex
Auditory association area

What do you see?

'A Walk On The Bike' By Alexandr Vlasyuk

What do you see?

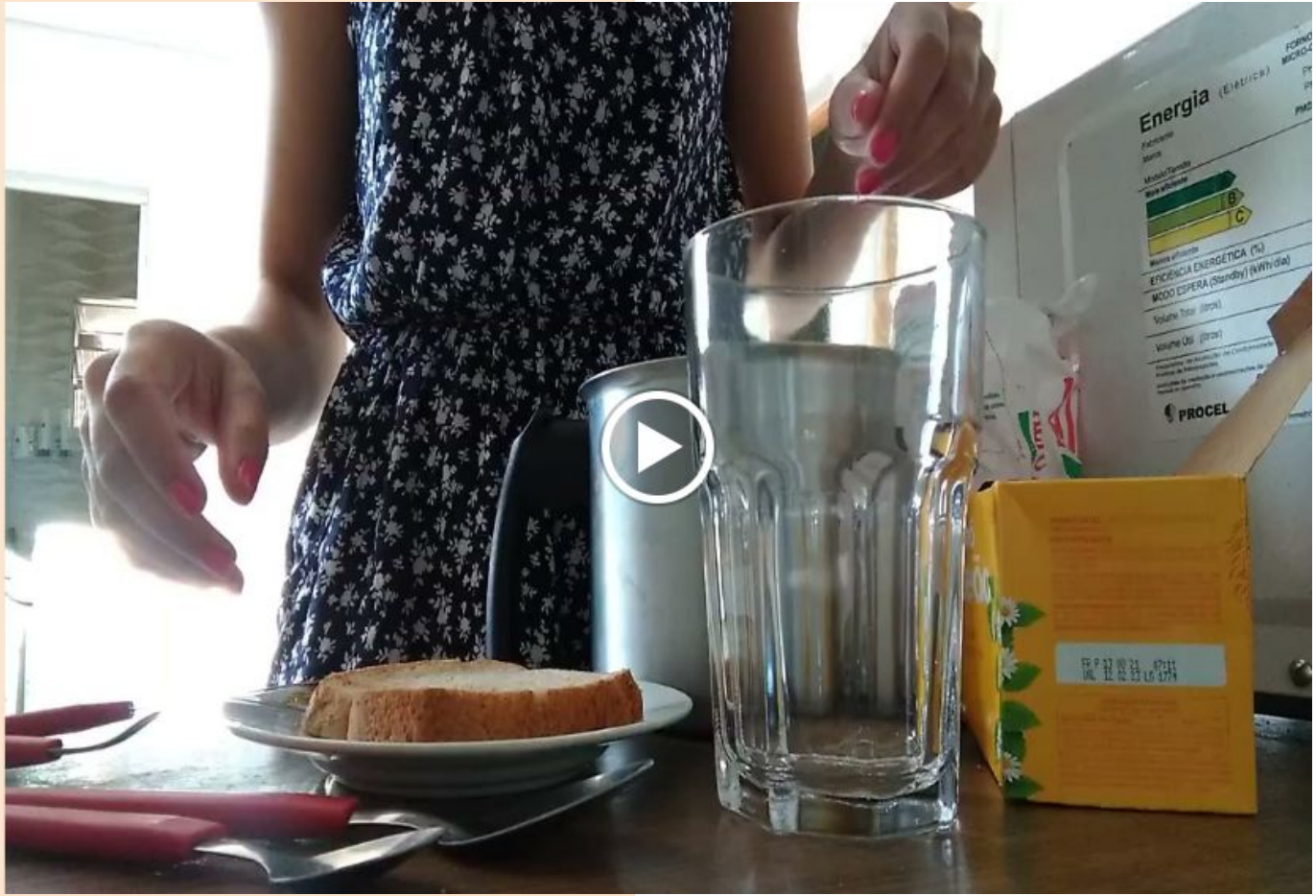
What do models see?

Resnet-50: bicycle, garden

CLIPCLAP: aerial view of a man walking on a path in the park

'A Walk On The Bike' By Alexandr Vlasyuk

What do you see?



Flamingo: This is a person making a sandwich.

To identify the strengths and weaknesses of our models, we need comprehensive & robust quantitative evaluation.

A bit of history in intelligence and perception evaluation

Measuring **human** intelligence/perception [1]

- (**1905**) Binet's intelligence test: 30 short tasks (naming parts of the body, comparing lengths and widths, counting coins, remembering digits, etc).
- (**1912**) William Stern coined the IQ term
- (**1916**) Stanford-Binet test: adaptation to English language and children age norms
- (**1936**) Raven matrices
- (**1955**) Wechsler Adult Intelligence Scale: combination of verbal and non-verbal tasks
- (**1961**) Frostig visual-perception test
- (**1972-2003**) Motor-free visual perception test
- (**1996-2017**) Test of visual perception skills

Measuring **machine** intelligence/perception [2]

- (**1950**) Turing Test / Imitation Game: A machine that exhibits intelligent behaviour indistinguishable from that of a human.
- **(mid-1980s)**: first benchmarks in AI/ML imposed by funding agencies to measure value of research grants
- (**1995, 1998**): MNIST dataset
- (**2006, 2009**): TinyImages* -> CIFAR dataset
- (**2009**) Imagenet
- **Nowadays**: thousands of datasets and benchmarks

[1] Universal Intelligence: A Definition of Machine Intelligence, Legg and Hutter, 2007
[2] A survey of 25 years of evaluation, Church and Hestness, 2019
*withdrawn because of offensive content

Perception-related datasets on papers-with-code

Number of datasets by modality

Image	2068
Video	658
Audio	433
3D	212
Speech	143
RGB-D	122
Point cloud	66
LIDAR	42
RBG video	31
Tracking	29
Stereo	27
Actions	21
TOTAL	**3868**

Number of datasets by tasks

Semantic segmentation	228
Object detection	215
Speech recognition	189
Image classification	186
Pose estimation	103
Visual question-answering	89
Action recognition	87
Face recognition	79
Instance segmentation	55
Image captioning	51
Object tracking	41
Scene understanding	39
TOTAL	**1362**

Shortcomings of existing benchmarks

- Often focus on a single task (e.g. image classification, object recognition, action recognition)
- Provide a single opaque score, no indication of model's strengths/weaknesses
- High fragmentation across computational tasks and modalities
- Poor coverage of some important areas: memory and physics skills, counterfactual reasoning, etc.

Goal: design a diagnostic multimodal perception benchmark using real-world videos

Desirable properties of an ideal intelligence test

Universal Intelligence: A Definition of Machine Intelligence, Legg and Hutter, 2007

- Repeatable – no overfitting
- Free of bias
- Validity – the test actually measures what it claims to measure
- Predictive power – performance on the test correlates with performance in real life
- Efficient – cost of running the test should be low

These apply to perception tests as well!

Goal: Comprehensive, robust, and efficient evaluation

Comprehensive:

- Diverse skills and tasks probed on real-world videos
- Probe for spatial and temporal understanding, both low-level and high-level, across multiple modalities
- Different types of reasoning (descriptive, explanatory, predictive, counterfactual)

Robust:

- Zero/few-shot or fine-tuning regimes (no in-dataset training from scratch)

Efficient:

- High density of diverse annotations on the same videos: extract features once, assess with multiple queries
- Enable cross-task analysis to diagnose the strengths and weaknesses of a model

DeepMind

Perception Test

multimodal perception benchmark

Outline

Part 1:

- Overview
- Try it yourself!
- Perception Test vs existing benchmarks

Part 2:

- Annotations and baselines
- Next steps

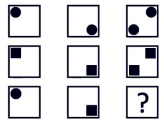
Comprehensive: wide coverage of skills in multiple areas

Memory

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

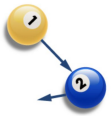
Abstraction

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



Physics (and Geometry)

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



Semantics

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

Videos

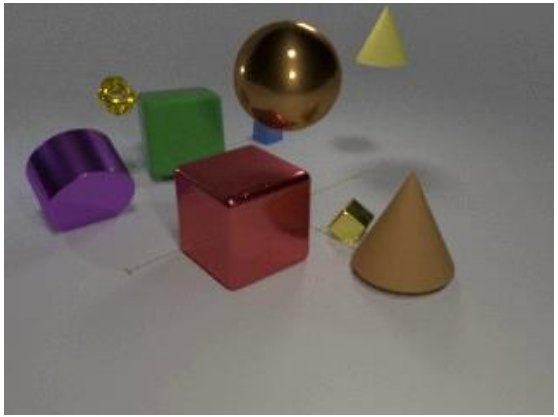
- Perceptually interesting videos
- Balanced and diverse dataset
- Enable language and non-language tasks
- Assessment of high-level understanding probed through low-level tasks (e.g. memory probed through object tracking) and vice versa (e.g. physics probed through VQA)

⟶ Design scripts and film them with crowd-sourced participants

Script design

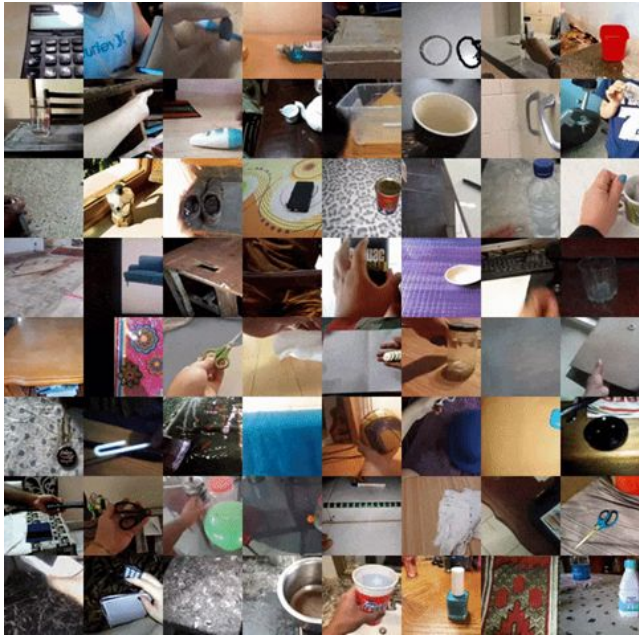
Inspiration:

- Perception tests for humans
- Synthetic datasets (CATER)
- Real-world datasets (SSv2)



CATER

Props from human perception tests



SSv2

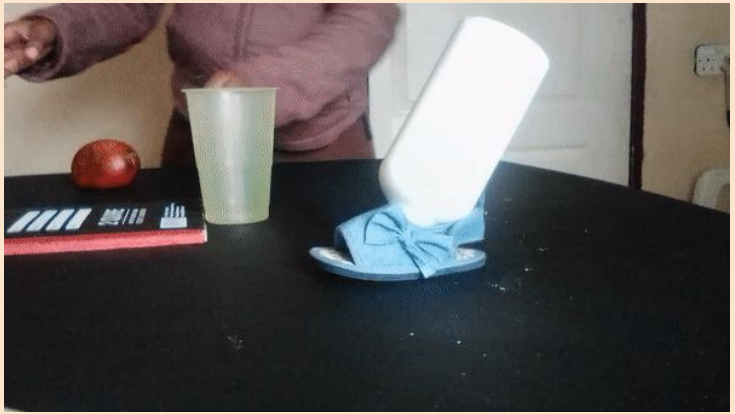
Script design

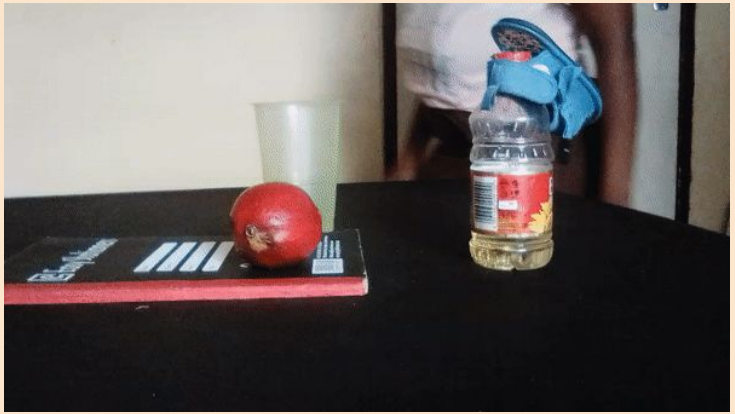
- Simple situations or games that can be performed by one person (w/ props shipped)
- Multiple variations per script for diversity; **37 scripts, 153 variations**
- No face or speech, focus on hands and their interaction with objects
- Distractor objects and distractor actions
- Adversarial object configurations and adversarial actions (violation of expectation)

Scene description	Actions description	Camera viewpoint
In field of view: N distractor objects; 3 identical objects (cups, glasses), hide one small object under one of the identical objects	lift each of the identical objects one at a time to show where the hidden object is, do some actions on the distractor objects for a few seconds; lift the identical objects again one at a time to show where the hidden object is	Static and moving
In field of view: N distractor objects; 4 identical objects (cups, glasses), hide one small object under one of the identical objects	lift each of the identical objects one at a time to show where the hidden object is, move the identical objects around for a few seconds; lift the identical objects again to show where the hidden object is	Static and moving

Loose script definition: room for variability and creativity!

Examples of collected videos: scripts with 2 variations

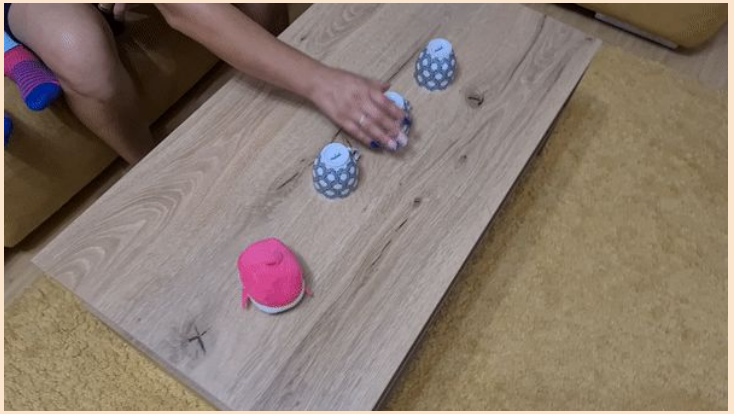


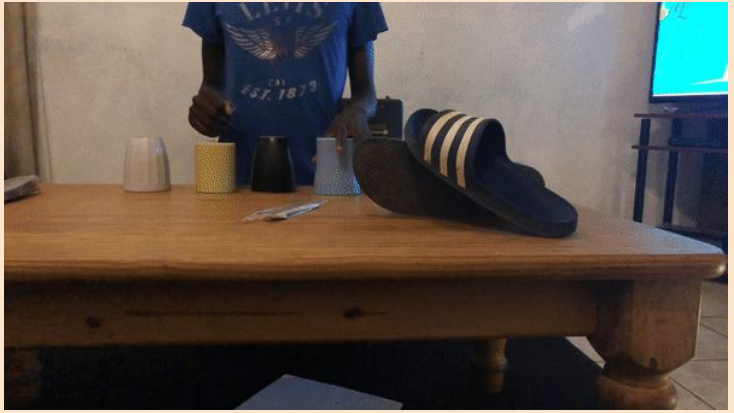


Assess stable configurations

Assess task completion

Examples of collected videos: script with 4 variations



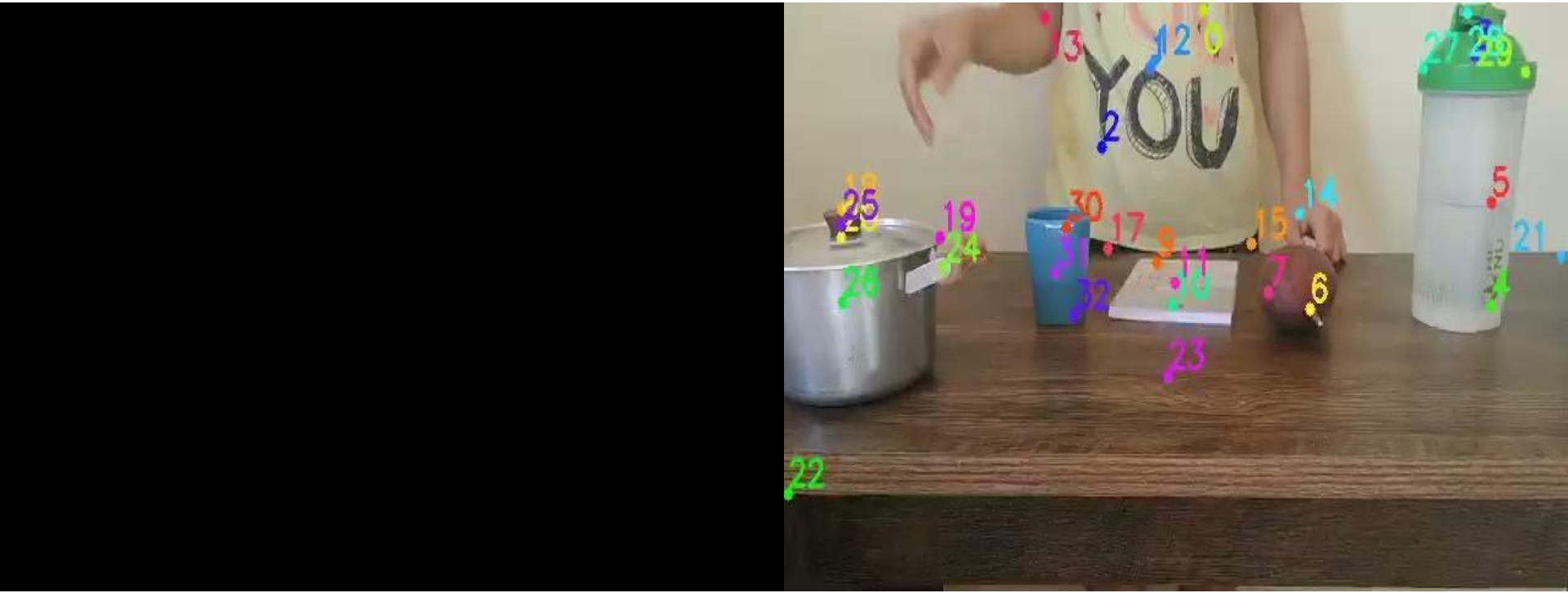


Cups-games

Comprehensive: low-level and high-level annotations

Annotation type	# classes	# annotated instances	# videos	Rate
Objects tracks	5125	191716	11672	1fps
Point tracks	NA	8574	145	30fps
Action segments	63	73859	11416	30fps
Sound segments	16	137756	11484	30fps
mc-vQA	132	38658	11672	NA
g-vQA	35	5890	3085	1fps

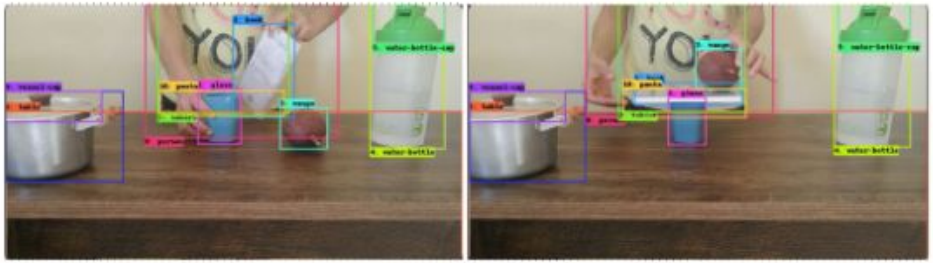
Comprehensive: low-level and high-level annotations



Box tracks

Point tracks

EXAMPLE 1



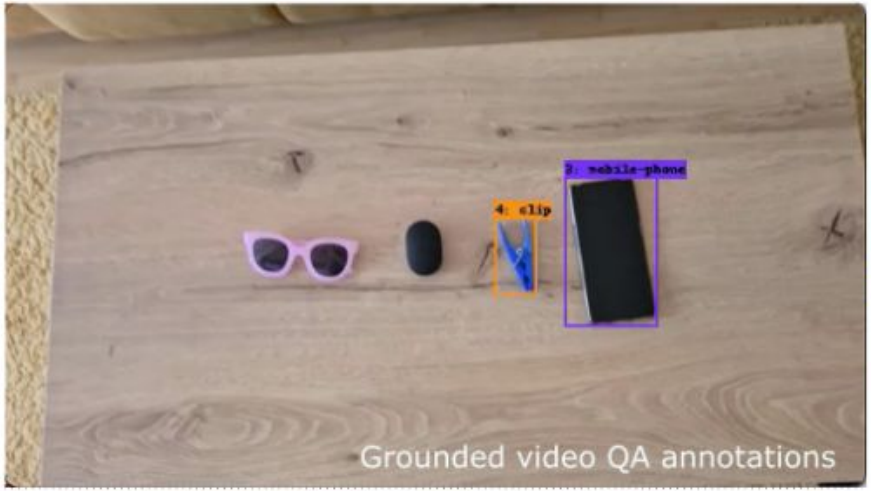
Object tracks

Point tracks

Multiple-choice video QA
Area: ***Physics***, Reasoning: ***Predictive***
Question: *Is the configuration of objects likely to be stable after placing the last object?*

Options:
a) *The configuration is likely to be stable.*
b) *The configuration is likely to be unstable.*
c) *One cannot judge the stability of this configuration.*

EXAMPLE 2



Multiple-choice video QA
Area: ***Memory***, Reasoning: ***Explanatory***
Question: *What changed on the table while the camera was looking away?*
Options:
a) *The mobile and clip swapped positions.*
b) *The bottle and watch were removed and a clip and mobile were added.*
c) *The mobile was added and a clip was removed.*

Grounded video QA
Area: ***Memory***, Reasoning: ***Descriptive***
Question: *Track the objects that were added to the table while the camera was looking away.*

EXAMPLE 3

Multiple-choice video QA
Area: ***Memory***, Reasoning: ***Counterfactual***
Question: *If the person had put the objects in the backpack in reverse order, which object or objects would have been put in second?*

Options: a) shirt b) pen c) laptop

EXAMPLE 4

Multiple-choice video QA
Area: ***Semantics***, Reasoning: ***Explanatory***
Question: *What action or actions did the person fail to complete and why?*

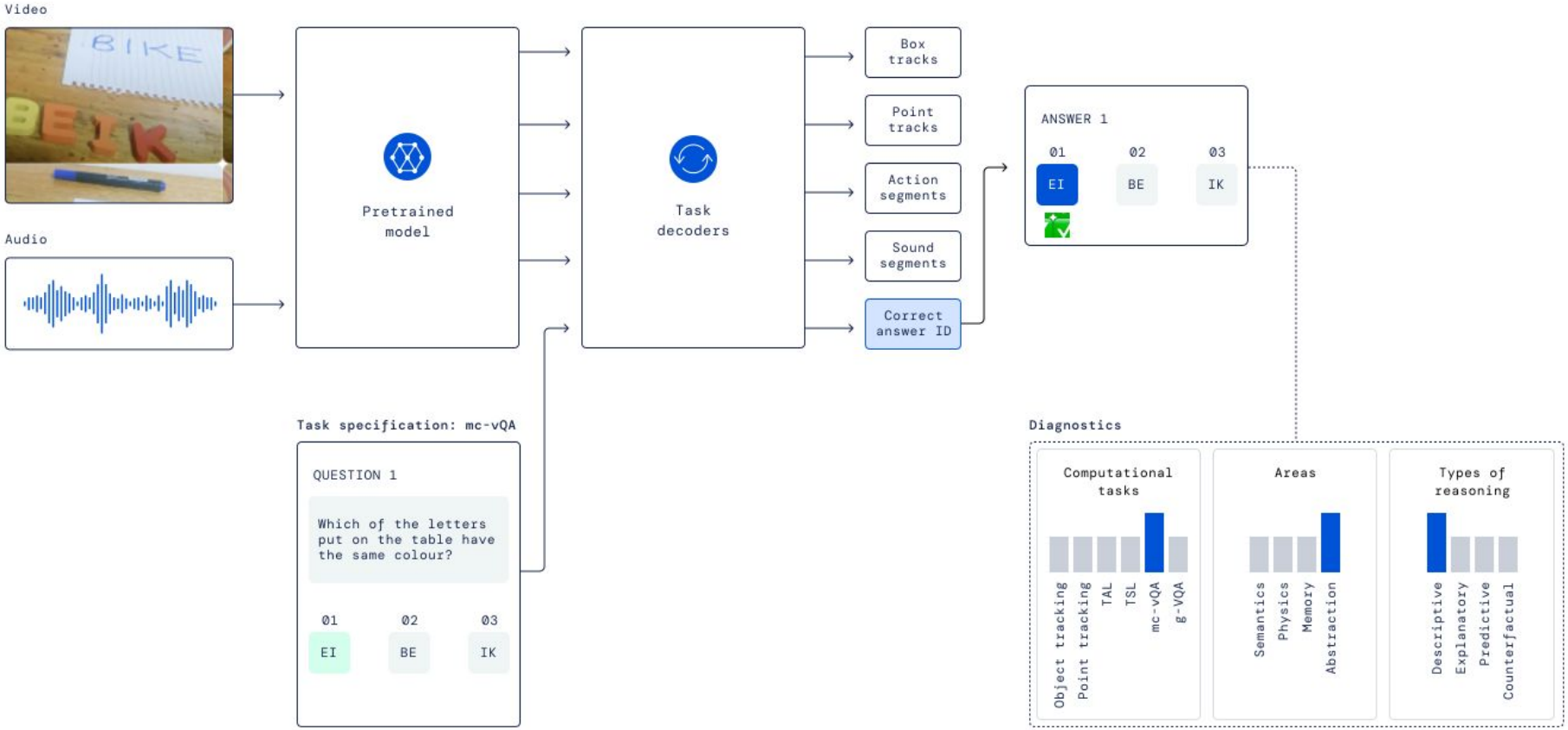
Options:
a) *The person put the teabag next to the cup instead of inside the cup.*
b) *The person tried to pour water, but failed because they didn't tilt the container enough.*
c) *The person tried to pour water, but failed because the water container seems empty.*

EXAMPLE 5

Multiple-choice video QA
Area: ***Abstraction***, Reasoning: ***Descriptive***
Question: *Which letters from the ones the person puts on the table have the same colour?*

Options:
a) *EI* b) *BE* c) *IK*

Robustness: Zero/few-shot or fine-tuning regime



Robustness: Zero/few-shot or fine-tuning regime

Train split	Validation split	Test split
2202 videos	3544 videos	5926 videos

11.6k videos (with audio), 23s average length

Efficiency: high density of annotations

Dataset	Data source	Skill area	#videos	Avg #annot. / video	Avg length (s)
Charades	scripted, real	S	10,000	14	30
SSv2	scripted, real	AS	108,499	1	4
Ego4D	real	MS	102,896‡	9*	492†
CLEVRER	scripted, synth	P	60,000	N/A	5
Perception Test	scripted, real	MAPS	11,672	907	23

- Widest coverage of skills areas (M=Memory, A=Abstraction, P=Physics, S=Semantics)
- Highest density of annotations
  - Efficient evaluation
  - Detailed analysis

Diversity of participants in visual data

Self-reported characteristics

Gender	%
Female, Other	53.60%
Male	46.40%

Ethnicity	%
White or Caucasian	28.97%
South and East Asian	25.49%
Black or African American	21.68%
Latino or Hispanic	9.25%
Middle Eastern	3.37%
Native Hawaiian or Pacific Islander	0.62%
Caribbean	0.22%
America Indian or Alaska Native	0.02%
Mixed	3.94%
Unknown	6.44%

Country	%
Philippines	31.38%
Brazil	11.27%
Kenya	10.02%
Indonesia	8.75%
Italy	8.03%
Romania	7.57%
South Africa	5.25%
Turkey	4.12%
India	3.72%
Mexico	1.45%
Bulgaria	1.37%
United States	0.70%
Egypt	0.48%
Other	5.87%

Around 100 participants, filmed a total of 11.600 videos.

Diversity of participants in visual data

Try it yourself

Link to Perception Test sample google form

DeepMind

The Perception Test in Context

Dima Damen, University of Bristol

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

Datasets for ...

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

Datasets for ...

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

Datasets for ...

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

Datasets for ...

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

Datasets for …

Training and Pretraining	Fine-Grained Actions	A…	
Large-scale Diversity Weak/sparse …	Fine-grain… … …ades	…lasses only Audioset	A split of the training set All
	Something-Something EPIC-KITCHENS Ego4D	VGG-Sound EPIC-KITCHENS Ego4D	

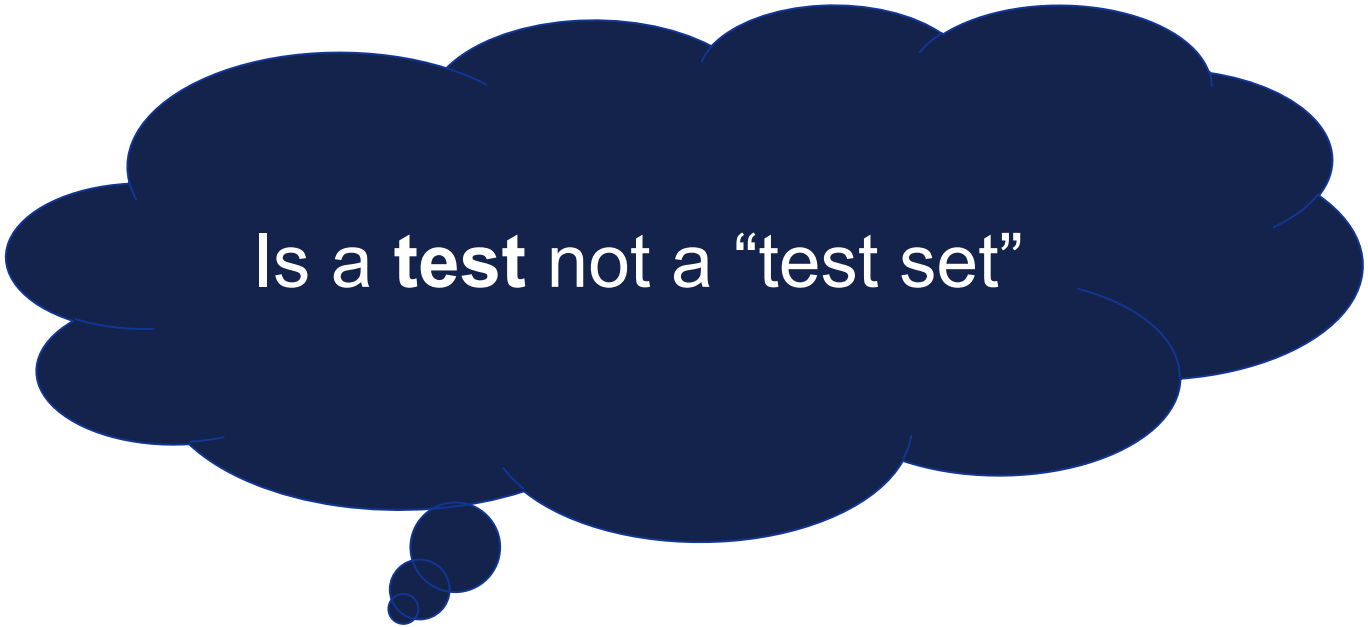
Focus on Training than evaluation….

But what about evaluation??

- Annotated in a consistent manner to the training set
- One or more metrics
- Considering all the test set at a time
- Number of training examples (head/tail)
- New environments (seen/unseen)

- But very limited insight into what to do next!

The perception test



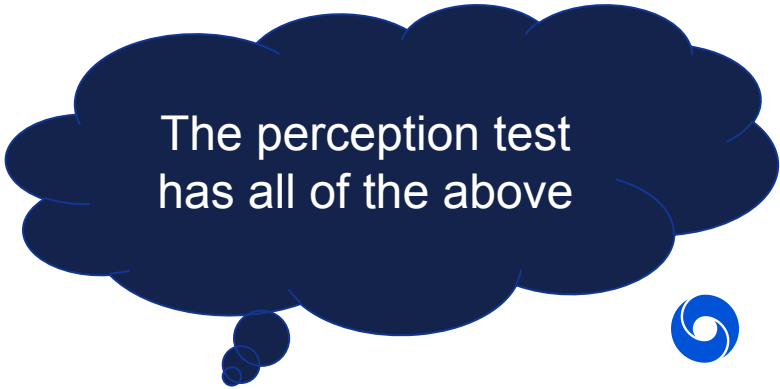
The perception test

Related previous efforts

In Collecting the data:

- Consent forms with informed users (e.g. EPIC-KITCHENS)
- Diverse participants and geolocations (e.g. Charades, Ego4D)
- User-selected objects for Diversity (e.g. Something-Something)
- User-selected camera viewpoint (e.g. Charades)
- Hand-Object Interactions (e.g. Something-Something)
- Subtle or incomplete actions (e.g. Something-Something)
- Irrelevant or random actions (e.g. Charades)

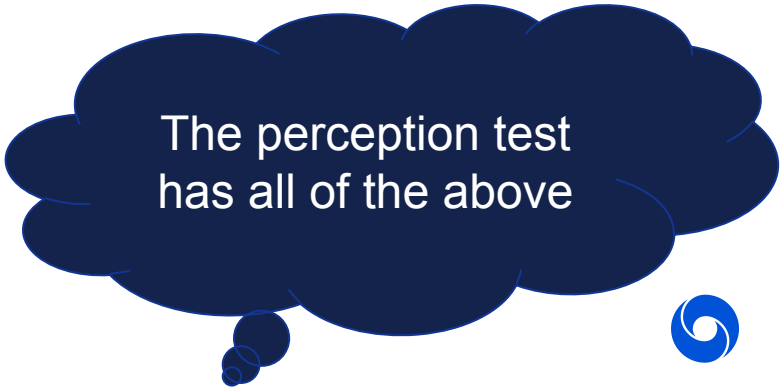
I



Related previous efforts

In Collecting Annotations:

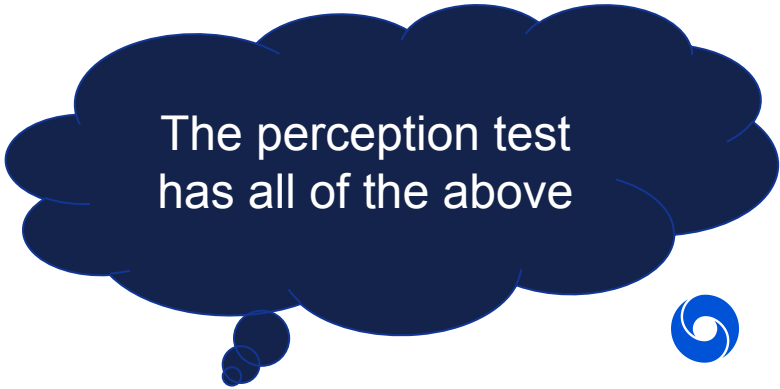
- Multiple levels of annotations (e.g. Kinetics)
- Untrimmed with action-level labels (e.g. ActivityNet)
- Sound-level labels (e.g. VGG-Sound)
- Bounding-Box Trajectories of actors and objects (e.g. Ava, Action Genome)
- Action-relevant objects bounding boxes (e.g. EPIC-KITCHENS)
- Point trajectories (e.g. TAPNet-Kinetics)



Related previous efforts

In Tasks:

- Action detection (e.g. THUMOS)
- Spatio-temporal action localisation (e.g. AVA)
- Moment Retrieval (e.g. Ego4D)
- Episodic memory tasks (e.g. Ego4D)
- VQA tasks (e.g. Ego4D)
- Sound recognition/detection (e.g. VGG-Sound)
- .....



Related previous efforts

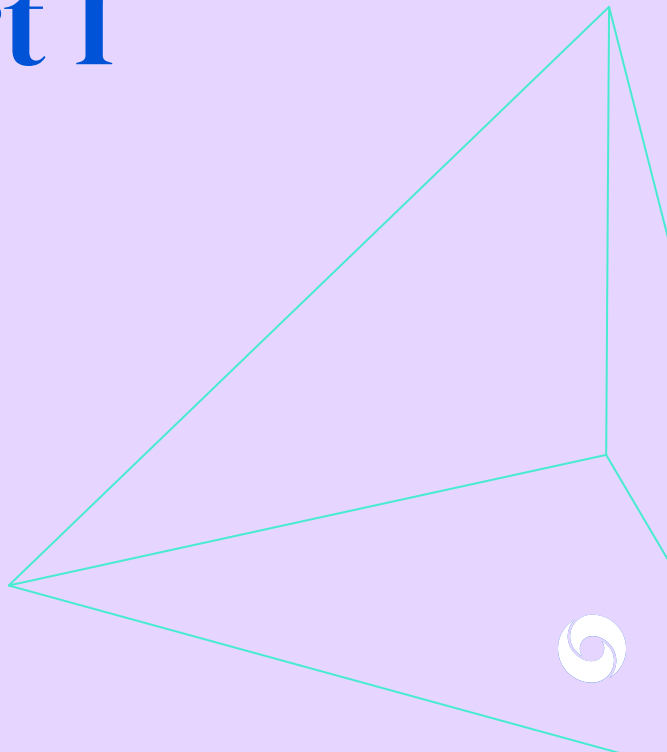
Dataset	Data source	Skill area	#videos	Avg #annot. / video	Avg length (s)
Charades	scripted, real	S	10,000	14	30
SSv2	scripted, real	AS	108,499	1	4
Ego4D	real	MS	102,896‡	9*	492†
CLEVRER	scripted, synth	P	60,000	N/A	5
Perception Test	scripted, real	MAPS	11,672	907	23

How to train models for the “Perception Test”?

- Use any/all data...
- Fine-Grained Models...
- Variable-length videos...
- Audio-Visual Input...
- Strong language/high-level reasoning...

DeepMind

End part 1



DeepMind

Agenda

12:00 - 12:15	**Opening notes**
12:15 - 13:10	**Perception Test (part 1)**
13:10 - 14:00	**Lunch break**
14:00 - 14:35	**Keynote Olga Russakovsky**
14:35 - 15:30	**Perception Test (part 2)**
15:30 - 16:00	**Coffee break**
16:00 - 16:35	**Keynote Aude Oliva**
16:35 - 17:10	**Keynote Matt Botvinick**
17:10 - 17:45	**Keynote Michael Auli**
17:45 - 18:00	**Coffee break**
18:00 - 18:35	**Keynote Daniel Yamins**
18:35 - 19:10	**Keynote Jitendra Malik**
19:10 - 19:55	**Panel discussion**
19:55 - 20:00	**Closing notes**

DeepMind

Agenda

12:00 - 12:15	**Opening notes**
12:15 - 13:10	**Perception Test (part 1)**
13:10 - 14:00	**Lunch break**
14:00 - 14:35	**Keynote Olga Russakovsky**
14:35 - 15:30	**Perception Test (part 2)**
15:30 - 16:00	**Coffee break**
16:00 - 16:35	**Keynote Aude Oliva**
16:35 - 17:10	**Keynote Matt Botvinick**
17:10 - 17:45	**Keynote Michael Auli**
17:45 - 18:00	**Coffee break**
18:00 - 18:35	**Keynote Daniel Yamins**
18:35 - 19:10	**Keynote Jitendra Malik**
19:10 - 19:55	**Panel discussion**
19:55 - 20:00	**Closing notes**

DeepMind

Perception Test Team

23 October 2022

Outline

Part 1:

- Overview
- Try it yourself!
- Perception Test vs existing benchmarks

Part 2:

- Annotations and baselines
- Next steps

Perception Test summary

11.6k purposefully designed videos to show interesting perceptual situations

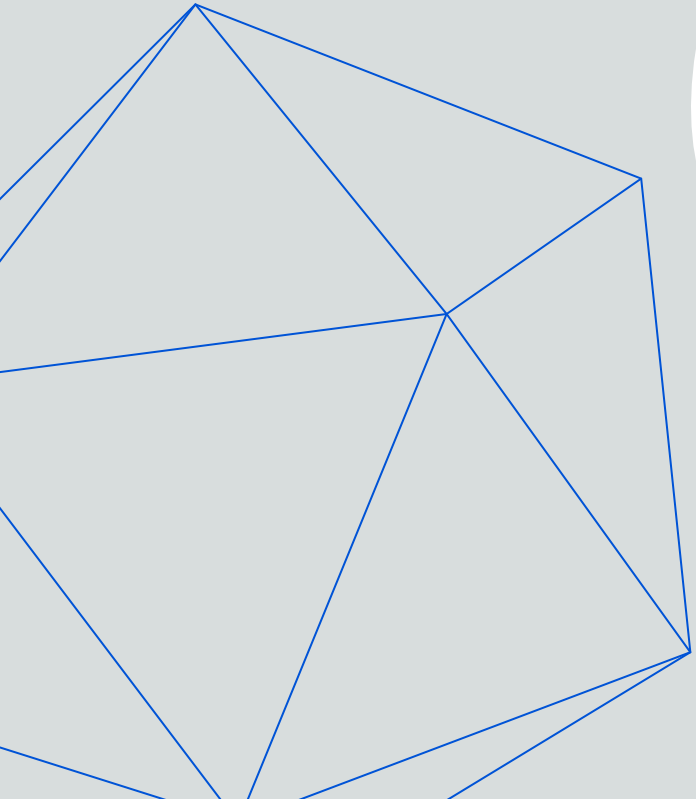
6 types of annotations

- Object tracks: Ankush Gupta
- Point tracks: Yi Yang, Carl Doersch
- Action segments } Adria Recasens Continente
- Sound segments
- Multiple-choice videoQA } Viorica Patraucean
- Grounded videoQA

Cleaning of annotations

DeepMind

Object tracking



Objects: core units of perception?

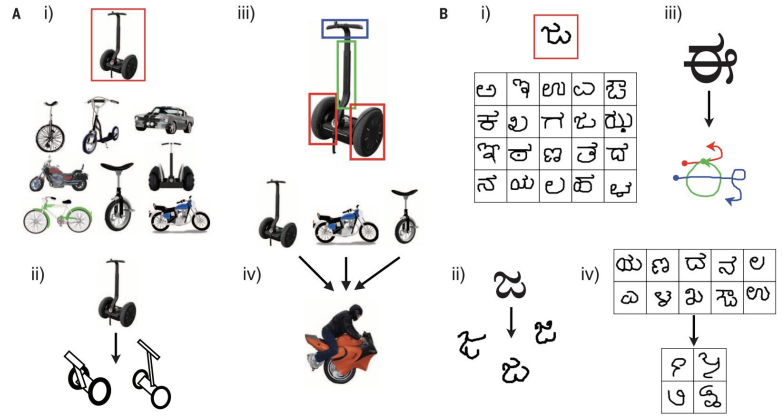
Core Knowledge

Elizabeth S. Spelke
Massachusetts Institute of Technology

Twenty years of research provides evidence that infants build representations of objects as complete, connected, solid bodies that persist over occlusion and maintain their identity through time (e.g., Baillargeon, 1993; Spelke & Van de Walle, 1993). One of the situations that reveal this ability was devised by Karen Wynn (1992). Wynn's studies used a preferential looking–expectancy violation method, based on the assumption that infants would look longer at an unex-

Human-level concept learning through probabilistic program induction

Brenden M. Lake,[1]* Ruslan Salakhutdinov,[2] Joshua B. Tenenbaum[3]

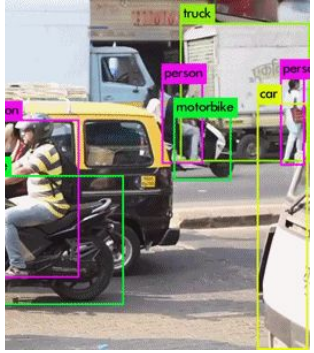


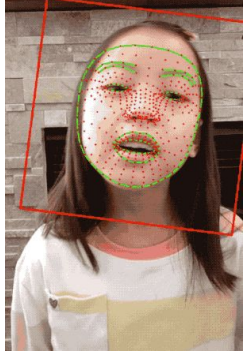
Flexible recombination of objects/concepts and their attributes/affordances is the foundation of perception at least in humans, perhaps also in machines.

Objects in Computer Vision

Key datasets/benchmarks in the community

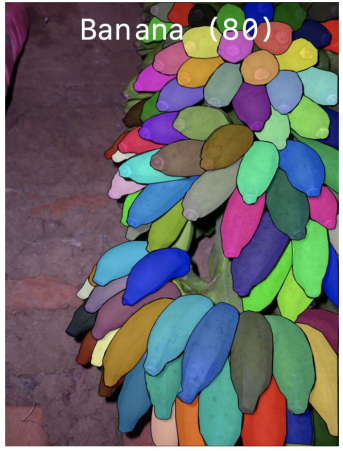
defined in terms of objects and their attributes.





Image/video object classification, detection, segmentation, tracking, pose estimation,

3D reconstruction, counting …



PASCAL2

Pattern Analysis, Statistical Modelling and Computational Learning

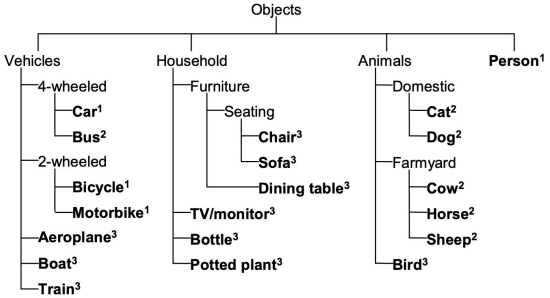


Fig. 2 VOC2007 Classes. Leaf nodes correspond to the 20 classes.

COCO

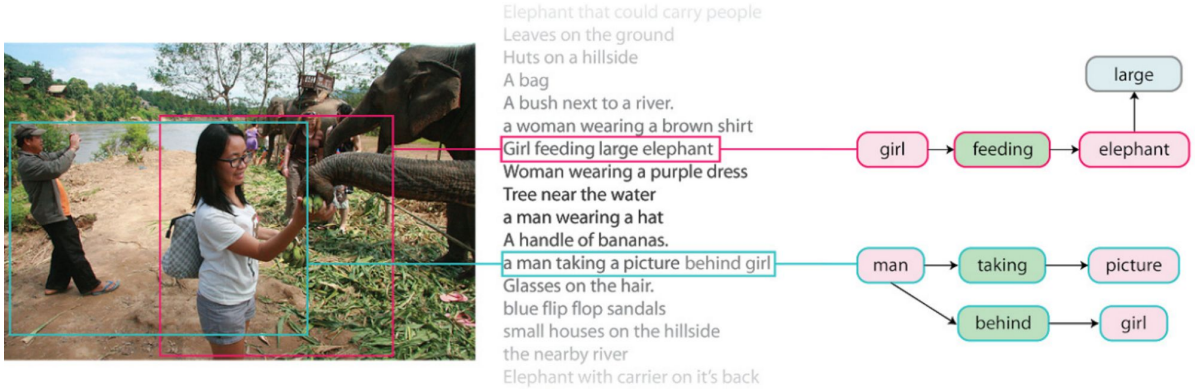
Common Objects in Context

COCO is a large-scale object detection, segmentation, and captioning dataset. COCO has several features:

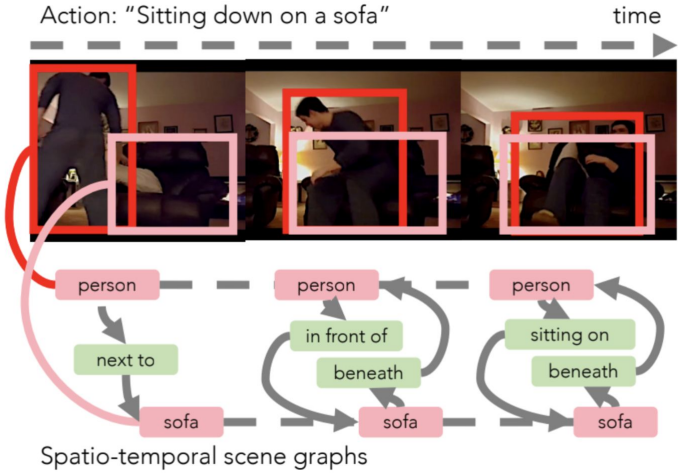
- Object segmentation
- Recognition in context
- Superpixel stuff segmentation
- 330K images (>200K labeled)
- 1.5 million object instances
- 80 object categories
- 91 stuff categories
- 5 captions per image
- 250,000 people with keypoints

Objects in Computer Vision (II)

Objects also used for grounding higher-order actions/attributes/relationships.

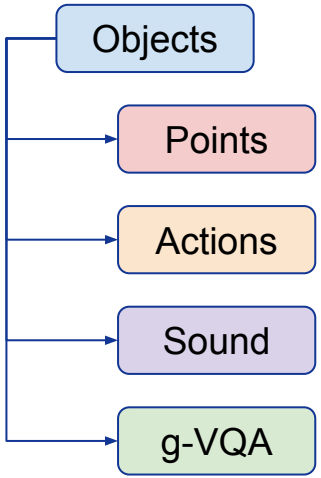


Visual Genome
[Krishna et al., 2016]



Action Genome
[Ji et al., 2020]

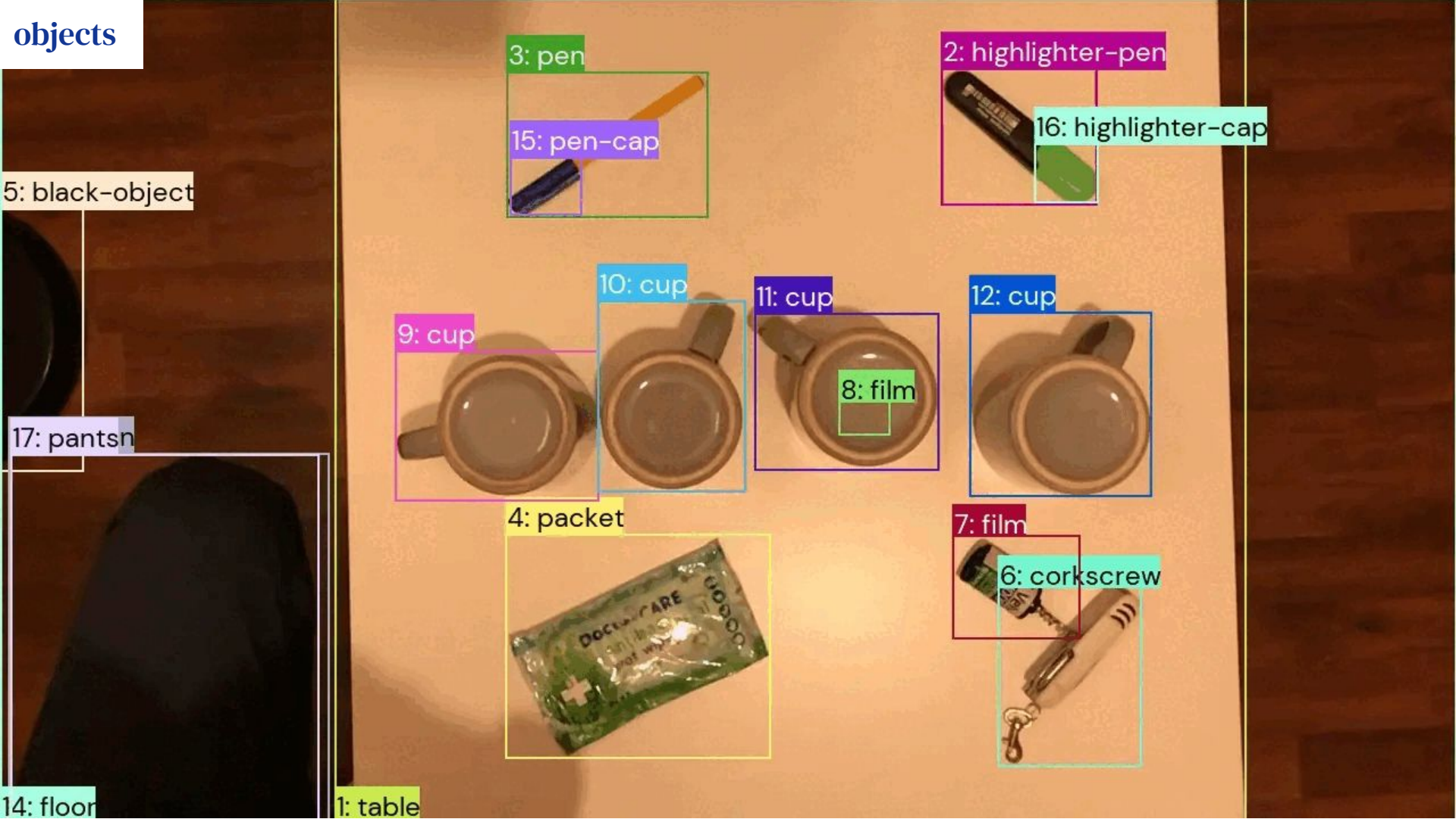
Root Annotation in the *Perception Test*

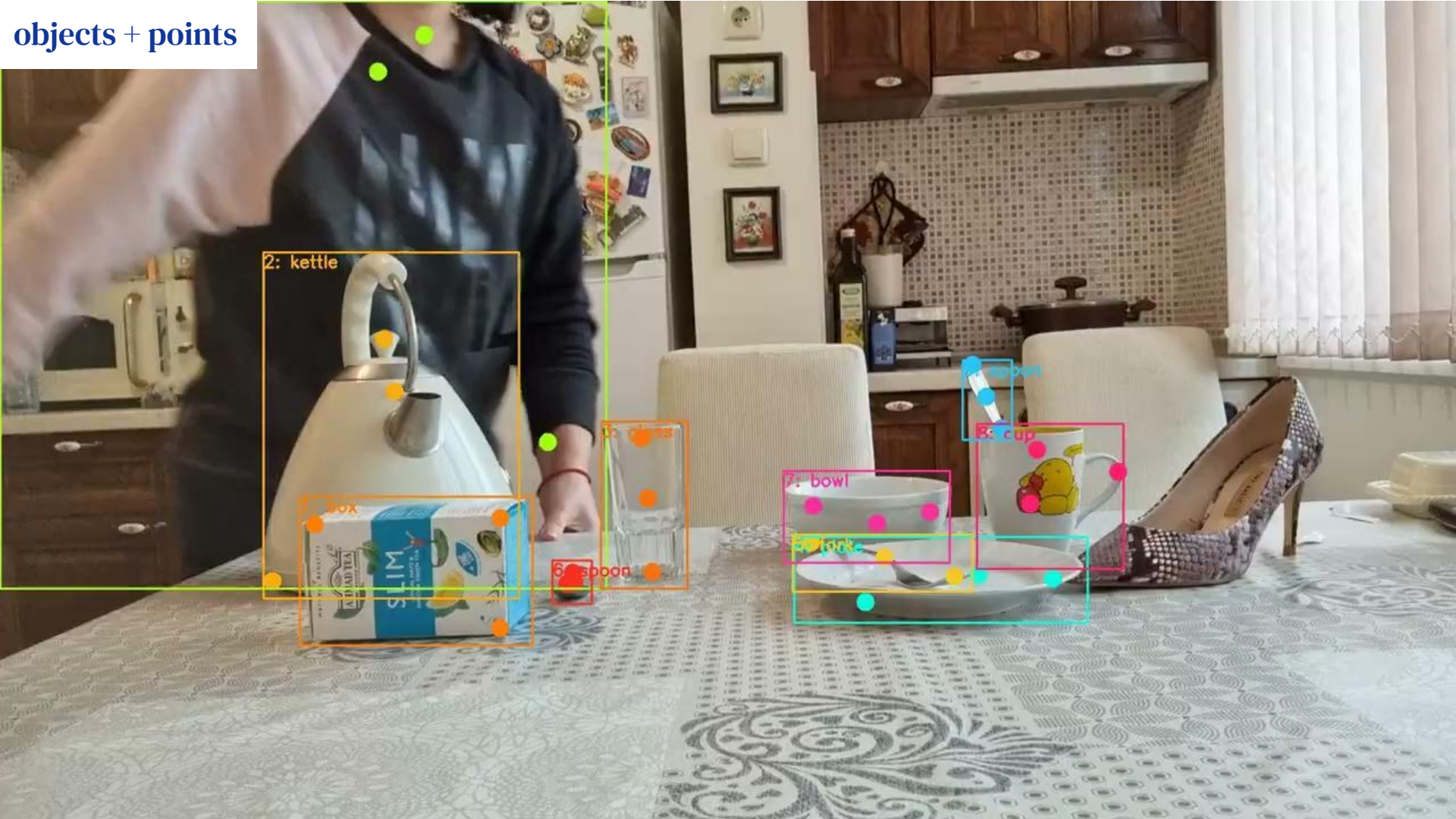


		No. of box tracks
Objects		**191,716**
Points	more spatially precise locations	**3003**
Actions	objects involved in actions	**62,267**
Sound	objects producing sound	**63,824**
g-VQA	objects grounding the answers	**9,374**

not just spatial boxes, but complete trajectories

Total number of videos: 11.6k


objects
3: pen
15: pen-cap
2: highlighter-pen
16: highlighter-cap
5: black-object
10: cup
11: cup
12: cup
9: cup
8: film
17: pantsn
4: packet
7: film
6: corkscrew
14: floor
1: table

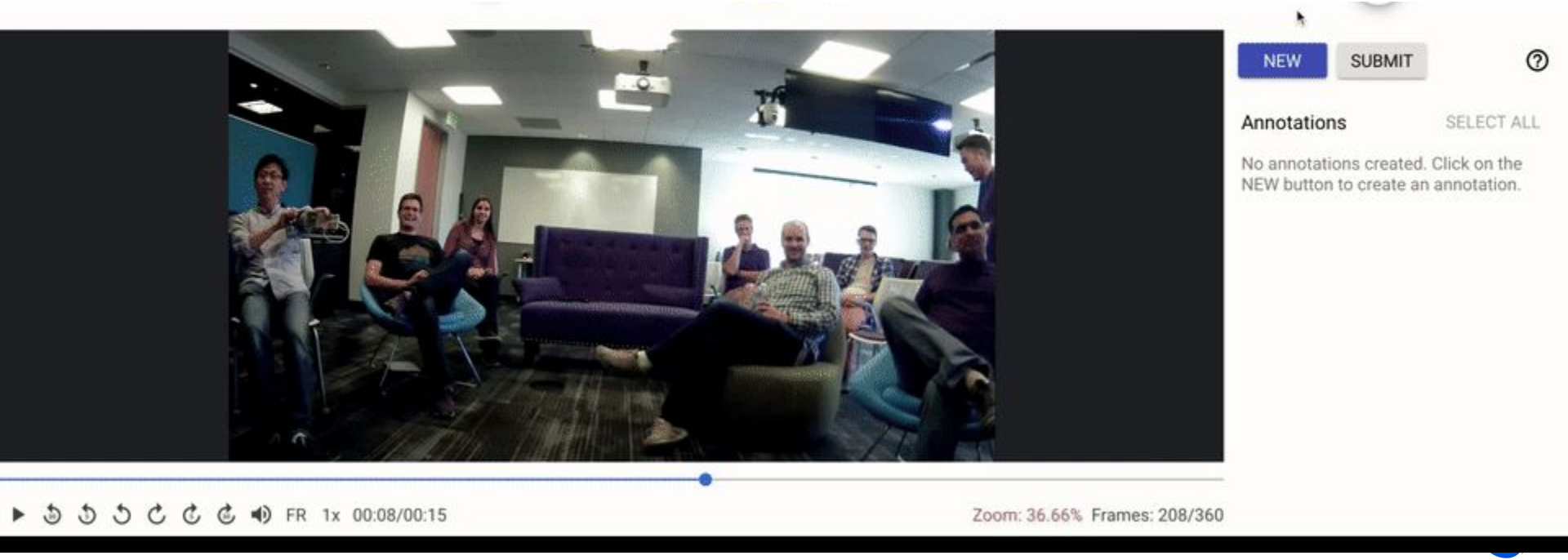

objects + points
2: kettle
1: box
SLIM
3: glass
5: spoon
7: bowl
8: cup

objects + actions

objects + sound

Annotating object tracks

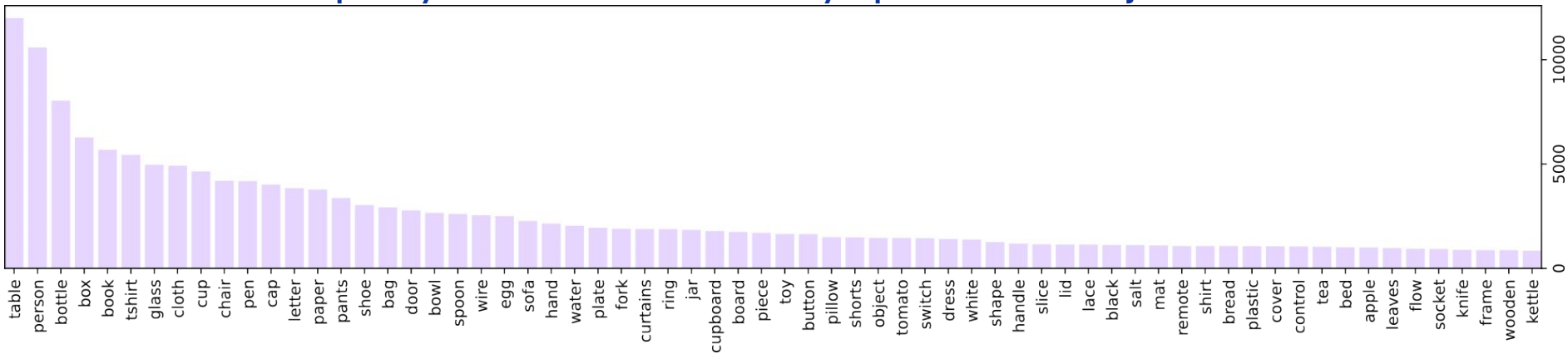
- Free-form (open vocabulary) object names — often with attributes, e.g., “red bag”, “blue bottle” etc.
- Object tracks — axis-aligned bounding-boxes @ 1-fps.



Objects in the *Perception Test*

- ~192,000 object tracks, 11.6k videos
- 5,125 unique object description strings

Frequency of “words” in the manually input free-form object names



Objects in tracking benchmarks

- Closed set of object categories / limited range of objects (e.g., 833 in TAO [Dave et al., 2020]).
- Specific to tracking / limited number of other tasks (e.g., ActionGenome [Ji et al., 2020]).

Dataset	single/ multi	# train/val/test	Avg. duration	# classes
OxUVA	single	200 / 166	142 sec	22
GOT-10k	single	9.5k / 500	15 sec	563
LaSOT	single	1.1k / 280	84 sec	70
TLP (IIIT Hyd)	single	(50)	**500 sec**	---
TAO	multi	500 / 2400	36.8 sec	833
PT	**multi**	**2200 / 3500 / 5900**	23 sec	**5125** (unique strings)

Challenges in Annotation (I)

- Standard definition of single object tracking (SOT) assumes **tracking from the first frame** when the object appears.

- **In ~20% of object tracks, first frame is not representative of object appearance.**
  E.g., when object is entering the frame, only a few pixels might be visible.

- **Use heuristics to find a good initial frame.**
  - object not touching boundaries.
  - frame from which a tracker successfully track for at least the next 2 secs.

Challenges in Annotation (II)

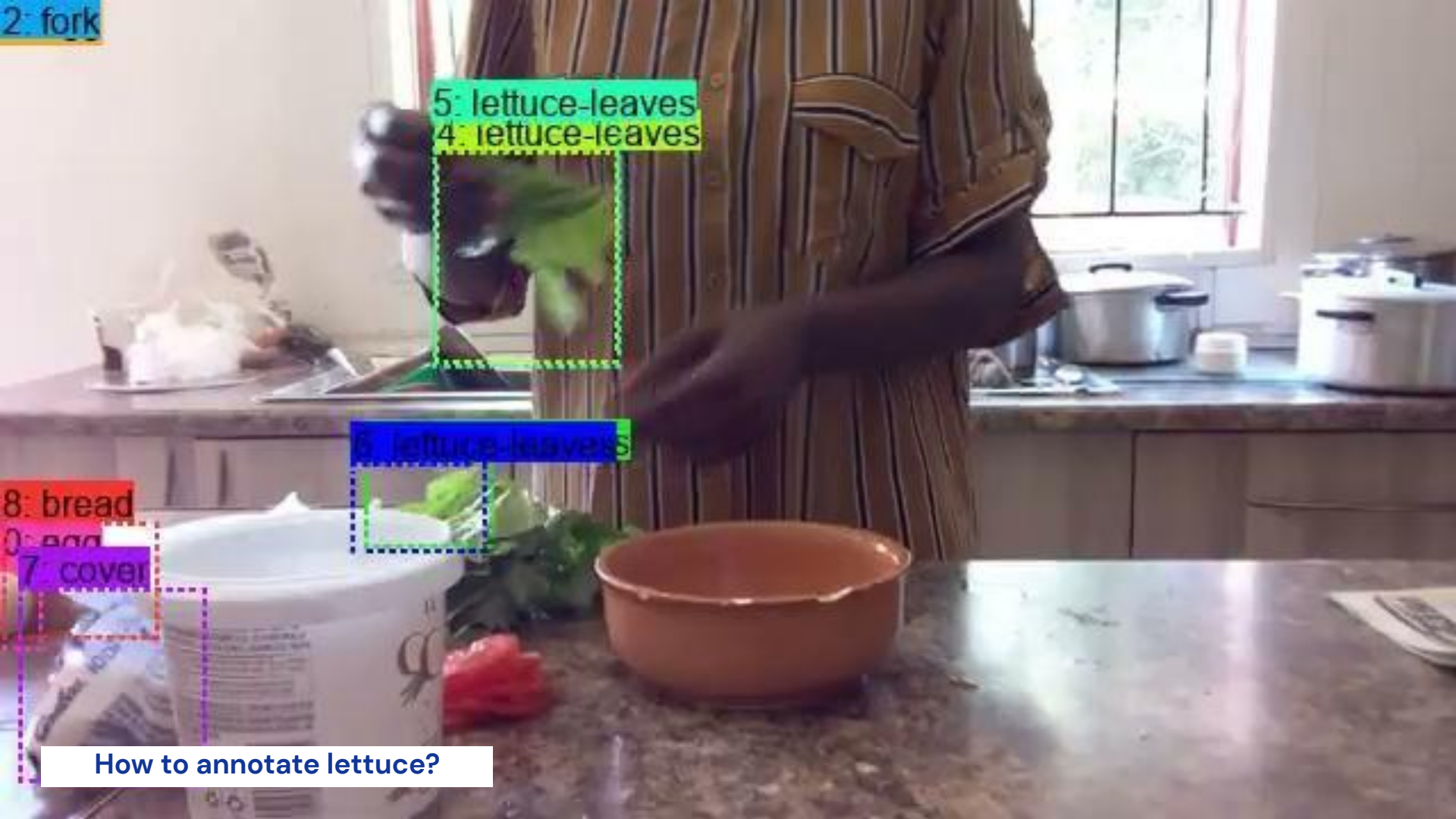
- **Objects that split.**
  - Chopping veggies / salad etc.
  - Lids on boxes / containers.
  - Breaking into sub-parts.

- **Liquids / semi-solids.**
  - Pouring from one container to another.
  - Breaking eggs.

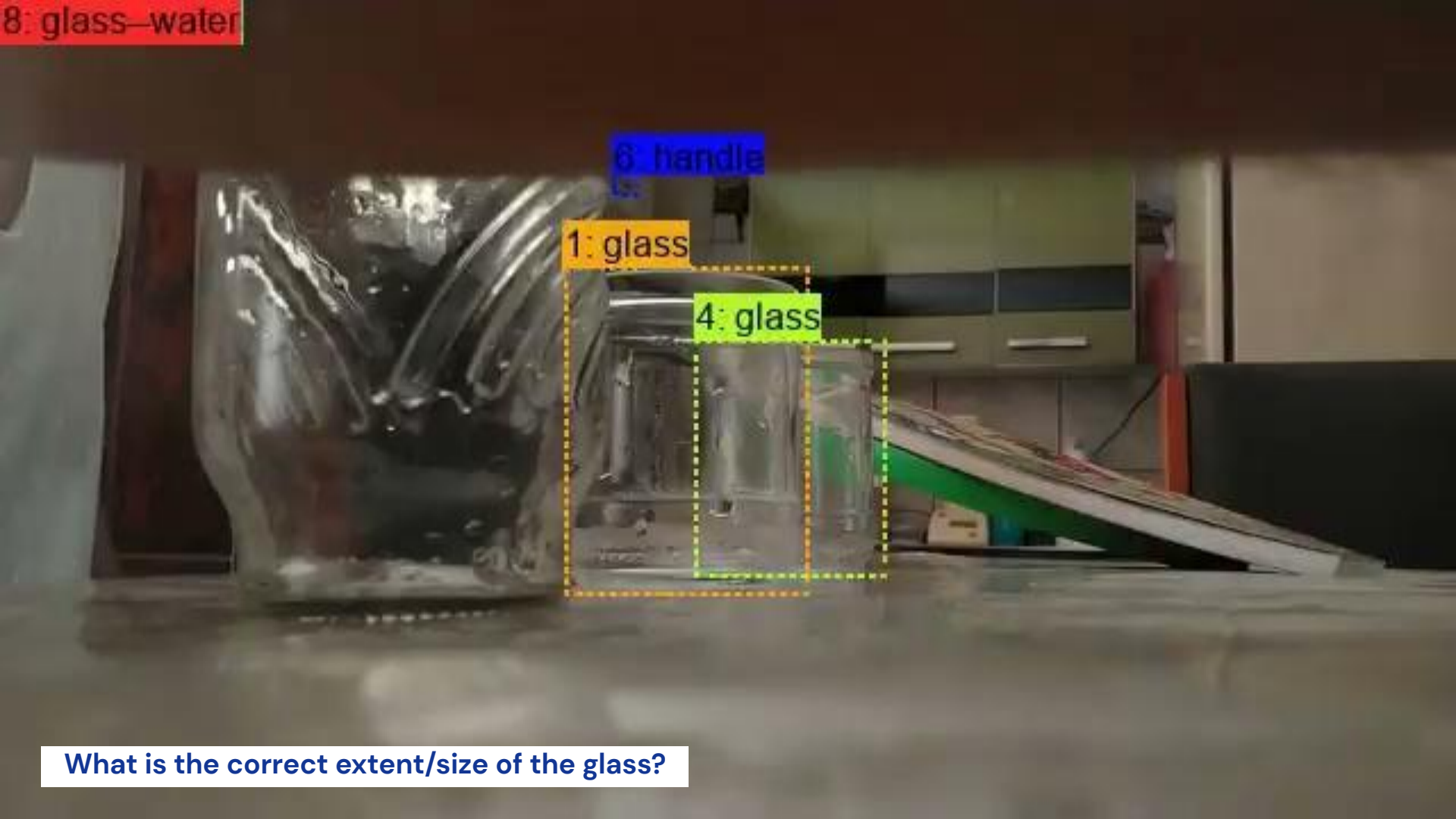
- **(semi-)Transparent containers?**
  - Which object to track after containment.
  - Is the object occluded?

- Difficult to get **consistent** annotations.
- **Currently: Excluded** from evaluation.
- Need better task definitions / models.

How to annotate: egg → egg shell, egg white etc.?



How to annotate lettuce?



What is the correct extent/size of the glass?

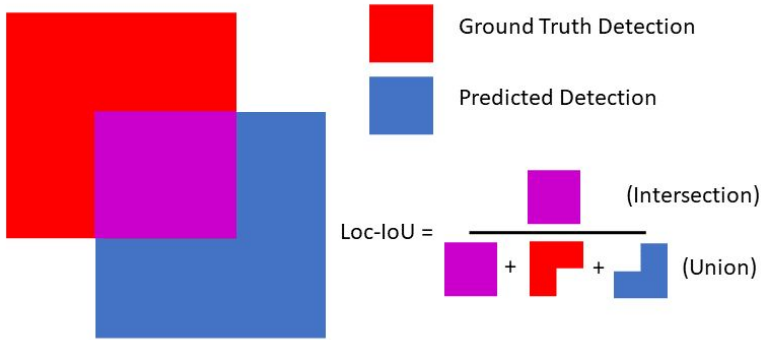
Object Tracking: task and metrics

- **Input:** initial box + all video frames
  Output: boxes in all video frames

- **First frame chosen to be:**
  - **0th human annotated frame: 80%**
  - **1st human annotated frame: 10%**
  - **2nd human annotated frame: 10%**

- **Evaluate using average IoU.**
  (also called *Average Overlap*)
  Intersection-over-union of predicted boxes and ground-truth boxes, averaged over all un-occluded frames.

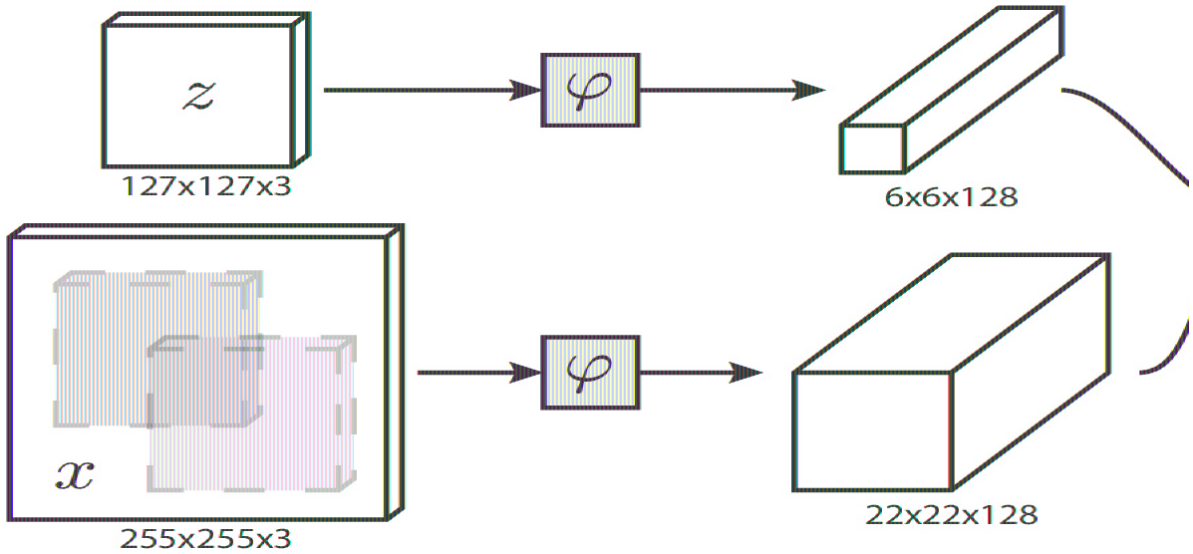
standard in single object tracking benchmarks, e.g., GOT10k [Huang et al.], LaSOT [Fan et al.], etc.



Average Overlap is a preferred metric due to simplicity and strong correlation with other tracking metrics:
L. Čehovin, A. Leonardis, and M. Kristan. Visual object tracking performance measures revisited.
IEEE Transactions on Image Processing, 2016.

SiamFC: baseline method for tracking

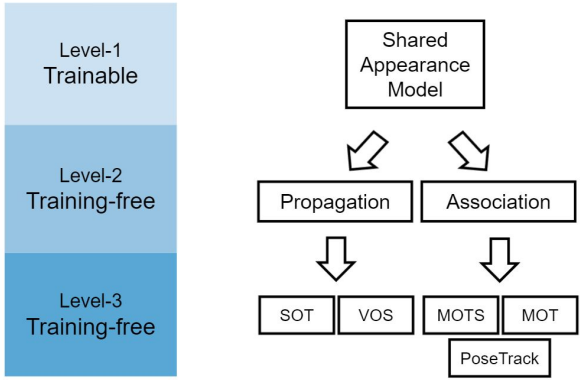
- **SiamFC** template based tracker.
- Cross-correlation based peak finding.
- Class-agnostic / no dataset-specific training.
- Good performance on a number of benchmarks w.r.t. other more complex ones.



Fully-Convolutional Siamese Networks for Object Tracking
Luca Bertinetto *, Jack Valmadre *, João F. Henriques, Andrea Vedaldi, Philip H.S. Torr, ECCVw 16

SiamFC: Implementation details

- Follow **UniTrack** [Wang et al., NeurIPS 2021].
- ImageNet pre-trained ResNet-50 features.
- Template size fixed to 520x520 pixels; maintain aspect ratio.
- Multi-scale search windows (3 scales).
- Spatial prior preferring closeness to previous object location.



Single Object Tracking (SOT) @ OTB-2015 [ref: UniTrack]

SiamFC: baseline results (I)

Object tracking	Static or slightly shaking camera	Moving camera
all objects	0.6854	0.5322
action objects	0.5396	0.4653
sound objects	0.6094	0.5269
g-vQA boxes	0.5130	0.4587

- Approx. 65% of all object tracks are static/ don't move much.
  - => 'all objects' + 'static camera' >> 'action object' + 'moving camera'
- Object occlusions/ panning cameras — failure modes for template based tracker.
- Small sized objects / fast moving objects — also challenging.

SiamFC: baseline results (II)

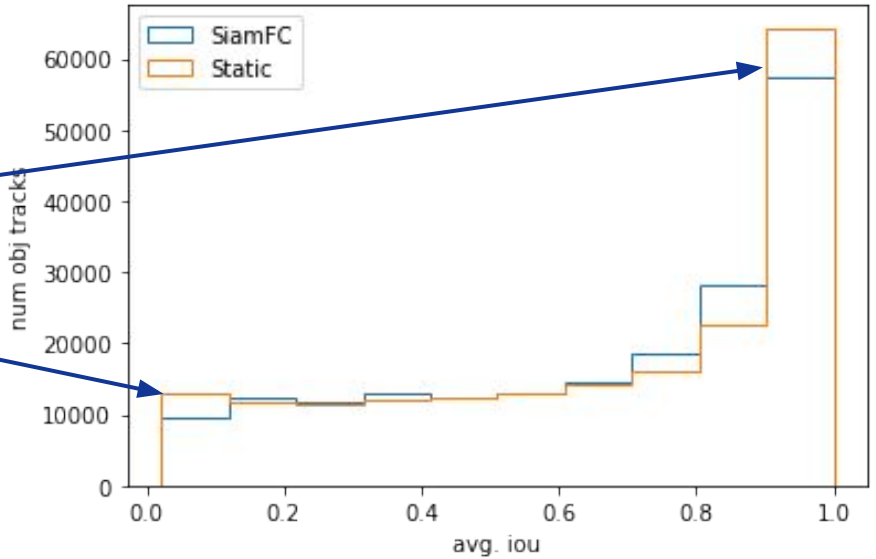
- Compare against ‘static’ baseline:
  use the initial box as the prediction.

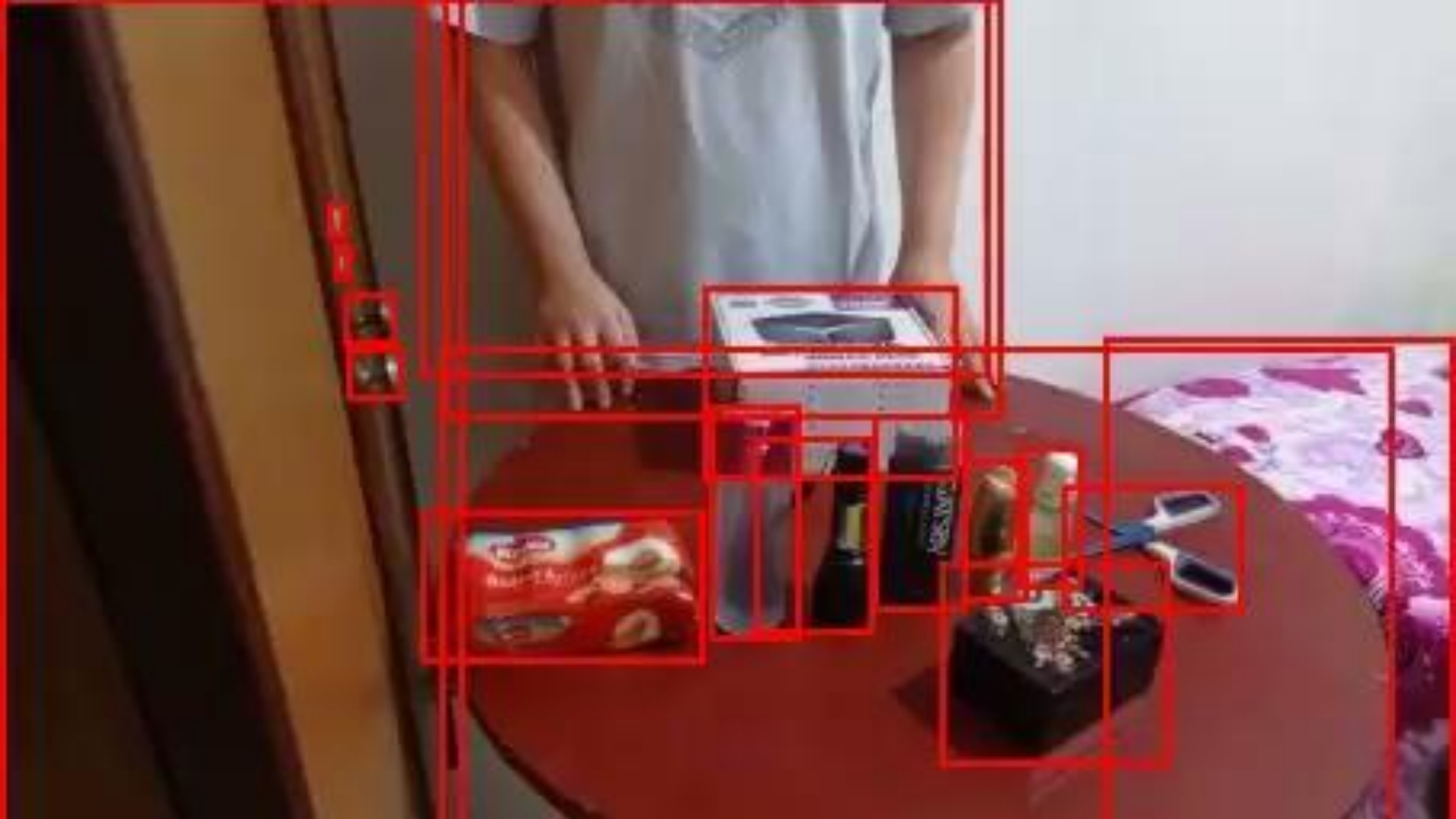
- SiamFC worse for ‘static objects’
- better for ‘moving objects’

  .

- **Note:** this is with a single initialization of the tracker, as opposed to multiple-restarts at failure also considered in the community.

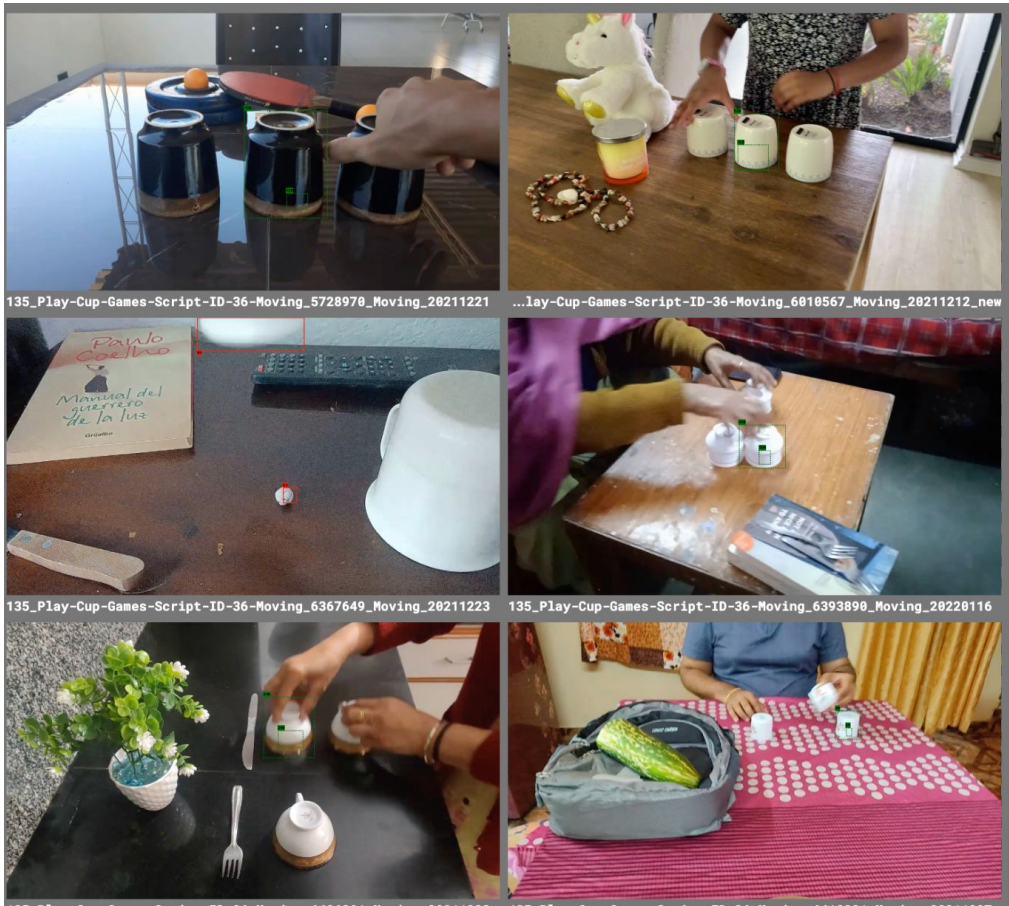
- Panning cameras / fast moving / small objects – challenging for trackers.





Probing object permanence with *Cup Games*

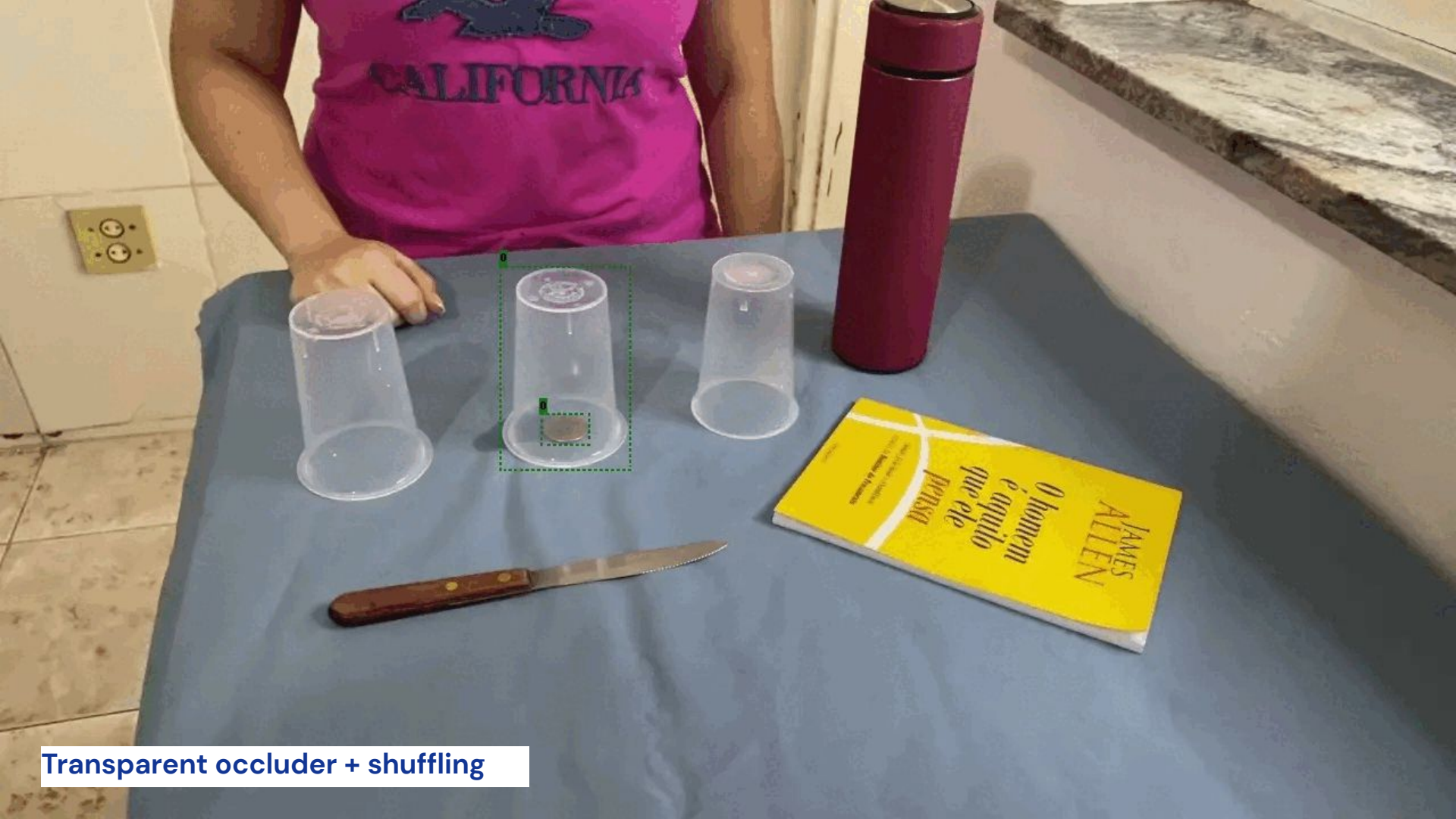
- Probe complex motion / occlusion interactions with **Cup Games** tracking tasks.



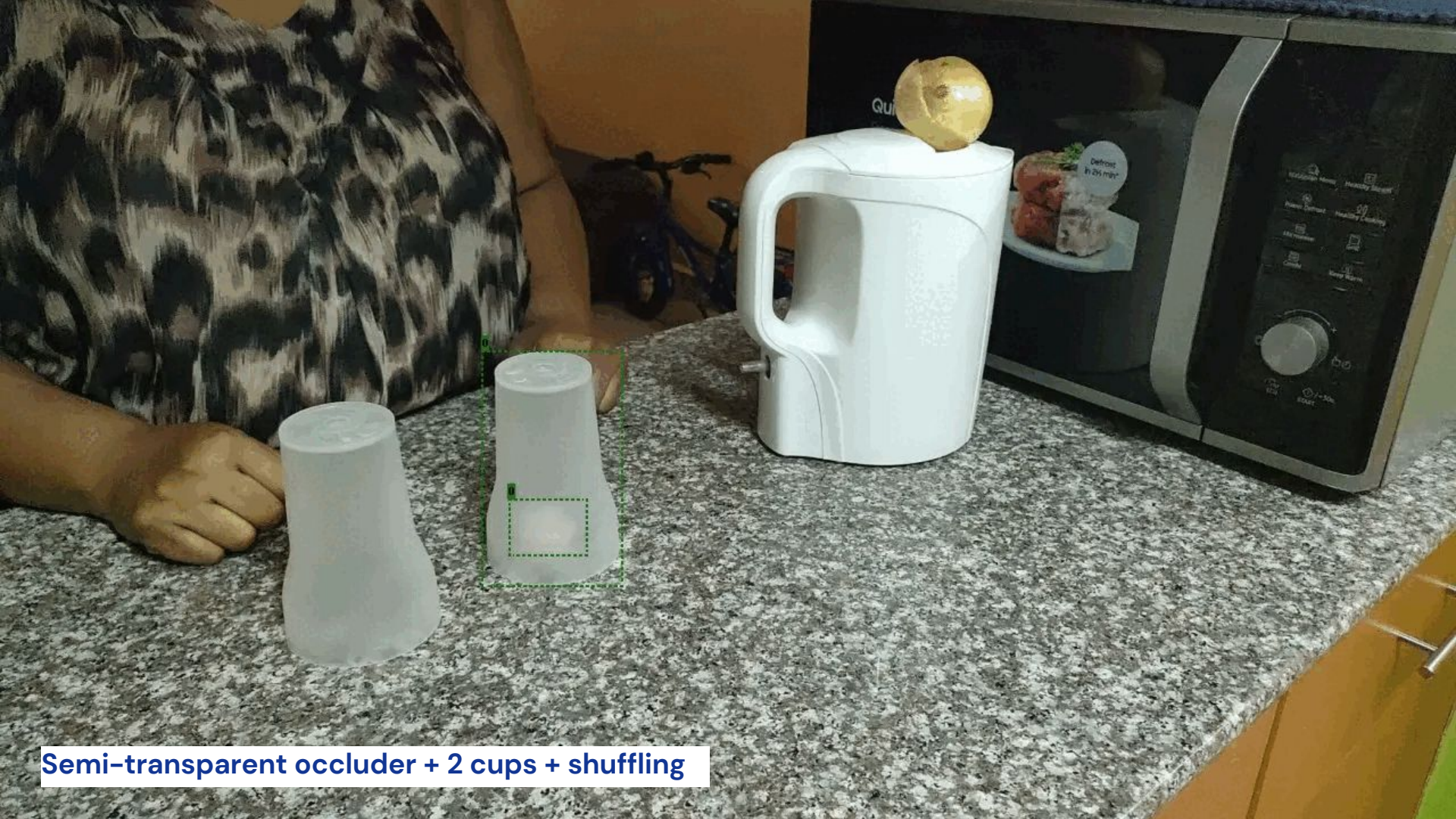
Statistics

- 598 cup-games videos.
  - 483 videos with identical cups (hard)
  - 113 videos with transparent cups (easier)

- 3 cups: 451 videos
  2 cups: 132 videos
  4 cups: 34 videos



Transparent occluder + shuffling


Semi-transparent occluder + 2 cups + shuffling

Opaque occluder + distractor action

Cup Games: Task definition + metrics

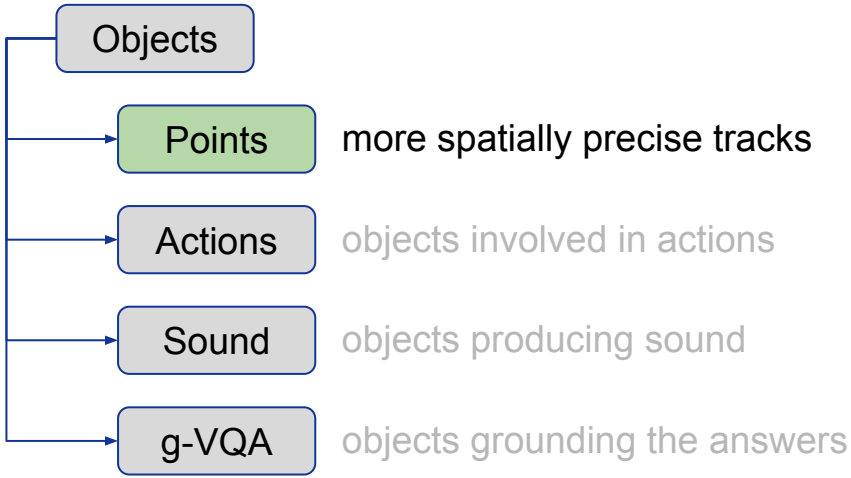
- **Given:** an initial box + all video frames
  but track through strong occlusions, motion, identical objects.

- Additional **visibility/occlusion** flag used to determine the target (occluding / occluded object) for tracking.

- **Metric:** intersection-area / predicted-box-area
  As when object is occluded, exact location/size is unknown.

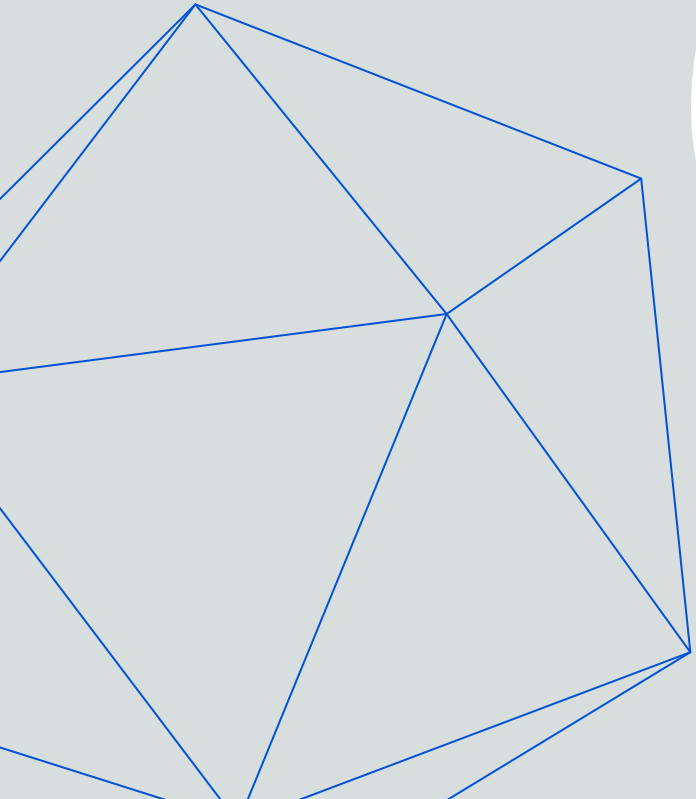
Challenge for tracking algorithms + finer-grained analysis.

Up Next..


Objects
Points
more spatially precise tracks
Actions
objects involved in actions
Sound
objects producing sound
g-VQA
objects grounding the answers

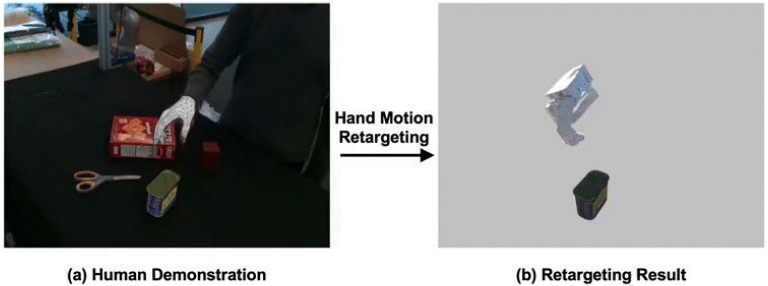
DeepMind

Point tracking

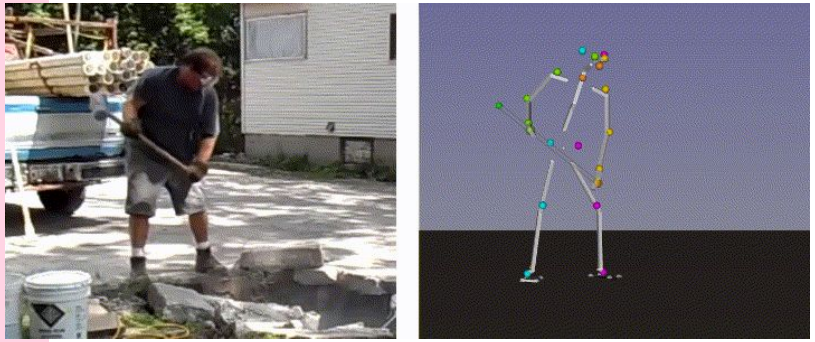


CV Grand Challenge: Physical Scene Understanding

Peng et al. 2018 “Learning Acrobatics by Watching YouTube”



Ye et al. 2022 “Learning Continuous Grasping from Human Demonstrations“

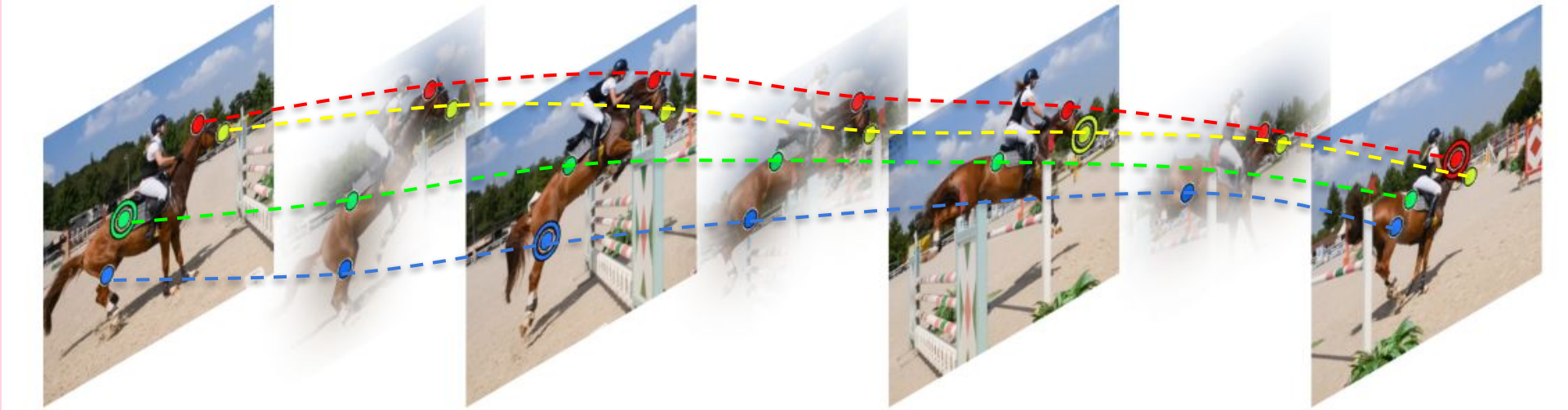


Sivic et al. 2019 “3D Motion and Forces of Person-Object Interactions from Video”

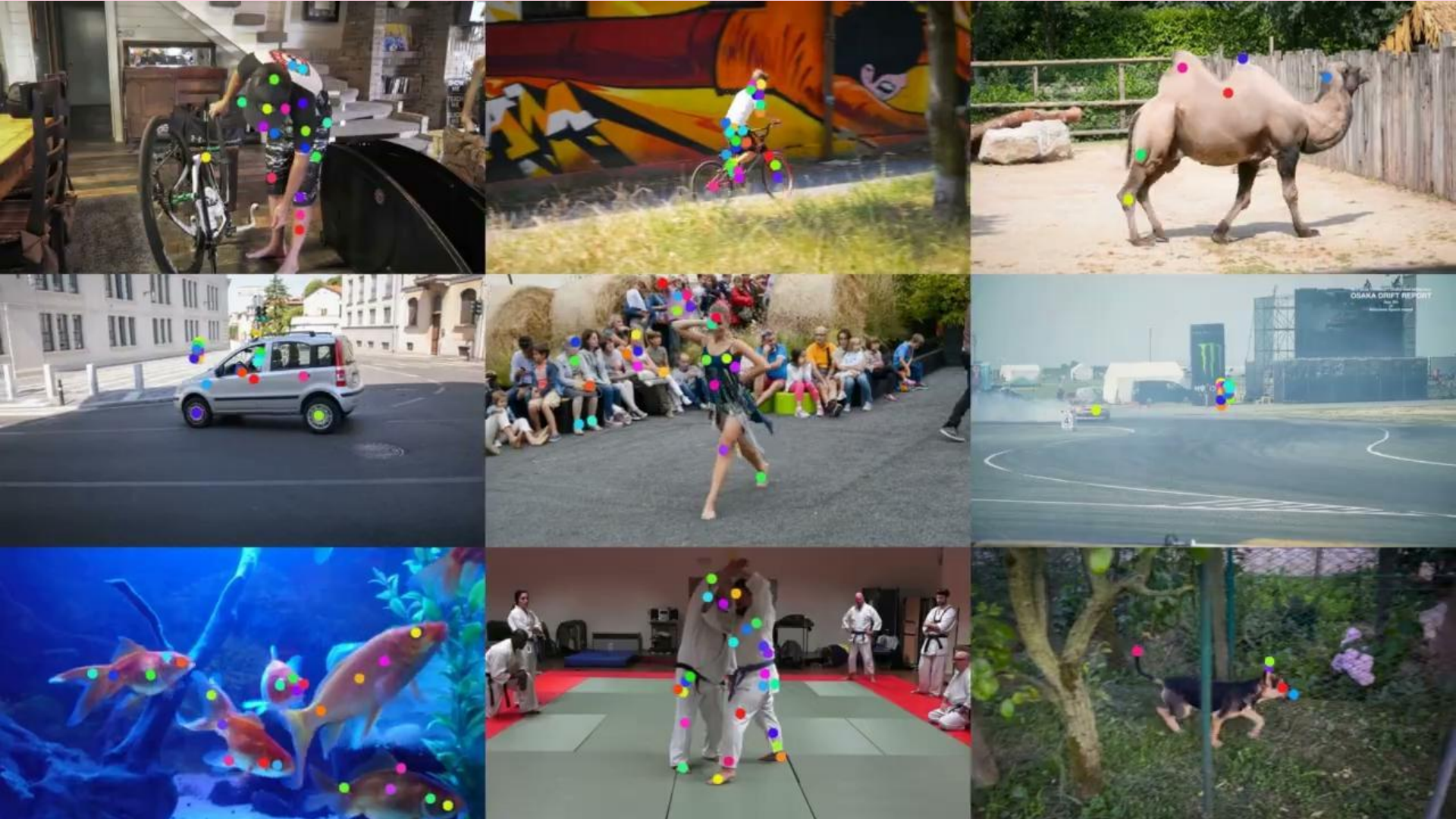
Ehsani et al. 2020 “Use the Force Luke! Learning to Predict Physical Forces”

Goal: track any point on any solid surface in video

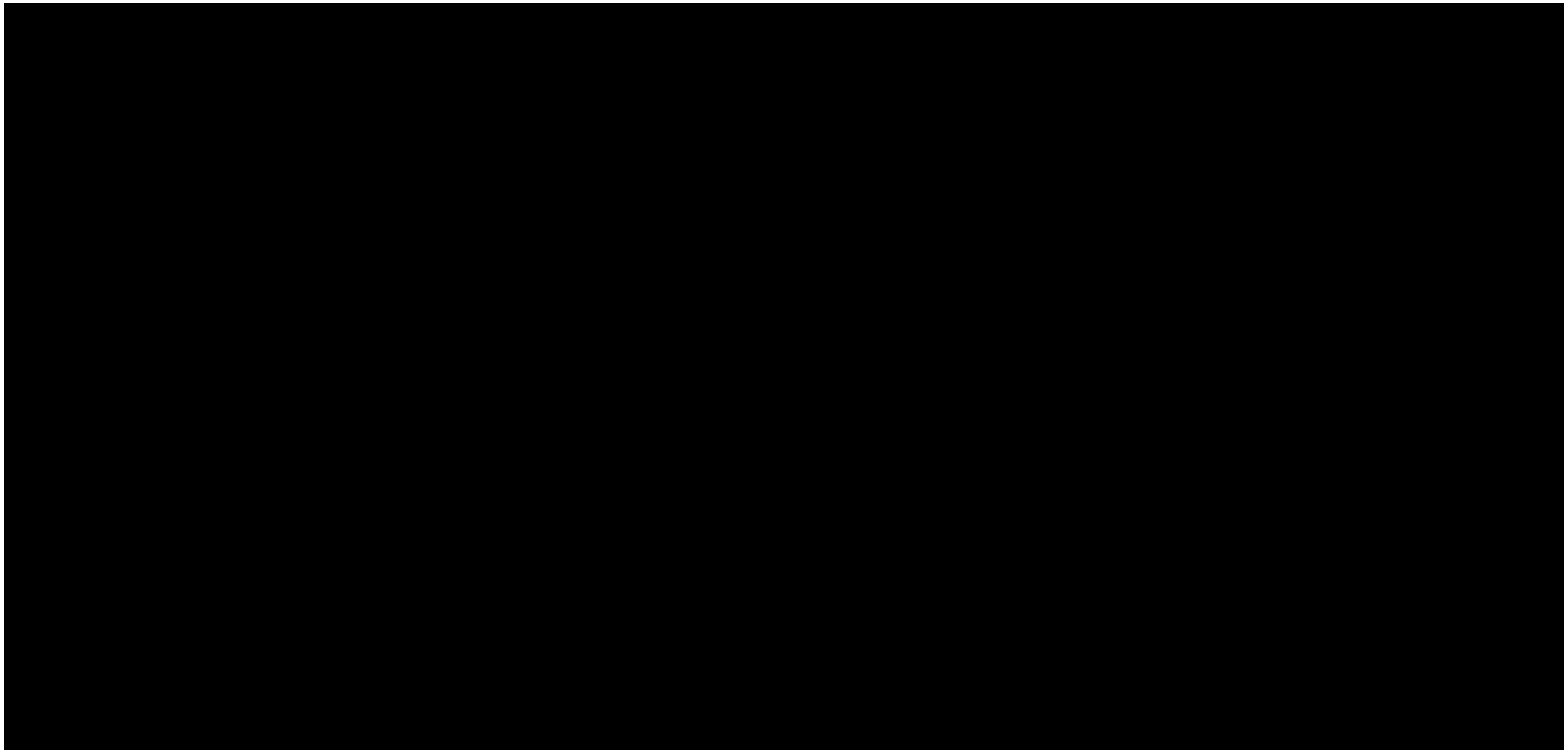
TAP: Tracking Any Point



- Input: a video
- Input: a set of query points on any frame
- Output: trajectories + occlusion masks for each query point across all other frames


OSAKA DRIFT REPORT

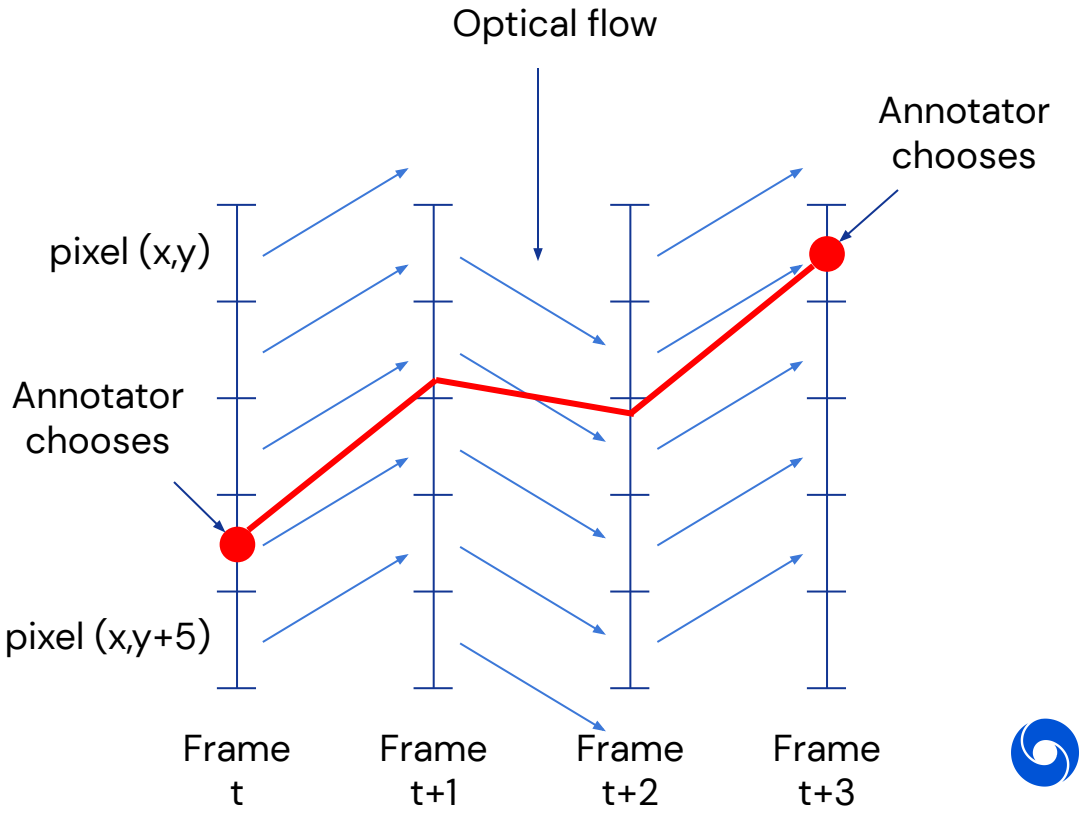
Annotation Interface



(Kuznetsova et al. 2021)

Optical Flow Interpolation

Optical flow (RAFT)



Improvements from Optical Flow

Without flow interpolation

With flow interpolation

TAP-Vid-DAVIS (30 videos, 650 points, 25 fps)

TAP-Vid-Kinetics (1189 videos, 31301 points, 25 fps)

TAP-Perception-Test (145 videos, 8574 points, 30 fps)

Point Annotation Quality

- Two humans annotate a real video on the same point track

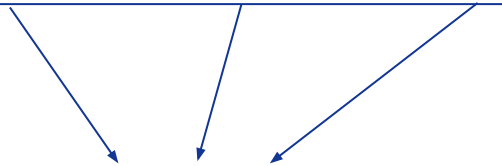
Human v.s. Human	Occlusion Agreement	Distance < 4 pixels	Distance < 8 pixels
Percentage	95.5%	92.5%	98.7%

- Human annotate a simulated video with ground truth point track

Human v.s. Groundtruth	Occlusion Accuracy	Distance < 4 pixels	Distance < 8 pixels
Percentage	96.9%	96.2%	99.5%

Dataset statistics

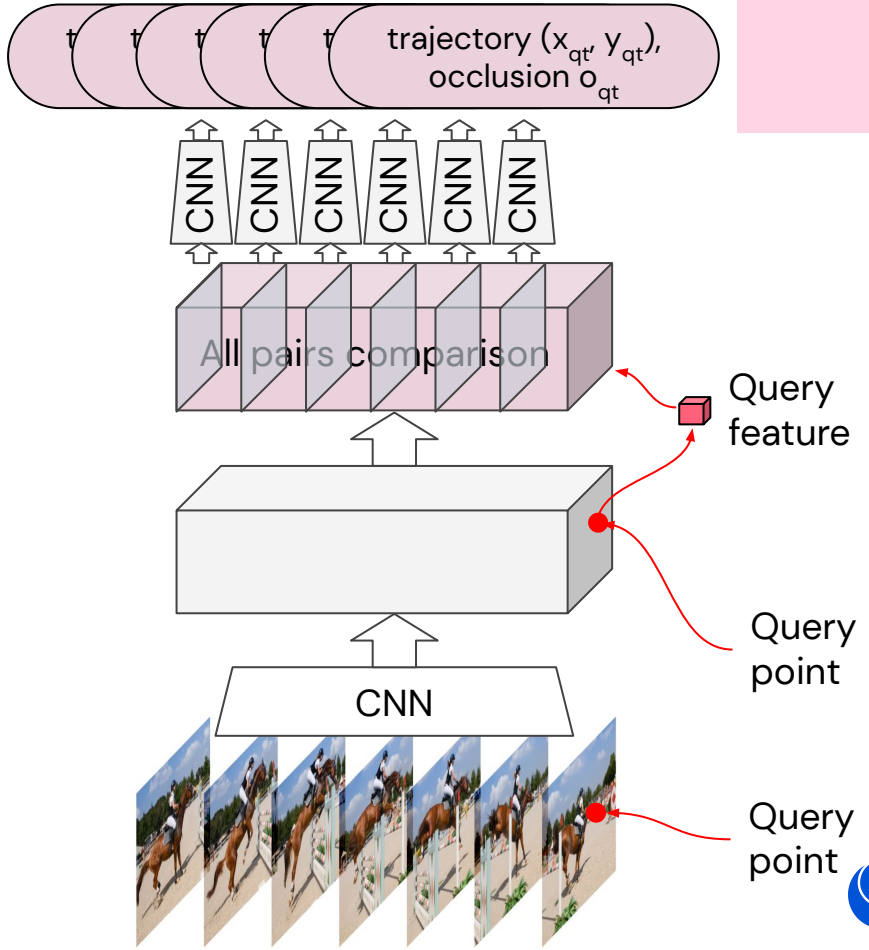
Dataset	# Videos	Avg. Points	# Frames	Resolution	Sim/Real	Train/Eval
TAP-Vid-Kinetics	1,189	26.3	250	≥720p	Real	Finetune/Eval
TAP-Vid-DAVIS	30	21.7	34-104	1080p	Real	Eval
TAP-Perception-Test	145	59.1	1000	1080p	Real	Finetune/Eval



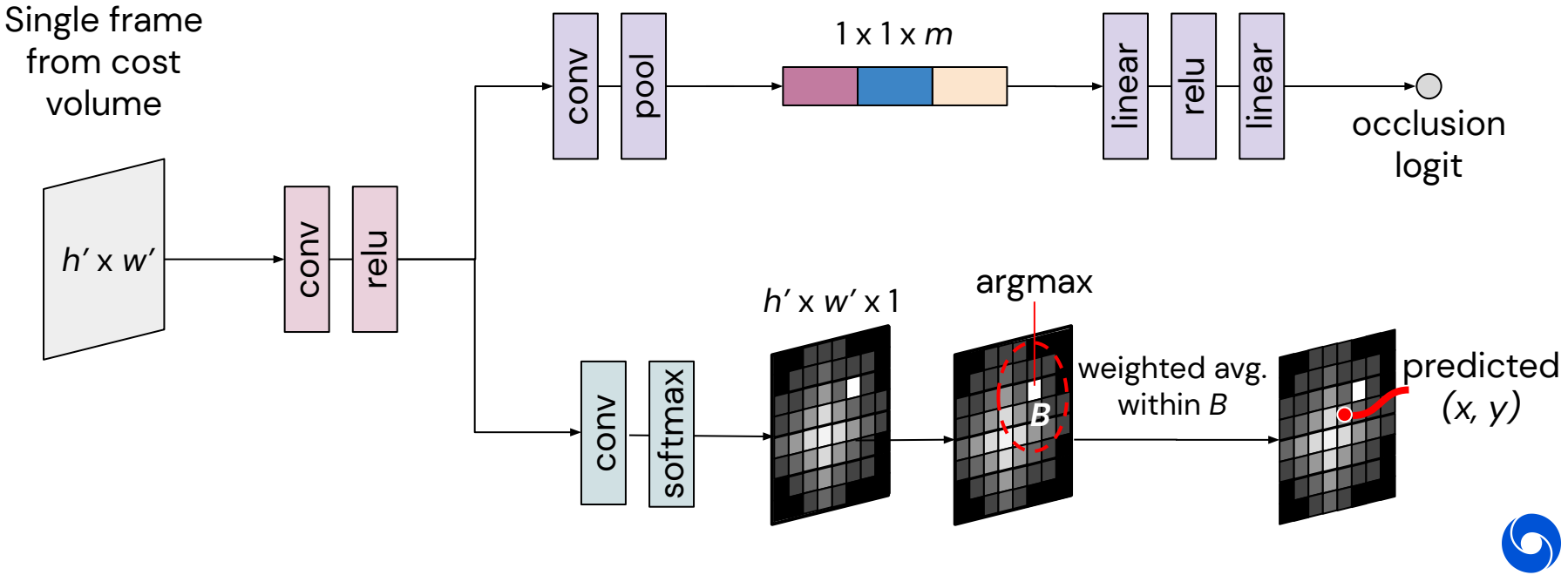
40,000 point tracks, 7,000,000 points

Baseline architecture: TAP-Net

- Input: a video and a query point
- Apply a ConvNet
- Extract the query point features
- Compute a “cost volume” comparing query feature to all others
- Post-process to make a prediction for every frame

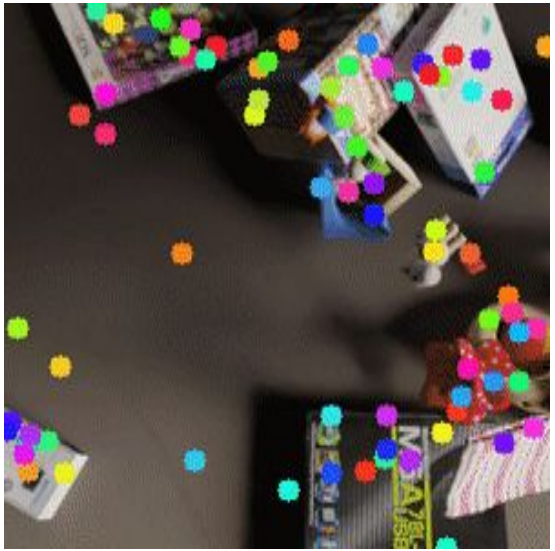


Predicting Position and Occlusion



Training data?

- Can’t get enough from the real world!
- But tracking is low-level, like optical flow.
- Kubric: a scalable dataset generator
- Bullet physics + Blender raytraced rendering
- 24-frame clips with a few objects thrown into a scene
- Perfect point tracking

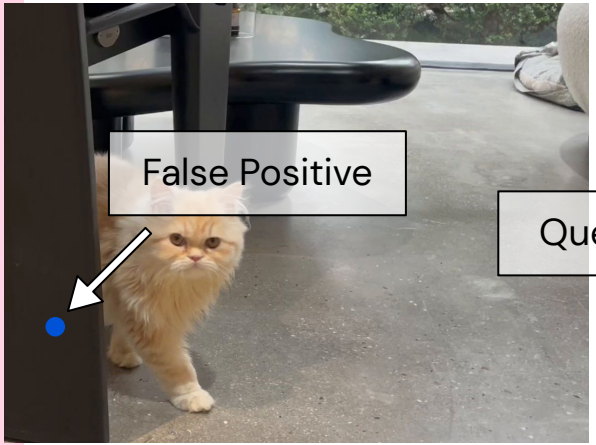


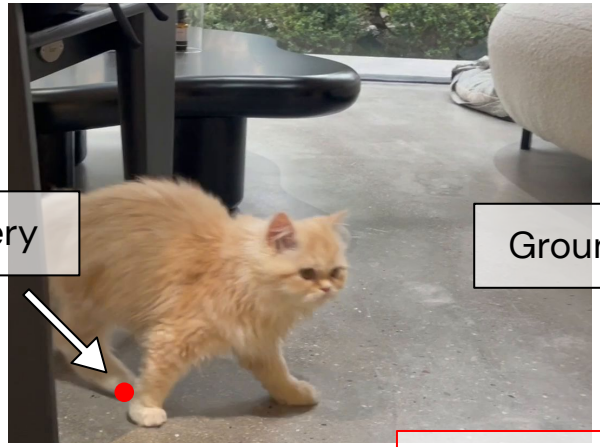
TapNet Prediction - Qualitative

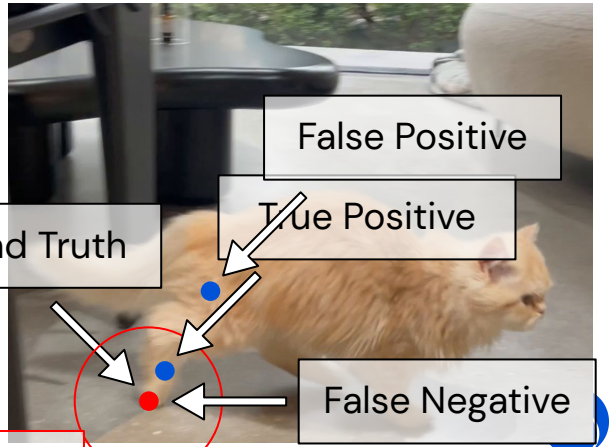
Performance Metric: Average Jaccard

- Follows prior box tracking work:

$$\text{Jaccard score} = \frac{\text{true\_positives}}{\text{true\_positives} + \text{false\_positives} + \text{false\_negatives}}$$

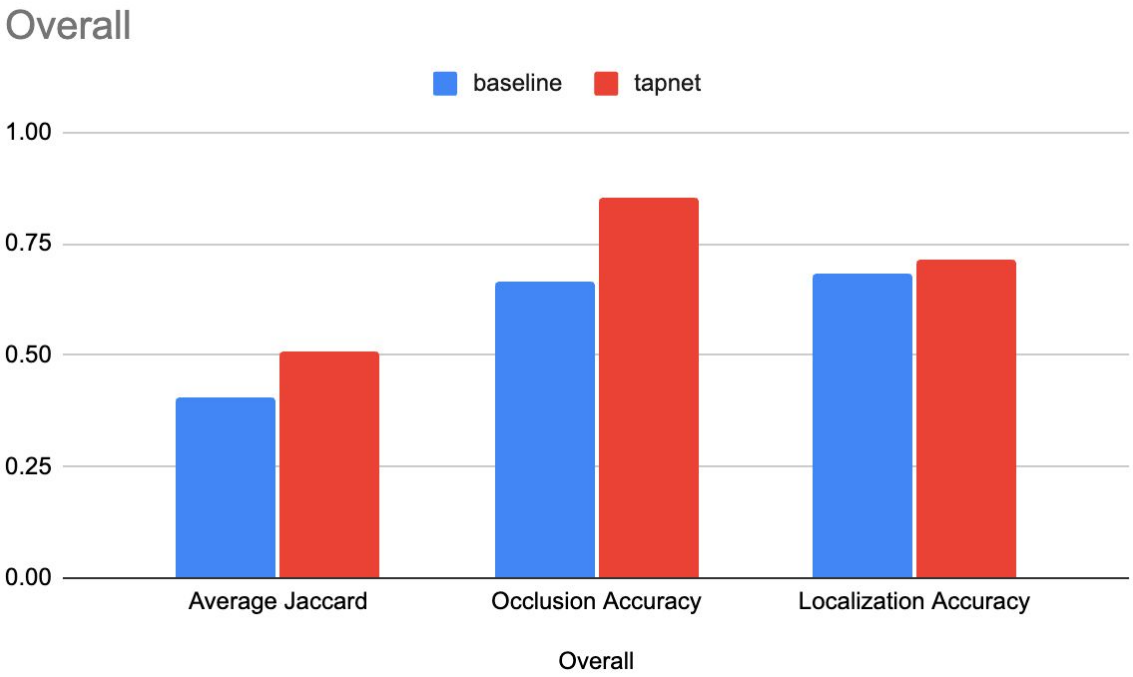






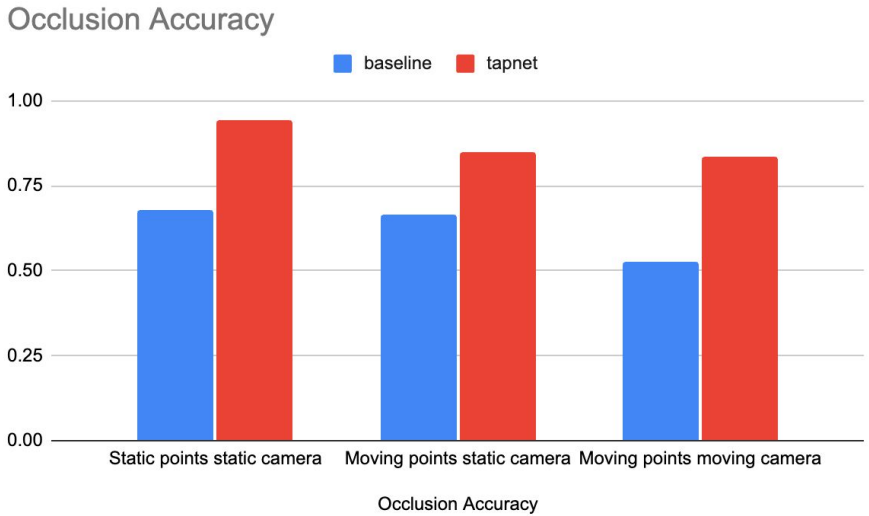
TapNet Prediction - Quantitative

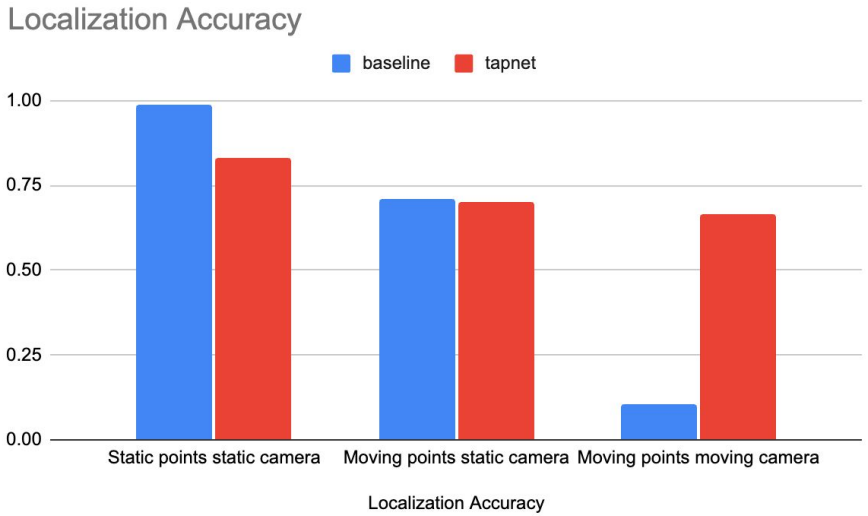
Baseline: assume query points never move



TapNet Prediction - Moving v.s. Static

Baseline: assume query points never move



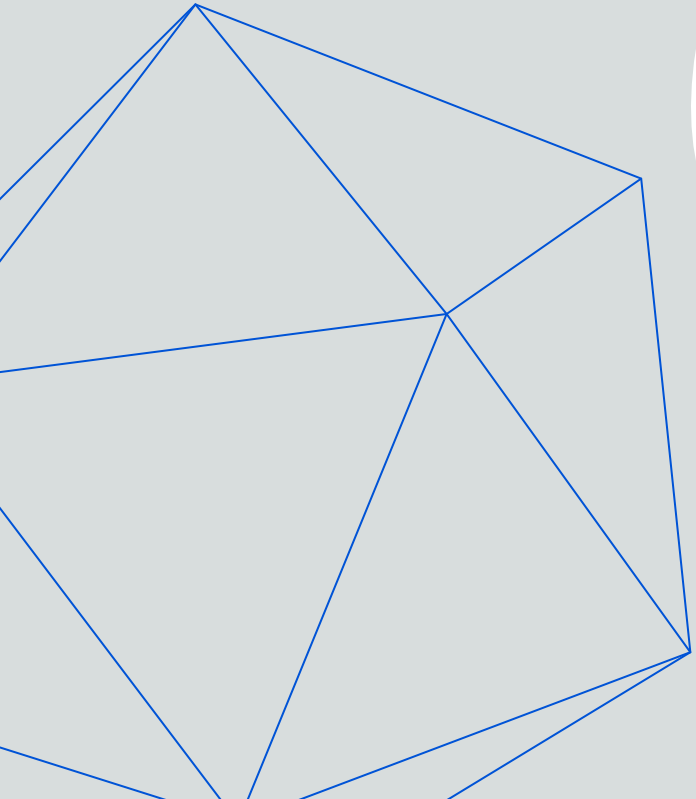


Conclusion

- Point/surface tracking is underexplored, but useful
- Non-trivial to annotate: track assist makes it possible
- First-of-its-kind dataset
- TAP-Net+Kubric performs better-than-chance
- But still a long way to go!

DeepMind

Action and sounds




URSASU

Describe the previous video

There is a person using a hair dryer to dry an orage while moving the orange and a pair of scissors with their other hand.

Describe the previous video

There is a **person** **hair dryer** to an **orage** while **the orange** and a **pair of scissors** with their **other hand**.

Describe the previous video

There is a **person using a hair dryer** to **dry** an **orage** while **moving the orange** and a **pair of scissors** with their **other hand**.

Actions in the Perception Test

3 types of actions annotations in the Perception Test

1

Segments of actions in videos

- We provide temporal boxes (initial timestamp, end timestamp) for any action in the video.

2

Class of the action segments

- We provide a class for every of the segment in the video.
- Important for action detection.

3

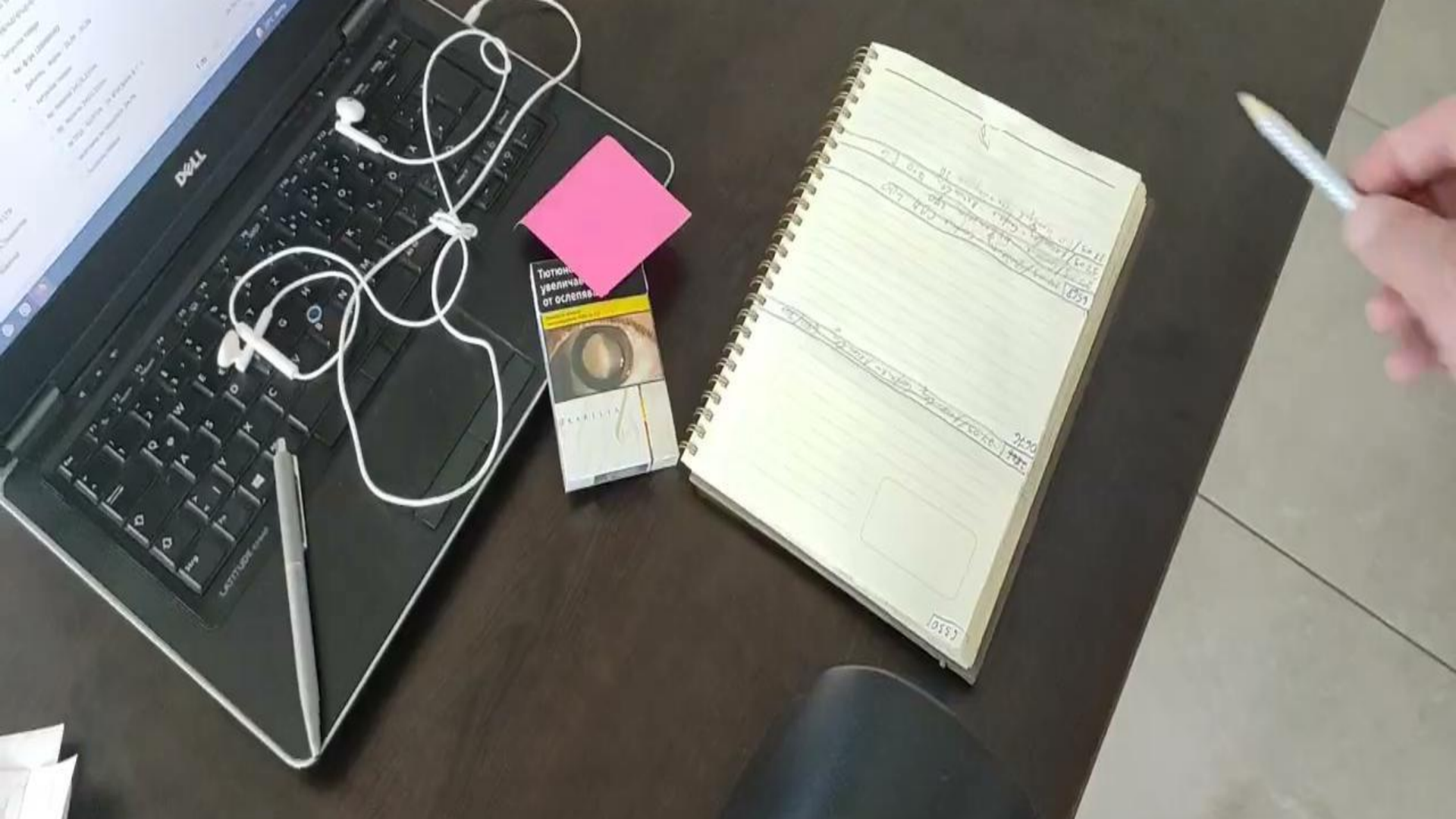
Action segments are tied to the objects that produce the actions

- This connects the action annotations with other annotations in the dataset.
- These annotations enable tasks such as action localization in video.

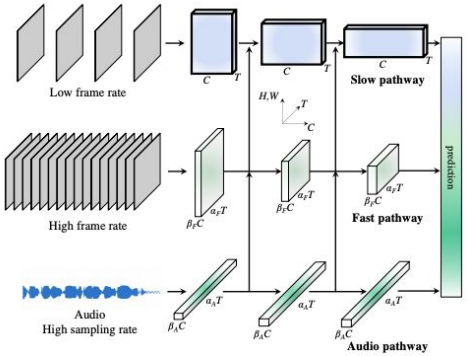

Signature
POTATO CHIPPER
Signature
Care & Use Instruct
MACMILLAN
English
DICTIONARY
GLAM SQUAD KE


URSASU

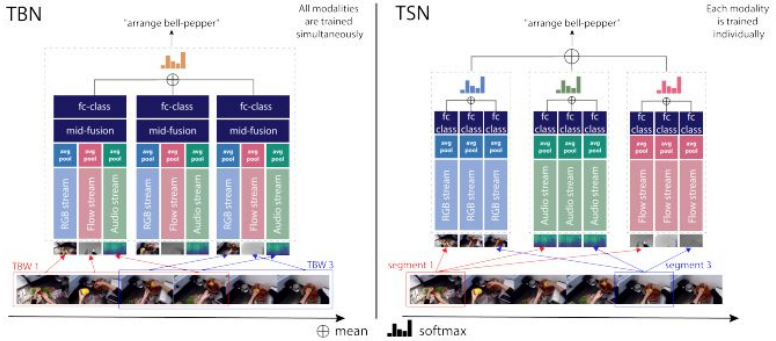
[Listen to the audio]


DELL

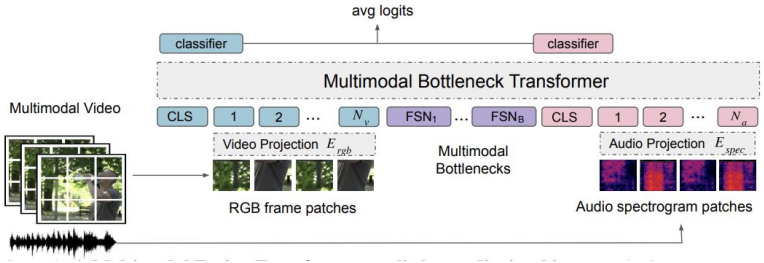
Lots of works on audio-visual modelling (supervised)



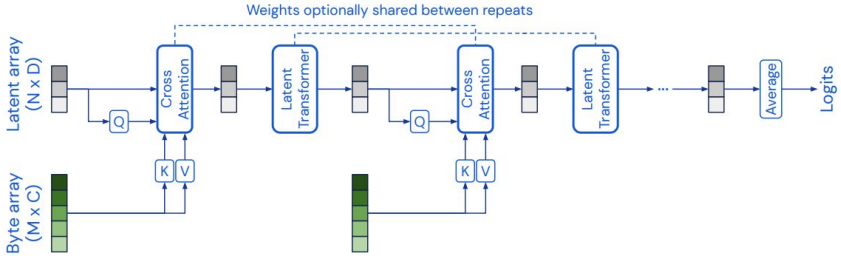
AVSlowFast: Xiao et al, 2020



TBN: Kazakos at al, 2019

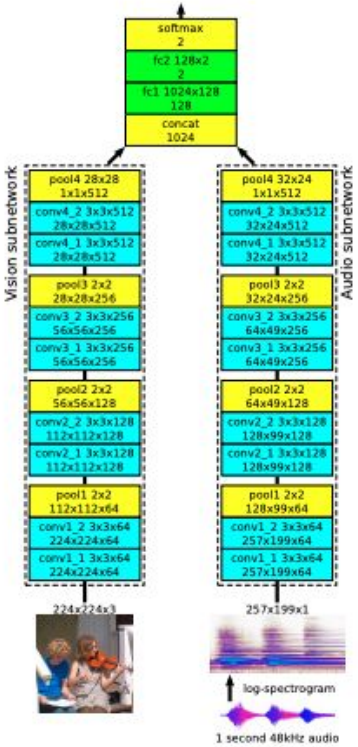


Attention Bottlenecks for Multimodal Fusion: Nagrani et al, 2022

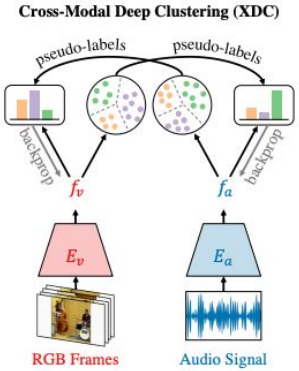


Perceiver: Jaegle at al, 2021

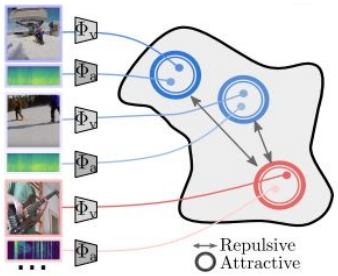
Lots of works on audio-visual modelling (self-supervised)



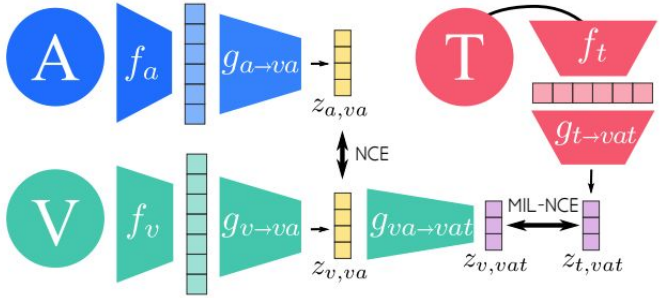
L3: Arandjelovic and Zisserman, 2017



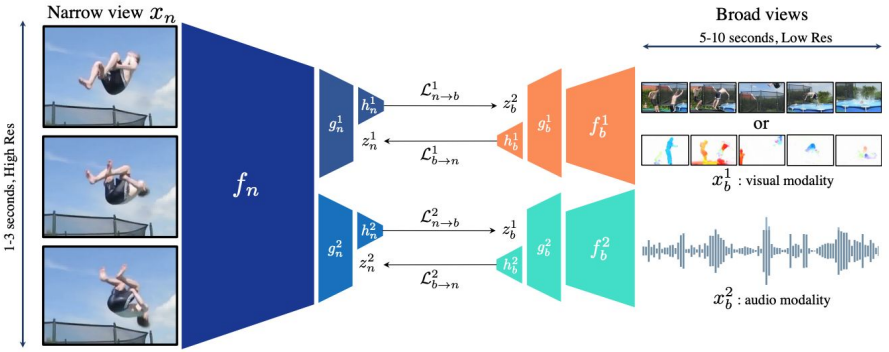
XDC: Alwassel at al, 2020



GDT: Patrick at al, 2020



MMV: Alayrac at al, 2020



Brave: Recasens at al, 2021

However...!

- Available benchmarks are mostly focused on classification.
- Benchmarks that include other tasks are speech-centric (e.g. Ego4D)
- We lack sound-centric benchmarks where audio annotations are linked to other tasks annotations.
- Available benchmarks do not include questions about audio using natural language.

Audio in the Perception Test

3 types of audio annotations in the Perception Test

1

Segments of sounds in videos

- We provide temporal boxes (initial timestamp, end timestamp) for any sound in the video.

2

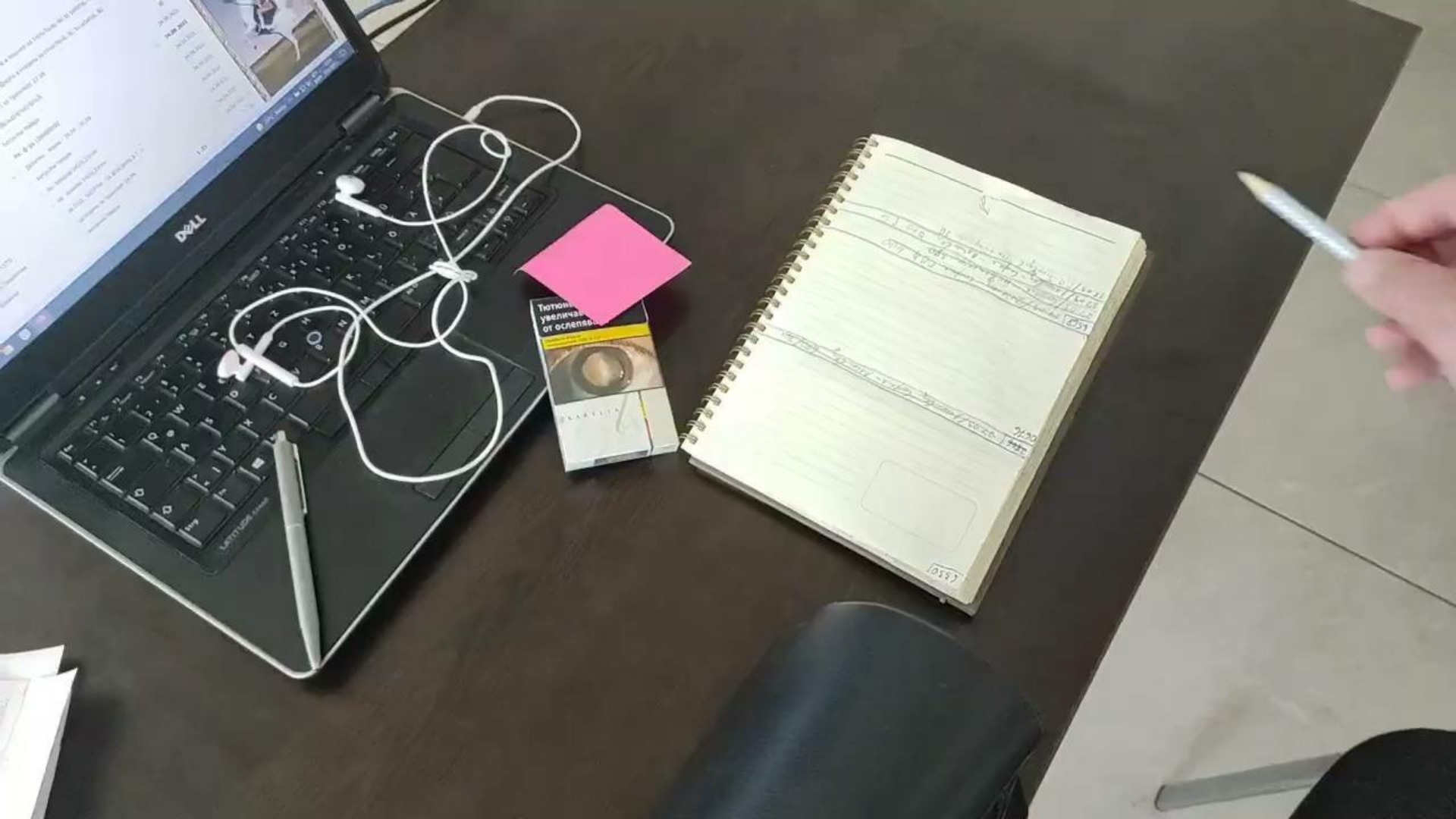
Class of the sound segments

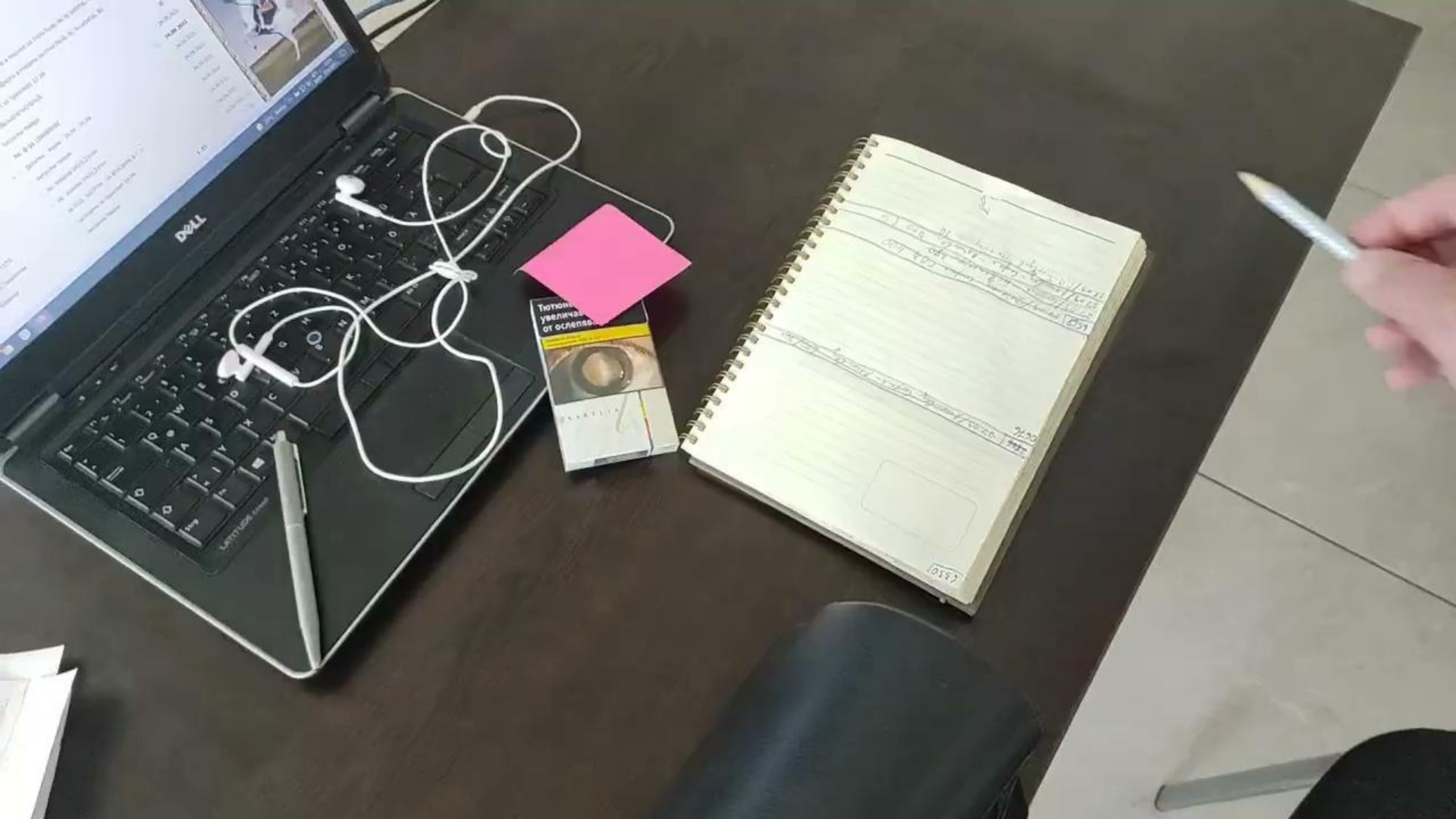
- We provide a class for every of the segment in the video.
- The list of classes enable distinguish between interactions, human sound, animal sound and object sound.

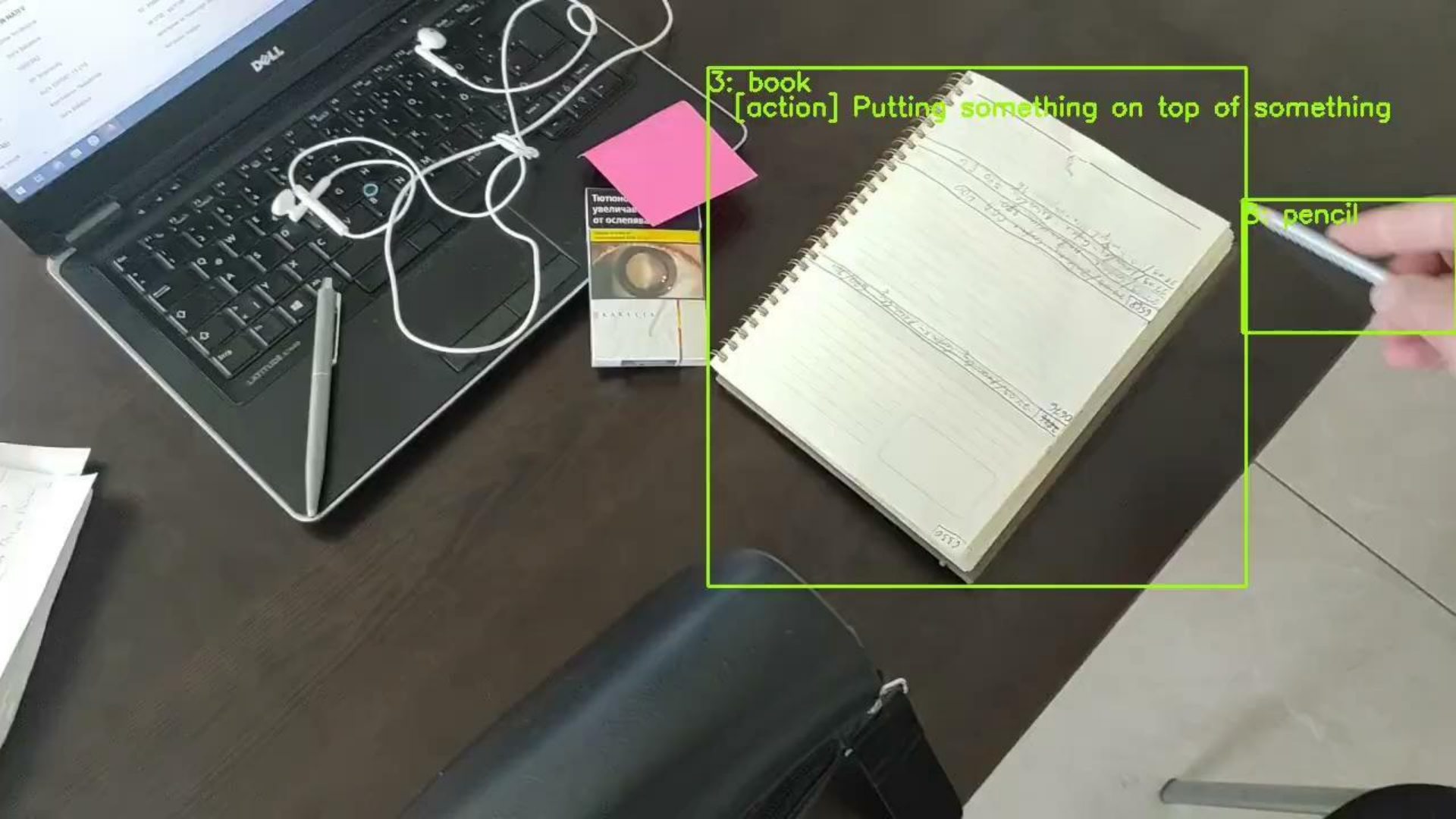
3

Sound segments are tied to the objects that produce the sound

- This connects the sound annotations with other annotations in the dataset.
- These annotations enable tasks such as sound localization in video.


DELL


DELL


3: book
[action] Putting something on top of something
5: pencil

EXTREME
POWER PLUS
460W
Power Supply

Statistics of audio and action segments

- Actions and sound segments are annotated at 30fps.
- Same set of videos (around 11k) have annotations of both actions and sounds.
- There are more annotations of sounds (137k) than actions (74k).

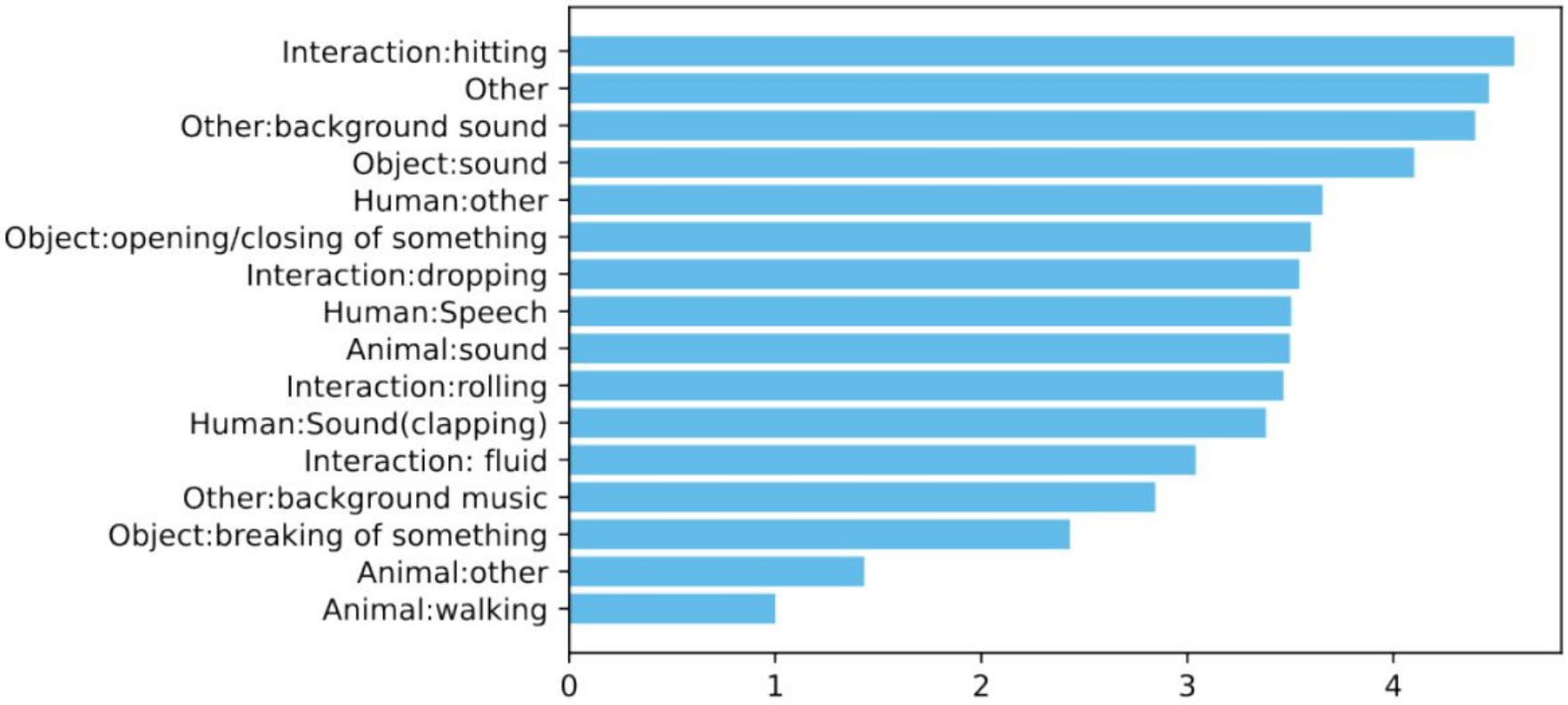
Annotation type	# classes	# annotated instances	# videos	Rate
Objects tracks	5125	191716	11672	1fps
Point tracks	NA	8574	145	30fps
Action segments	63	73859	11416	30fps
Sound segments	16	137756	11484	30fps
mc-vQA	132	38658	11672	NA
g-vQA	35	5890	3085	1fps

Statistics of audio and action segments

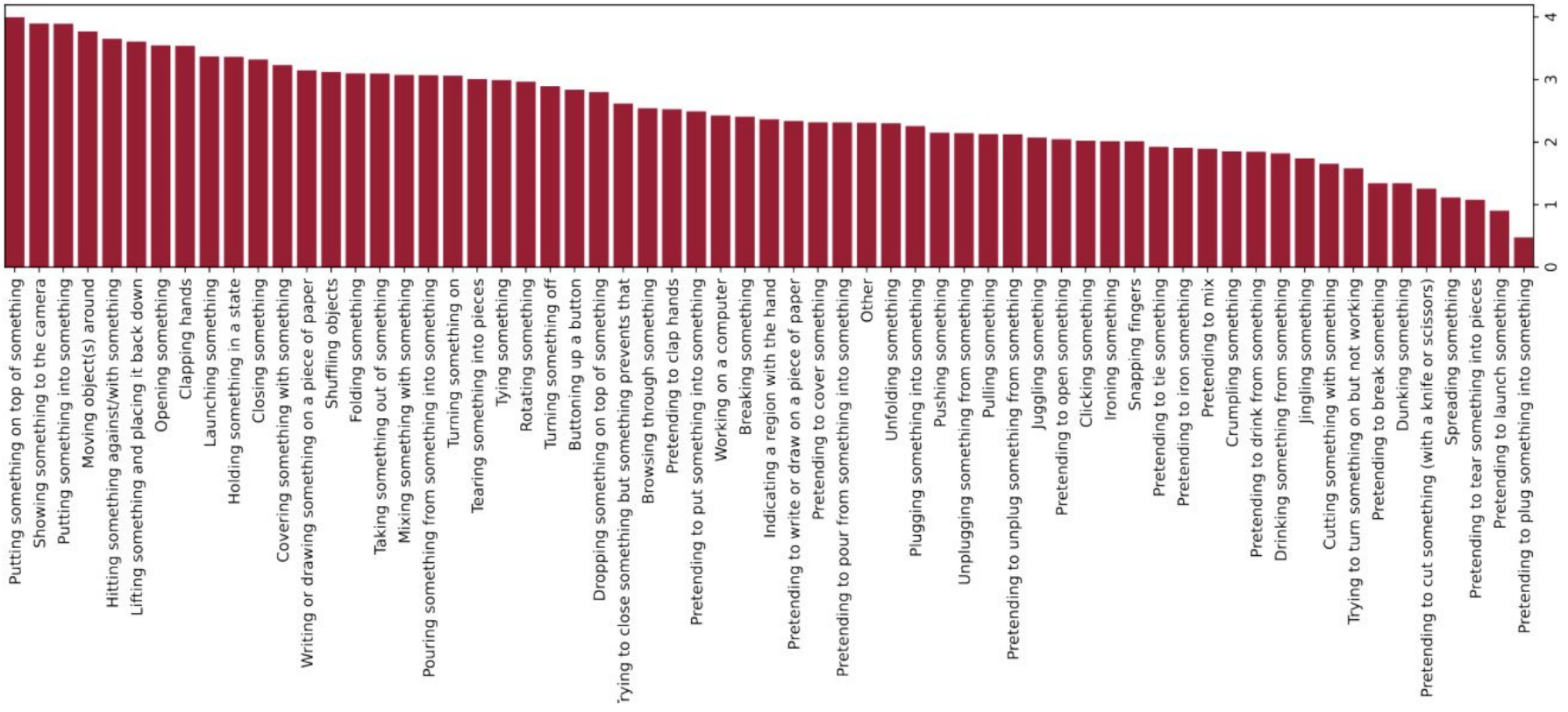
- Using the parent objects we can establish a count of objects tied to sounds and actions in the videos.
- Around 63k objects are involved on actions and sounds out of the total of 191k.
- It is expected that the number of objects that make sound is similar to the number of objects involved on actions, as actions make sound and typically every sound is the result of an action.

	Static or shaking camera	Moving camera	Total
Num total objects	165552	26164	191716
Num action objects	55344	6923	62267
Num sound objects	56158	7666	63824
Num g-vQA boxes	6795	2579	9374

Sound classes in the Perception Test

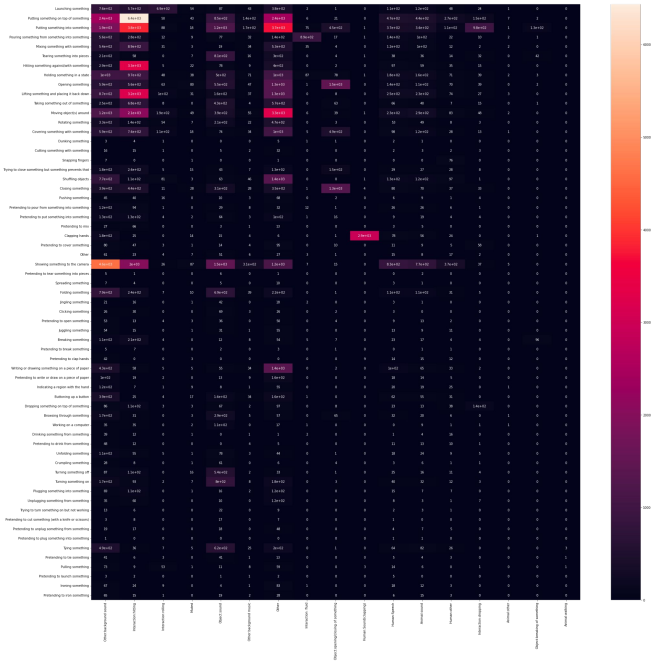


Action classes in the Perception Test



Correlation between actions and sounds

Using the parent objects we can establish a count of objects tied to sounds and actions in the videos.



Action: Pouring something on something
Sound: Interaction: Fluid

Action: Clapping Hands
Sounds: Human (clapping)

Action: Lifting something and putting back down
Sound: Object: Hitting

Action: Moving something around
Sound: Object: Rolling


Camomila


Lipton
YEŞİL ÇAY
Lipton


8: person
[sound] Muted
SAPPORO


LinTea
Chá de
Camomila

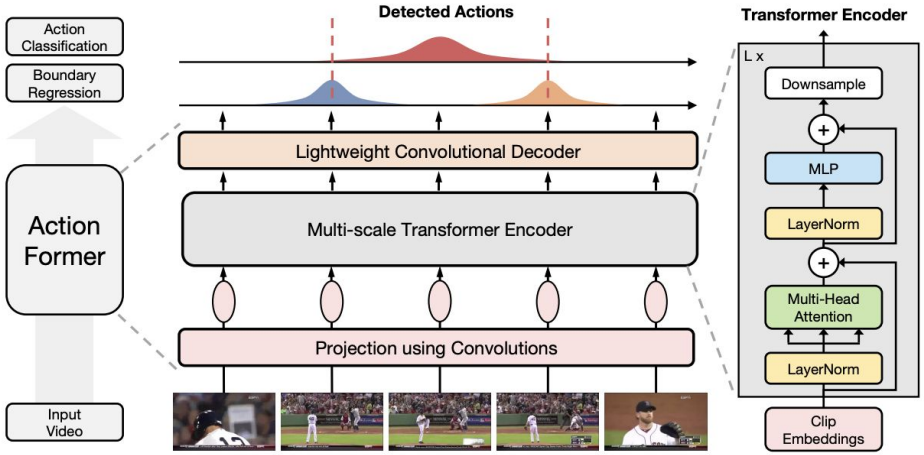
Considerations about sound and actions annotations

Some learning from the annotation process:

- The most important thing is **to be consistent**.
- Distinguishing one vs many (one clap vs clapping) is one challenge.
- It is useful if classes are hierarchical, easier for the annotators.
- Some sounds might be hard to parse / background, it’s better to have everything annotated.
- Storing frame number and times.

Audio and action detection

- Our annotations enable many interesting tasks.
- For now, we focus on temporal detection and classification.
- We are considering ActionFormer (adapted for audio/actions) as our initial baseline.



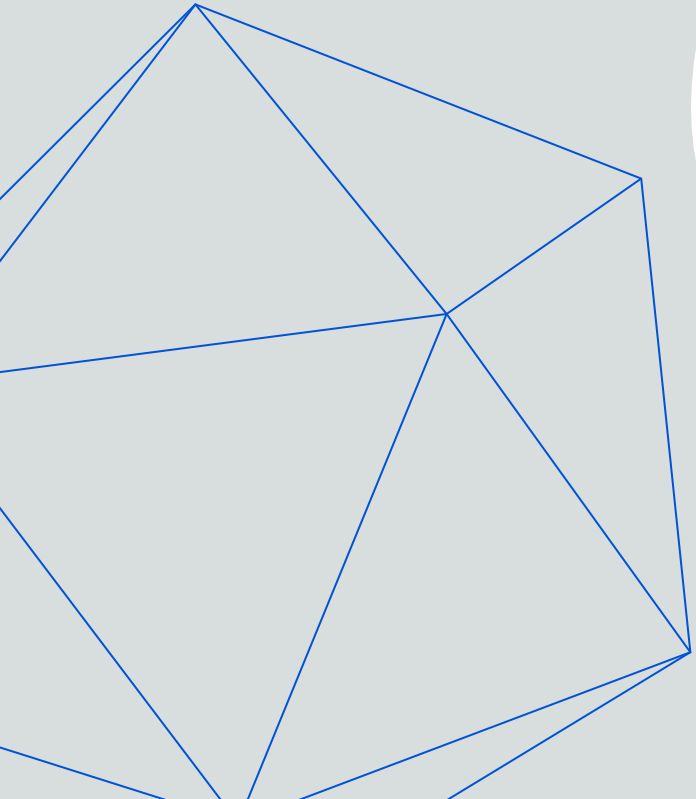
ActionFormer: Localizing Moments of Actions with Transformers, Zhang et al., ECCV 2022

Conclusions

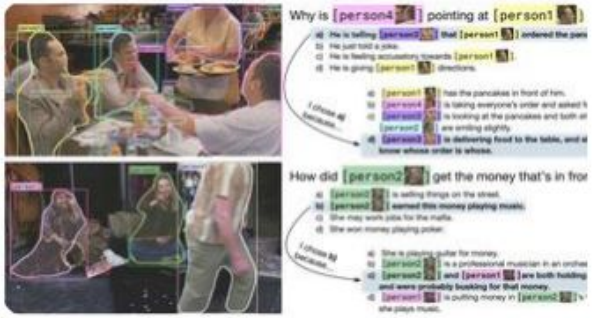
- Actions are the basic pieces of video understanding.
- Sounds are necessary for full parsing of videos.
- We provide a very comprehensive set of annotations for both actions and sounds.
- We hope the Perception Test will push forward research on audio-visual learning.

DeepMind

Multiple-choice videoQA



Many existing VQA datasets

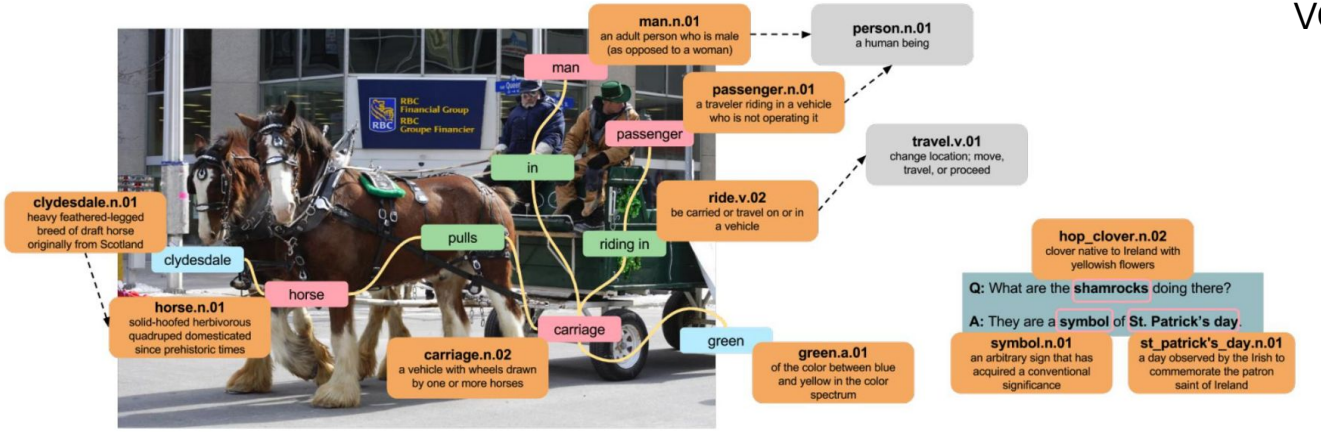


Visual Common Sense Reasoning

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

VQAv2



Visual Genome

Video question-answering

Questions designed by the research team

- Diverse questions
- Coverage of areas and types of reasoning

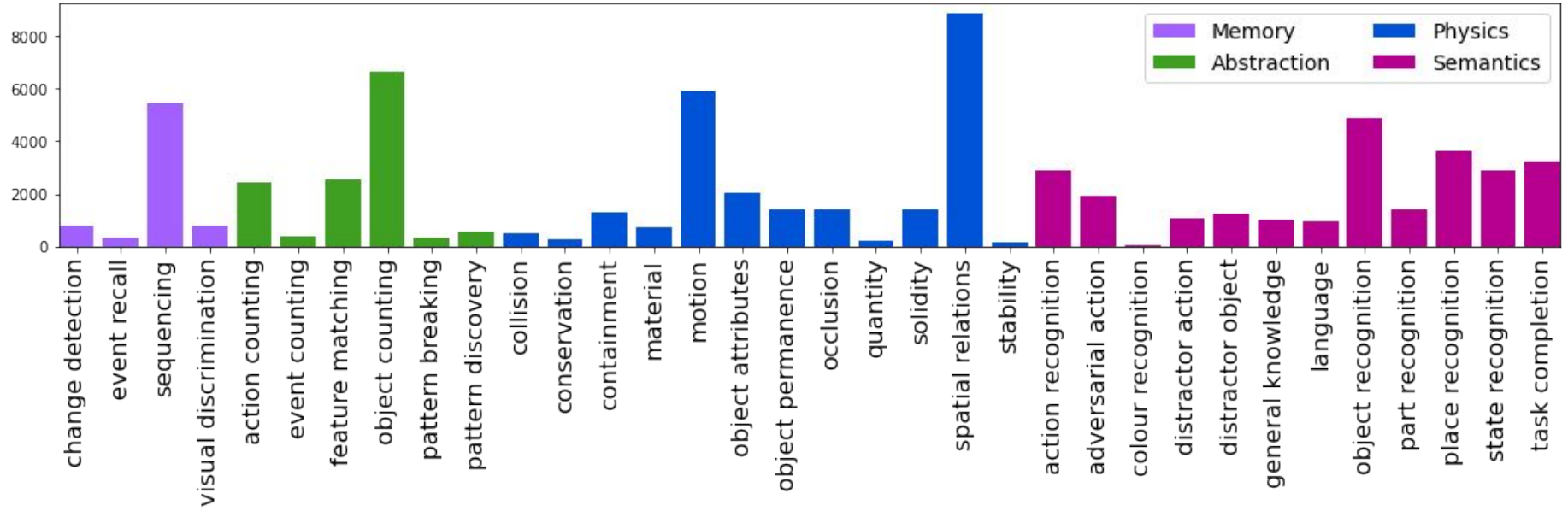
Answers provided by crowd-sourced raters per video

Unambiguous evaluation

- Multiple-choice question-answers with hard distractors
  - Collected with human raters or
  - Correct answers from other videos in the same type of script
- 3 options: 1 correct + 2 distractors
- Metric: top-1 accuracy; 33.3% random baseline

Multiple-choice video QA

4 areas: Memory, Abstraction, Physics (and Geometry), Semantics



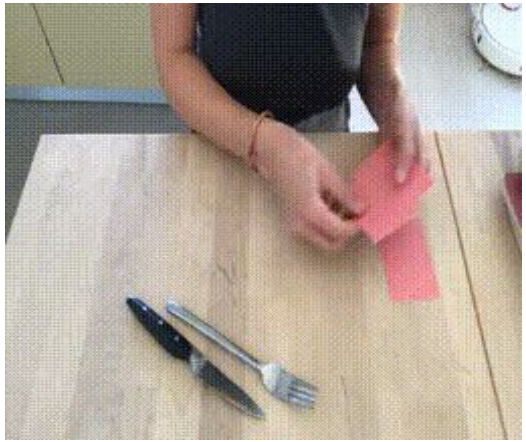
Area	# videoQA pairs	# unique Qs
Memory	7314	36
Abstraction	12853	58
Physics	24158	80
Semantics	25297	82

Reasoning	# videoQA pairs	# unique Qs
Descriptive	32068	106
Explanatory	4558	14
Predictive	1294	7
Counterfactual	738	5

Multiple-choice video QA

Annotation type	# classes	# annotated instances	# videos	Rate
Objects tracks	5125	191716	11672	1fps
Point tracks	NA	8574	145	30fps
Action segments	63	73859	11416	30fps
Sound segments	16	137756	11484	30fps
mc-vQA	132	38658	11672	NA
g-vQA	35	5890	3085	1fps

Multiple-choice video QA



From the objects that the person interacts with, which interaction has a different outcome?

Options:

a) The 1st from the left from the person's perspective
b) The 3rd from the left from the person's perspective
c) <u>The 5th from the left from the person's perspective</u>

How will the person place the last object?

Options:

a) Facing down
b) <u>Facing up</u>
c) I don't know

What action(s) not related to making tea did the person do?

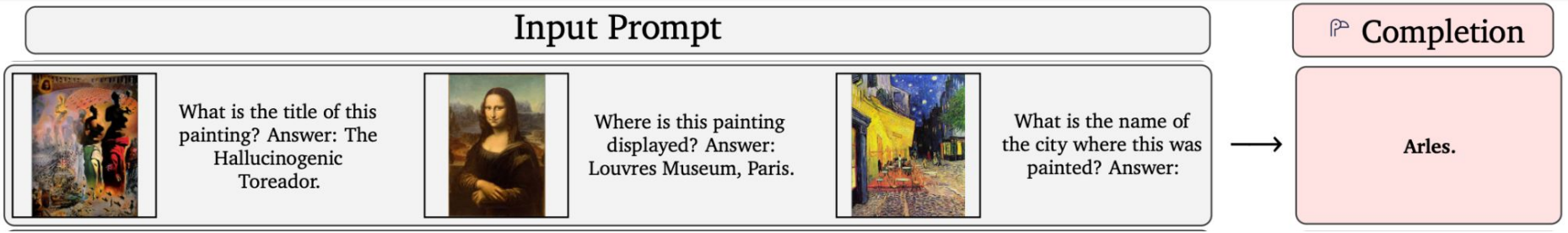
Options:

a) <u>Clapping hands</u>
b) Putting a tomato into a kettle
c) Shuffling letters

The Flamingo baseline for mc-vQA

What is Flamingo?

Flamingo [1] is a visual-language model (which can also ingest videos) adapter for few-shot learning.



[1] **Flamingo: a Visual Language Model for Few-Shot Learning,** J.-B. Alayrac, J. Donahue, P. Luc, A. Miech et al., to appear at NeurIPS'22

The Flamingo baseline for mc-vQA

What is Flamingo?

Flamingo [1] is a visual-language model (which can also ingest videos) adapter for few-shot learning.

Implementation details:

1. Audio was discarded.
2. Flamingo can only work with short videos. Limited the duration of 30 seconds by sampling the centered clip.
3. Framerate: 1 fps
4. Frame resolution: 320x320
5. 4-shots and 8-shots results.
6. No fine-tuning.

[1] **Flamingo: a Visual Language Model for Few-Shot Learning,** J.-B. Alayrac, J. Donahue, P. Luc, A. Miech et al., to appear at NeurIPS'22

The Flamingo baseline for mc-vQA

[1] **Flamingo: a Visual Language Model for Few-Shot Learning,** J.-B. Alayrac, J. Donahue, P. Luc, A. Miech et al., to appear at NeurIPS'22

Flamingo size	Number of shot	Val acc	Test acc
3B	4	46.2	46.2
	8	50.6	50.3
9B	4	47.3	47.4
	8	**52.0**	**52.0**
80B	4	42.4	42.5
	8	47.5	47.9

Baselines for mc-vQA

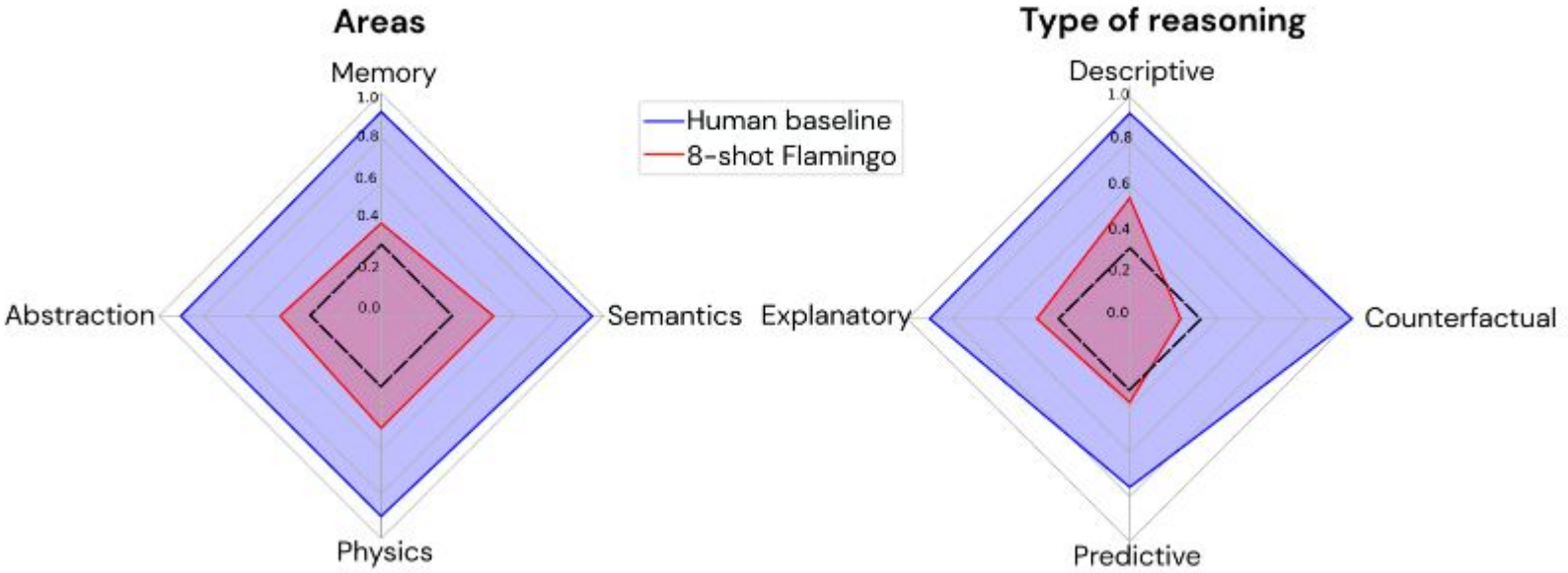
Multiple-choice video QA	0-shot	4-shot	8-shot	Full train set
Flamingo-3B	36.1	46.2	50.6	-
Flamingo-9B	34.7	**47.3**	**52.0**	-
Flamingo-80B	34.1	42.4	47.5	-
Frequency baseline	33.3	40.1	41.5	**47.0**
Human baseline	**91.4**	-	-	-

Human baseline:

- 30 participants (half male, half female, advanced English skills)
- 42 video-question-options, median time 30 minutes
- No training needed for humans: zero-shot

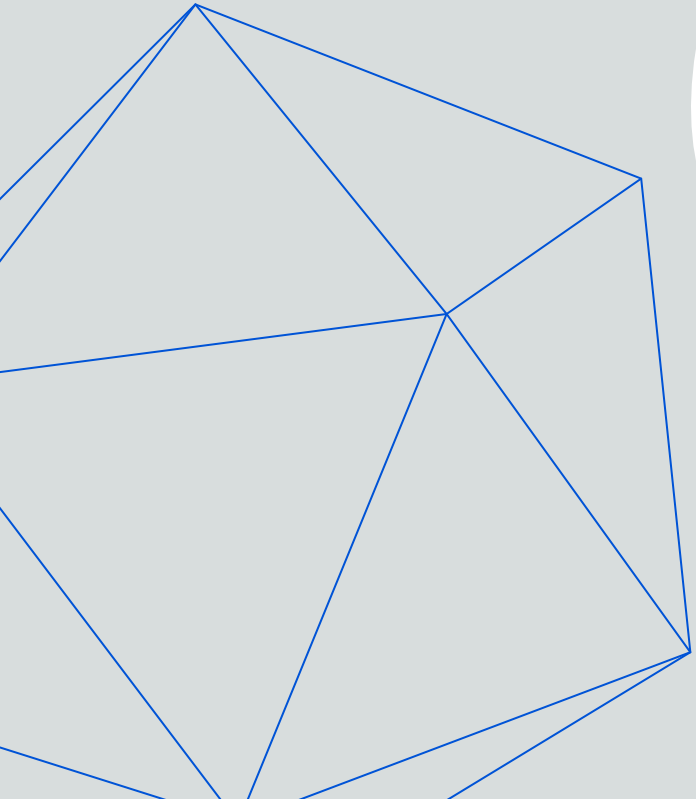
Baselines

Zero-shot human vs 8-shot Flamingo vs random baseline

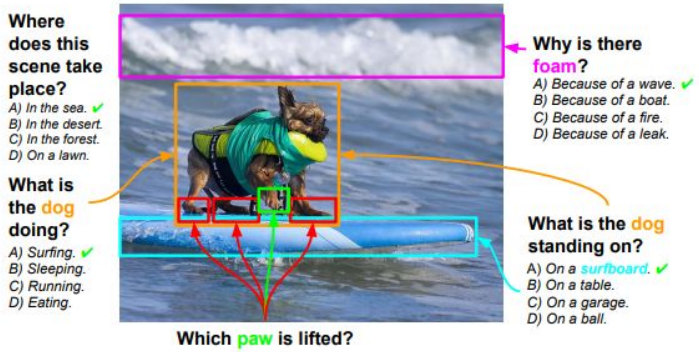


DeepMind

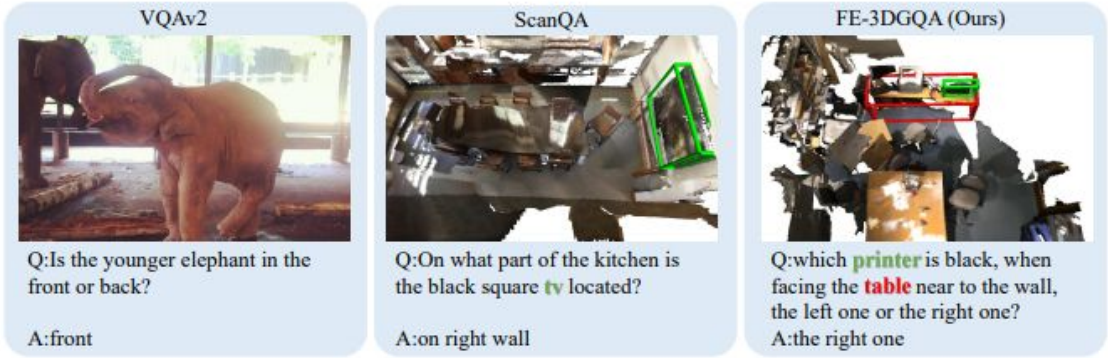
Video QA with grounding



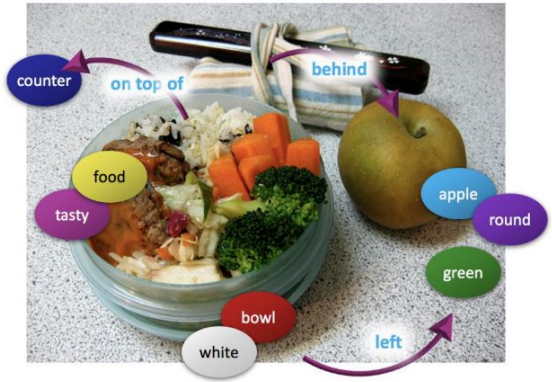
Grounded video QA



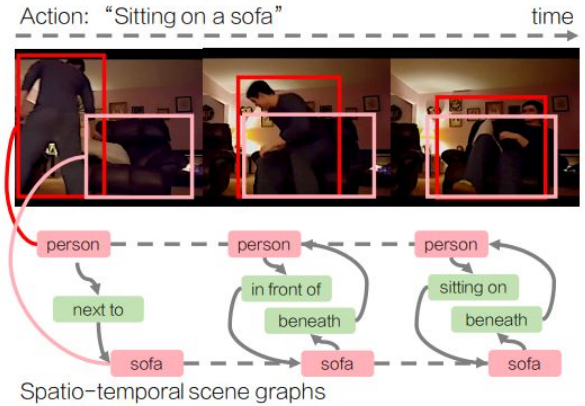
Visual7W



FE-3DGQA



GQA



Action Genome

Task definition and metric

Input:

- Video + Audio + Text question

Output:

- List of object tracks that represent the answer to the question/task

Metric:

- HOTA (higher-order metric for evaluating multi-object tracking)

Annotation type	# classes	# annotated instances	# videos	Rate
Objects tracks	5125	191716	11672	1fps
Point tracks	NA	8574	145	30fps
Action segments	63	73859	11416	30fps
Sound segments	16	137756	11484	30fps
mc-vQA	132	38658	11672	NA
g-vQA	35	5890	3085	1fps

Examples

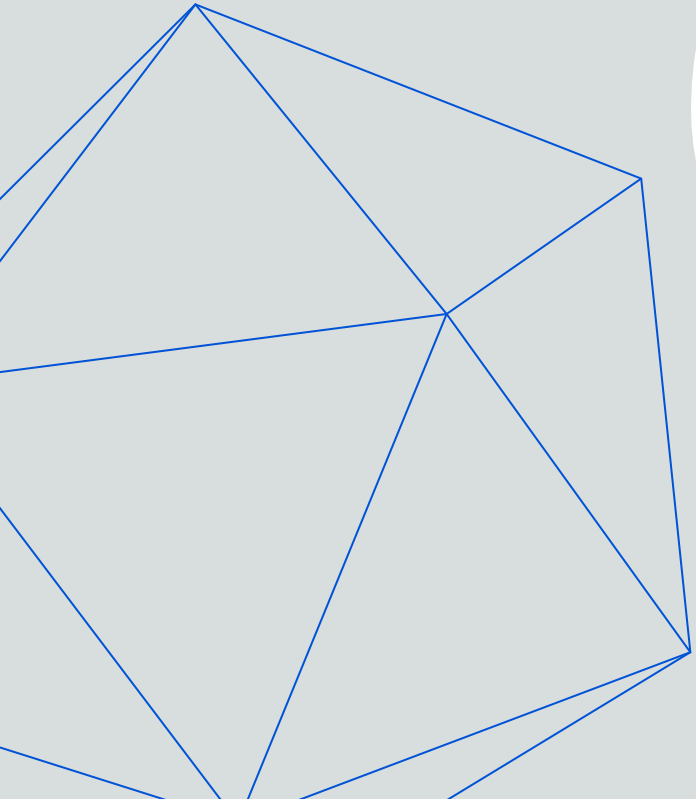
Track the hidden object

Track the launched object

Track the objects added to the table while the camera was looking away.

DeepMind

Cleaning the videos and annotations



Cleaning the videos and annotations

- Filmings delivered in batches, continuous feedback from the team to participants
- Removed videos that had full faces or incorrect execution of scripts
- Some speech still remains, e.g. participants narrating the script

- Clean object tracks if action/sound parents are missing
- Check actions against script type (e.g. for pretend actions)
- Iterate multiple times over difficult cases (e.g. strong occlusions in object tracking, point tracking on non-textured surfaces)

DeepMind

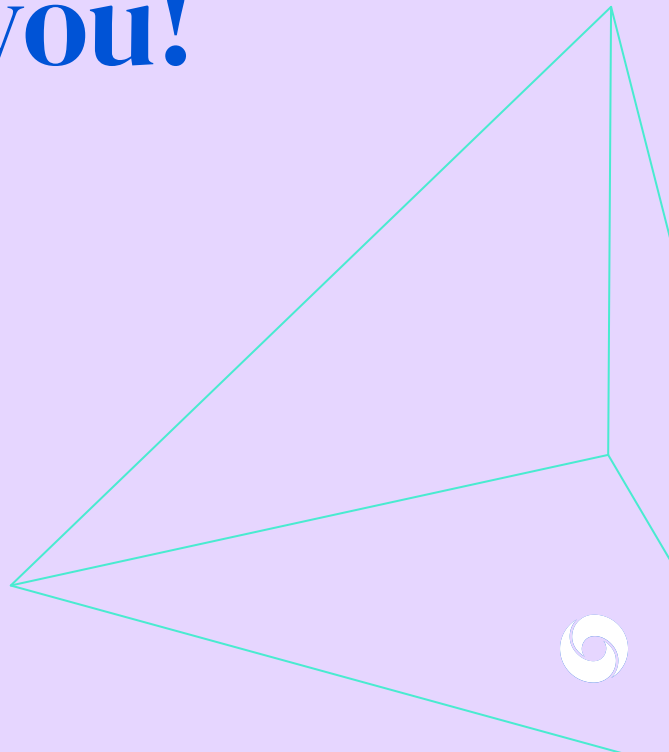
Next steps

Next steps

- Set up a challenge server (Kaggle) and release the test videos
- Populate github with baselines and evaluation code/metrics
  https://github.com/deepmind/perception_test
- Collaborations are very welcome perception-test@google.com

DeepMind

Thank you!



DeepMind

Agenda

12:00 - 12:15	**Opening notes**
12:15 - 13:10	**Perception Test (part 1)**
13:10 - 14:00	**Lunch break**
14:00 - 14:35	**Keynote Olga Russakovsky**
14:35 - 15:30	**Perception Test (part 2)**
15:30 - 16:00	**Coffee break**
16:00 - 16:35	**Keynote Aude Oliva**

16:35 - 17:10	**Keynote Matt Botvinick**
17:10 - 17:45	**Keynote Michael Auli**
17:45 - 18:00	**Coffee break**
18:00 - 18:35	**Keynote Daniel Yamins**
18:35 - 19:10	**Keynote Jitendra Malik**
19:10 - 19:55	**Panel discussion**
19:55 - 20:00	**Closing notes**

DeepMind

Agenda

12:00 - 12:15	**Opening notes**
12:15 - 13:10	**Perception Test (part 1)**
13:10 - 14:00	**Lunch break**
14:00 - 14:35	**Keynote Olga Russakovsky**
14:35 - 15:30	**Perception Test (part 2)**
15:30 - 16:00	**Coffee break**
16:00 - 16:35	**Keynote Aude Oliva**
16:35 - 17:10	**Keynote Matt Botvinick**
17:10 - 17:45	**Keynote Michael Auli**
17:45 - 18:00	**Coffee break**
18:00 - 18:35	**Keynote Daniel Yamins**
18:35 - 19:10	**Keynote Jitendra Malik**
19:10 - 19:55	**Panel discussion**
19:55 - 20:00	**Closing notes**

Panel discussion

Questions:

- How long before we reach human-level perception ?
- What is the one thing that would convince you that we are getting there?
- Is the computer vision community (i.e. people focusing mostly on images and video) best positioned to lead on the next steps of multi-modal (vision, audio, language) perception?
- Biggest challenges and how they evolved over time ?
- Human inspiration – is it worth it ?
- Passive (dataset) evaluation vs active (robotics); Perception Test uses passive evaluation, what are we missing?
  Active operation important for learning, is it necessary for evaluation?

DeepMind

Closing notes

12:00 - 12:15	**Opening notes**
12:15 - 13:10	**Perception Test (part 1)**
13:10 - 14:00	**Lunch break**
14:00 - 14:35	**Keynote Olga Russakovsky**
14:35 - 15:30	**Perception Test (part 2)**
15:30 - 16:00	**Coffee break**
16:00 - 16:35	**Keynote Aude Oliva**
16:35 - 17:10	**Keynote Matt Botvinick**
17:10 - 17:45	**Keynote Michael Auli**
17:45 - 18:00	**Coffee break**
18:00 - 18:35	**Keynote Daniel Yamins**
18:35 - 19:10	**Keynote Jitendra Malik**
19:10 - 19:55	**Panel discussion**
19:55 - 20:00	**Closing notes**

Perception Test

- 11.6k purposefully designed videos to probe perception
- 6 types of annotations (object & point tracks, action & sound segments, multiple-choice and grounded videoQA)
- Train and validation splits released
- Populate github with baselines and evaluation code/metrics https://github.com/deepmind/perception_test
- Set up a challenge server (Kaggle) and release the test videos
- Collaborations are very welcome perception-test@google.com

  E.g. more tasks, videos, languages