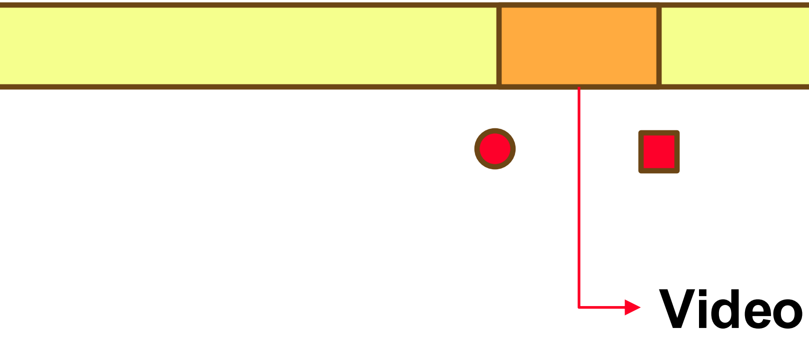
Beyond Long Video Understanding

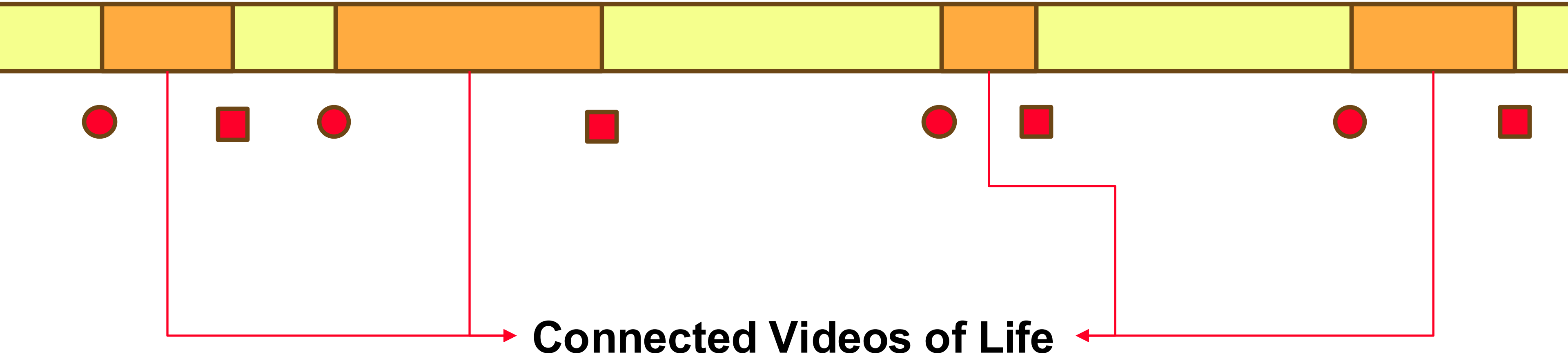
University of BRISTOL

Dima Damen
HVU - CVPRW'25

Previous Video Understanding…



Upcoming Video Understanding…



Eventually…

- No current model has the context required for this ...
- Impossible to store and process this influx of data …

But….

- Immense potential …

In this talk…

Unique Captioning

Visual Instructions

Learning from Continuous Streams

Out of Sight, Not Out of Mind

HD-EPIC

HD-EPIC: A Highly-Detailed Egocentric Video Dataset

In this talk…

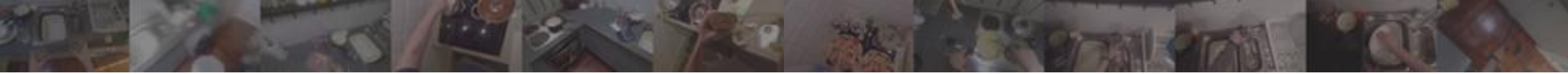
Unique Captioning

Visual Instructions

Learning from Continuous Streams

Out of Sight, Not Out of Mind

HD-EPIC: A Highly-Detailed Egocentric Video Dataset



It’s Just Another Day: Unique Video Captioning by Discriminative Prompting

Toby Perrett, Tengda Han, Dima Damen, Andrew Zisserman

Best Paper Award

ACCV HANOI VIETNAM 2024 DEC 8-12

University of BRISTOL

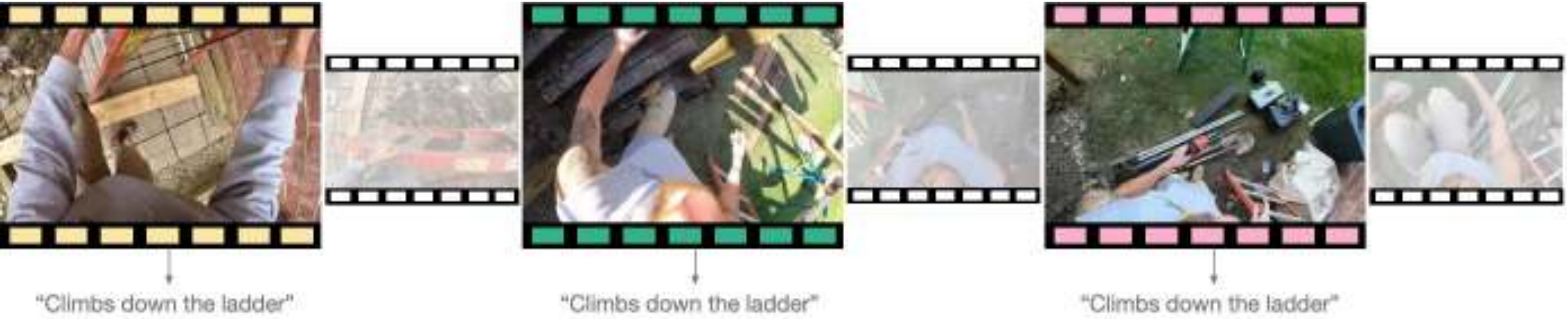
with: Toby Perrett
Tengda Han
Andrew Zisserman

Life is repetitive…

Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Unique Video Captioning

with: Toby Perrett
Tengda Han
Andrew Zisserman



- Current methods caption clips independently
- They generate the same caption for similar clips

Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

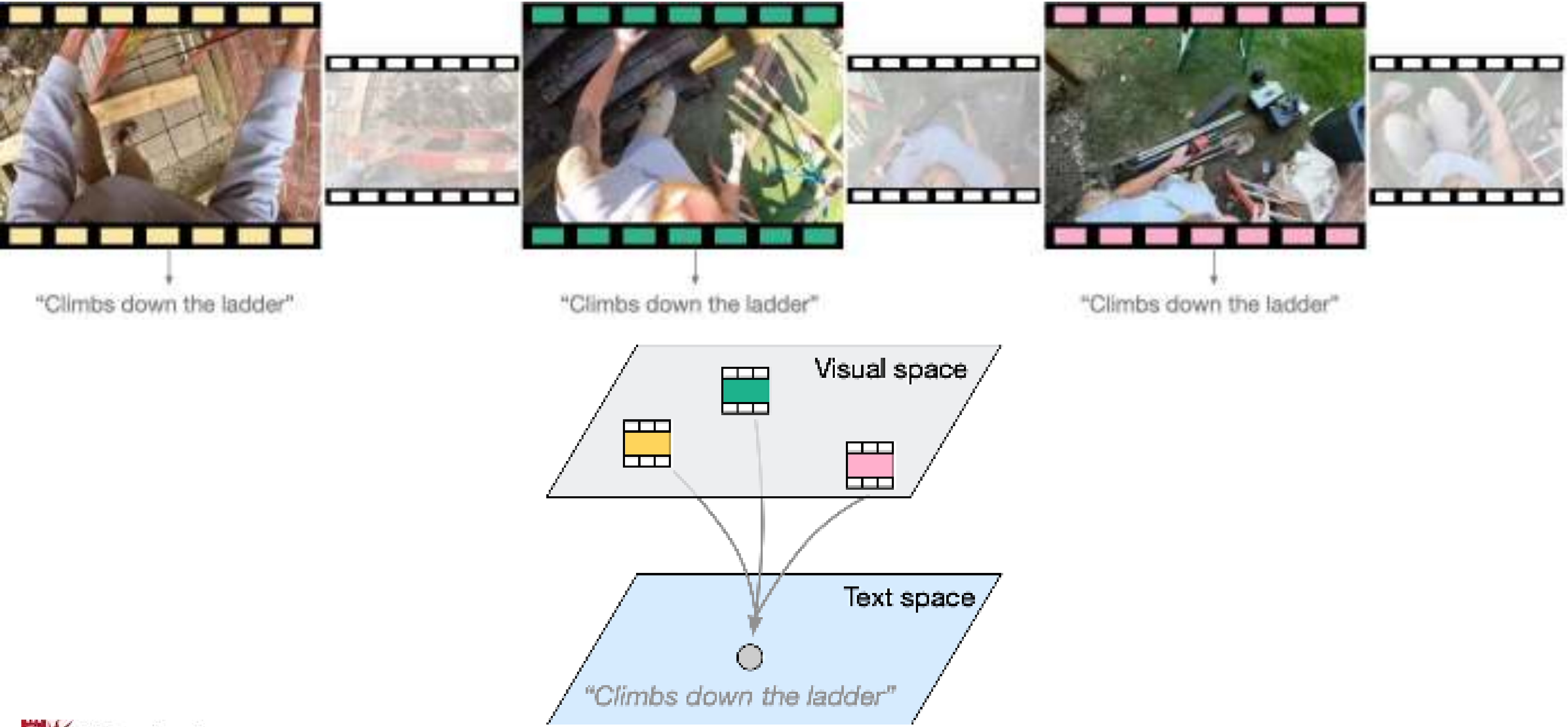
with: Toby Perrett
Tengda Han
Andrew Zisserman

Goal:
Generate a unique caption for every clip in a set

Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Unique Video Captioning

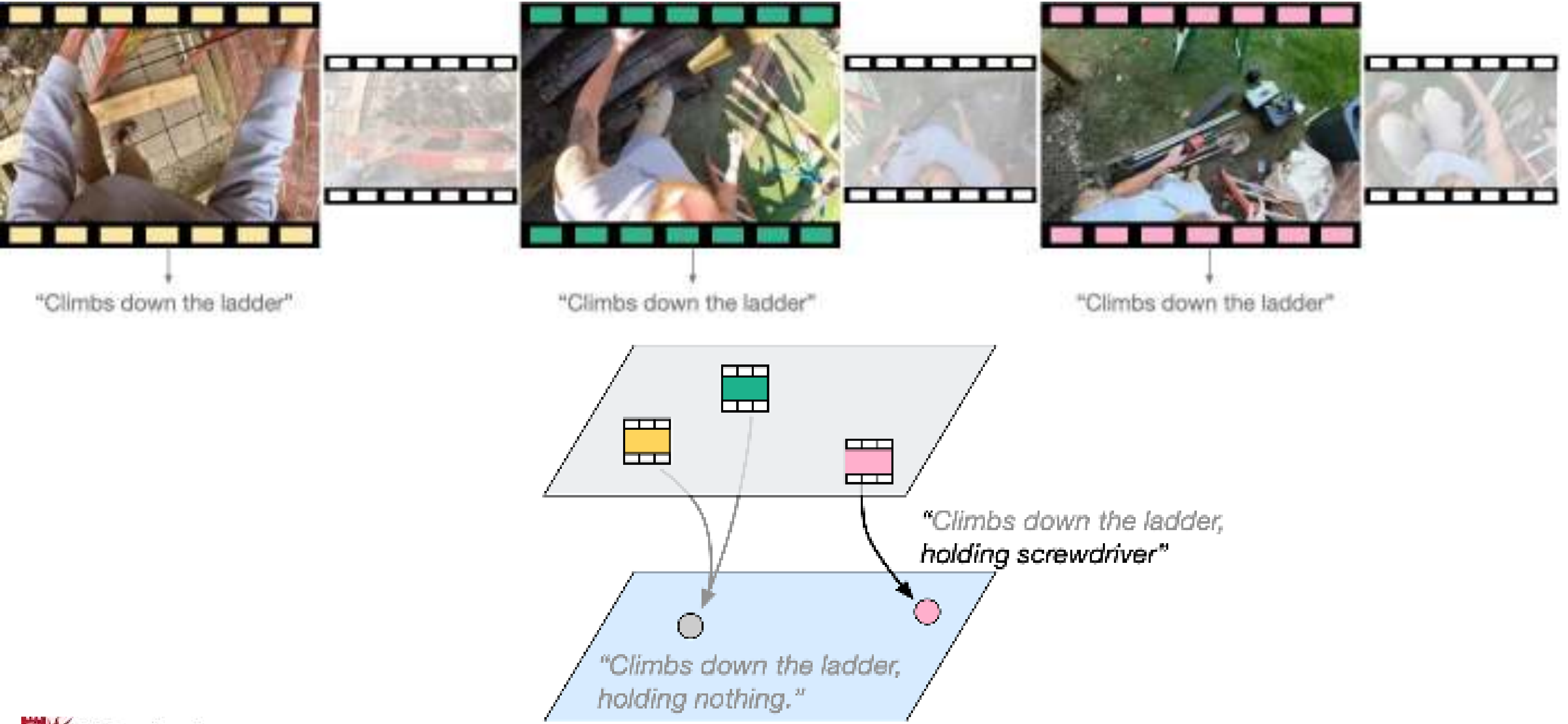
with: Toby Perrett
Tengda Han
Andrew Zisserman



Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Unique Video Captioning

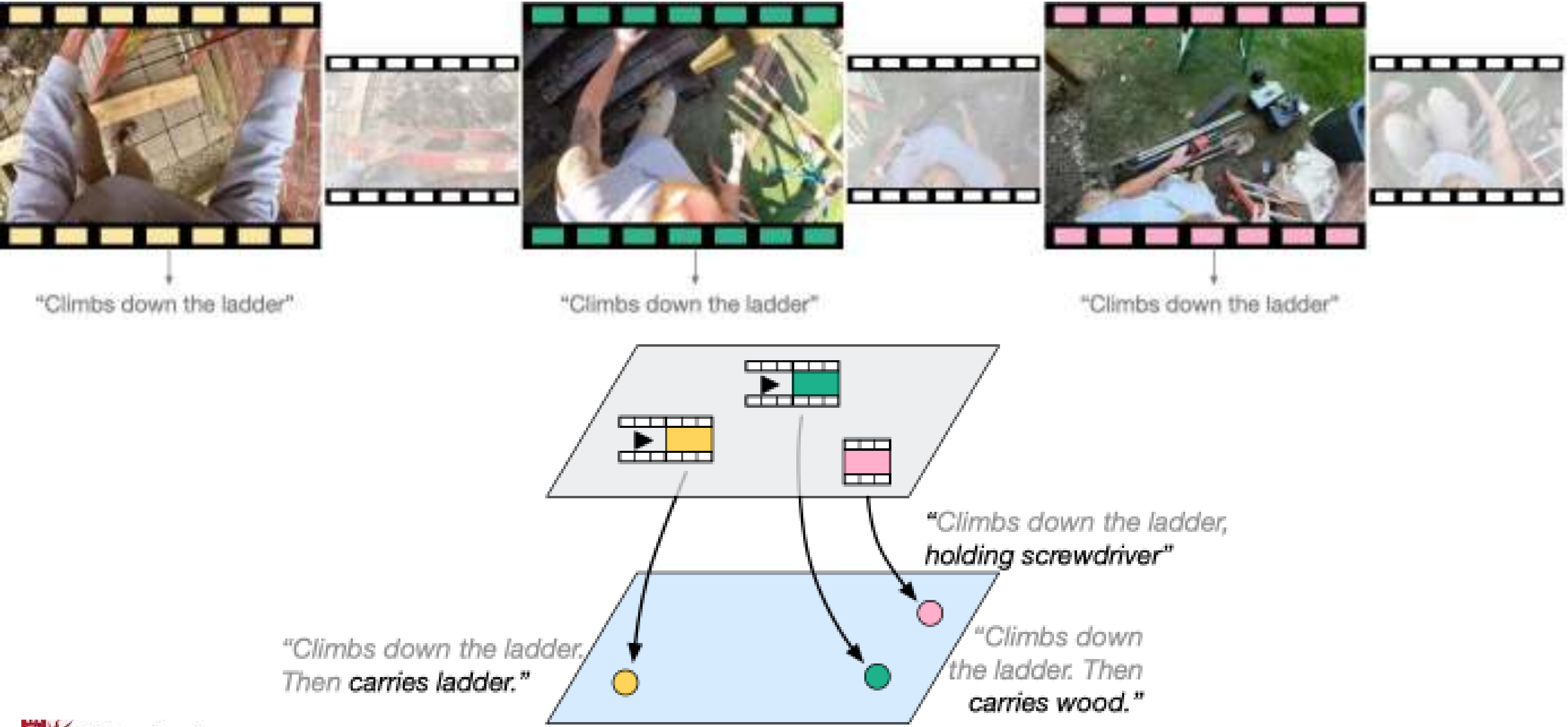
with: Toby Perrett
Tengda Han
Andrew Zisserman



Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Unique Video Captioning

with: Toby Perrett
Tengda Han
Andrew Zisserman

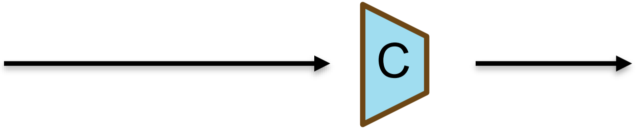


Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Captioning

with: Toby Perrett
Tengda Han
Andrew Zisserman

Captioner *Response*



Climbs down a ladder

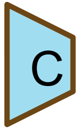
Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Captioning

with: Toby Perrett
Tengda Han
Andrew Zisserman

Discriminative prompt *Captioner* *Response*

The person walks around



Climbs down a ladder and walks around

a building site.

Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Captioning by Discriminative Prompting

with: Toby Perrett
Tengda Han
Andrew Zisserman

Discriminative prompts | *Responses*

The person walks around → C → a building site

The person holds → C → a screwdriver

The person is wearing → C → a t-shirt

…

The person walks around → C → a building site

The person holds → C → the ladder

The person is wearing → C → a jumper

…

Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Captioning by Discriminative Prompting

with: Toby Perrett
Tengda Han
Andrew Zisserman

Discriminative prompts — *Responses*

The person walks around → C → a building site

The person holds → C → a screwdriver

The person is wearing → C → a t-shirt

...

The person walks around → C → a building site

The person holds → C → the ladder

The person is wearing → C → a jumper

Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Captioning by Discriminative Prompting

with: Toby Perrett
Tengda Han
Andrew Zisserman

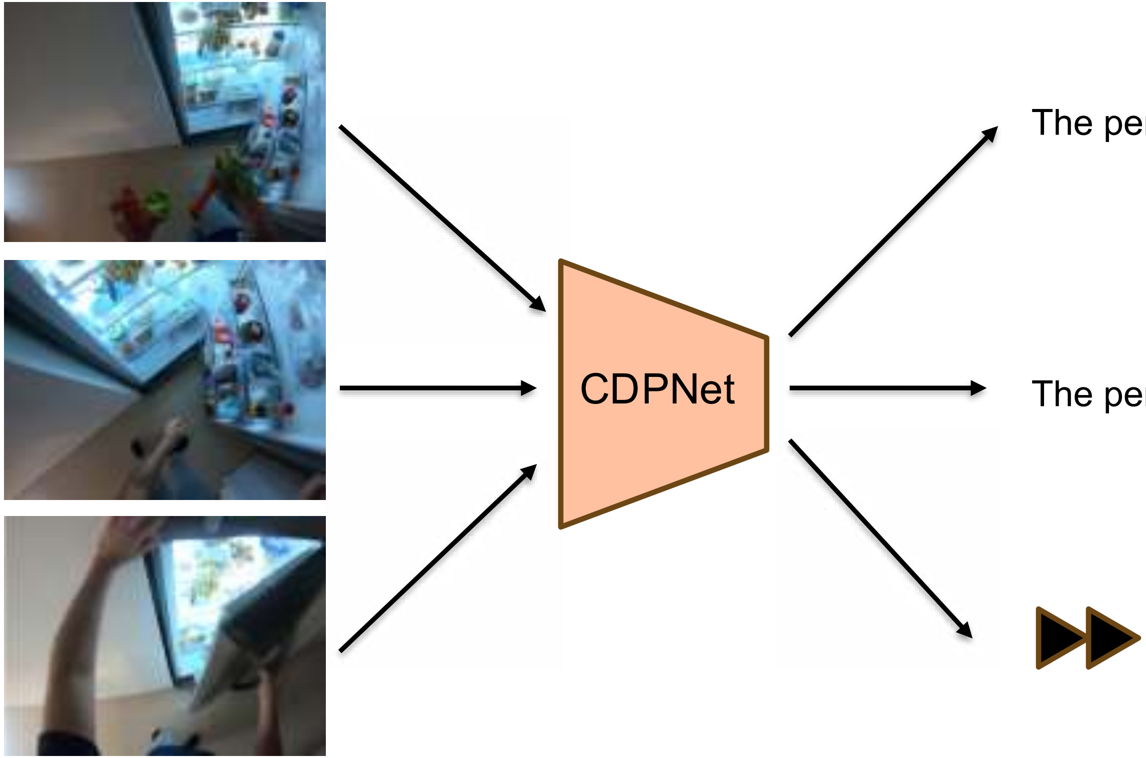
We propose to…
consider clips jointly
use a bank of discriminative prompts

But…
Expensive £££
What if there isn't a suitable prompt?

Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Captioning by Discriminative Prompting

with: Toby Perrett
Tengda Han
Andrew Zisserman



Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Unique Video Captioning

with: Toby Perrett
Tengda Han
Andrew Zisserman

Average recall @ 1

Egocentric	**+0s**
LaViLa	37
LaViLa + CDP	**45**

Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Unique Video Climbing

with: Toby Perrett
Tengda Han
Andrew Zisserman

Climbs the stairs

Unique Video Captioning

with: Toby Perrett
Tengda Han
Andrew Zisserman

Climbs the stairs

Climbs the stairs and

holds the phone

Climbs the stairs and

picks up the drill

Climbs the stairs and

holds a tape measure

Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Unique Video Captioning

with: Toby Perrett
Tengda Han
Andrew Zisserman

Looks around the shelves

Perrett et al (2024). It's Just Another Day: Unique Video Captioning by Discriminative Prompting. Asian Conference on Computer Vision (ACCV)

Unique Video Captioning

with: Toby Perrett
Tengda Han
Andrew Zisserman

Looks around the shelves

Looks around the shelves and

the other man picks up a packet of biscuits from the shelf with his left hand

Looks around the shelves and

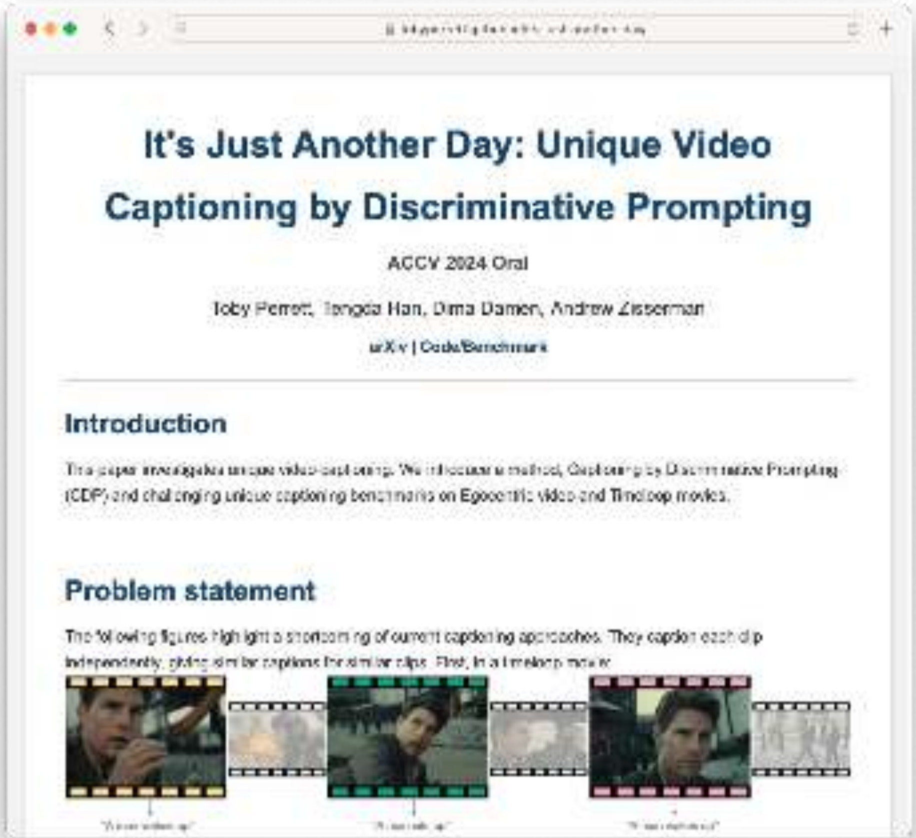
looks at the list

Looks around the shelves and then

picks up a packet of cough rubs

Unique Video Captioning

with: Toby Perrett
Tengda Han
Andrew Zisserman



It's Just Another Day: Unique Video Captioning by Discriminative Prompting

ACCV 2024 Oral

Toby Perrett, Tengda Han, Dima Damen, Andrew Zisserman

arXiv | Code/Benchmark

Introduction

This paper investigates unique video captioning. We introduce a method, Captioning by Discriminative Prompting (CDP) and challenging unique captioning benchmarks on Egocentric video and Timeloop movies.

Problem statement

The following figures highlight a shortcoming of current captioning approaches. They caption each clip independently, giving similar captions for similar clips. First, in a timeloop movie:

Dima Damen

In this talk…

Unique Captioning

Visual Instructions

Learning from Continuous Streams

Out of Sight,
Not Out of Mind

HD-EPIC

HD-EPIC: A Highly-Detailed Egocentric Video Dataset

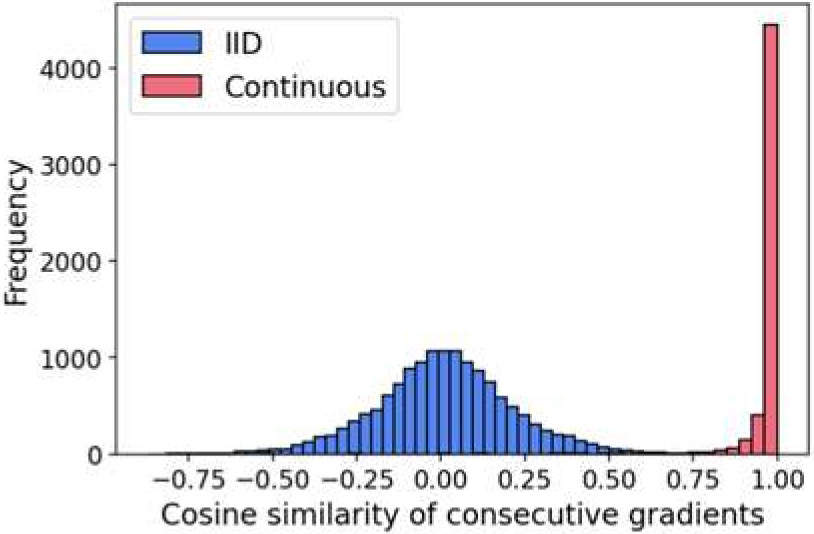
Learning from a continuous video stream

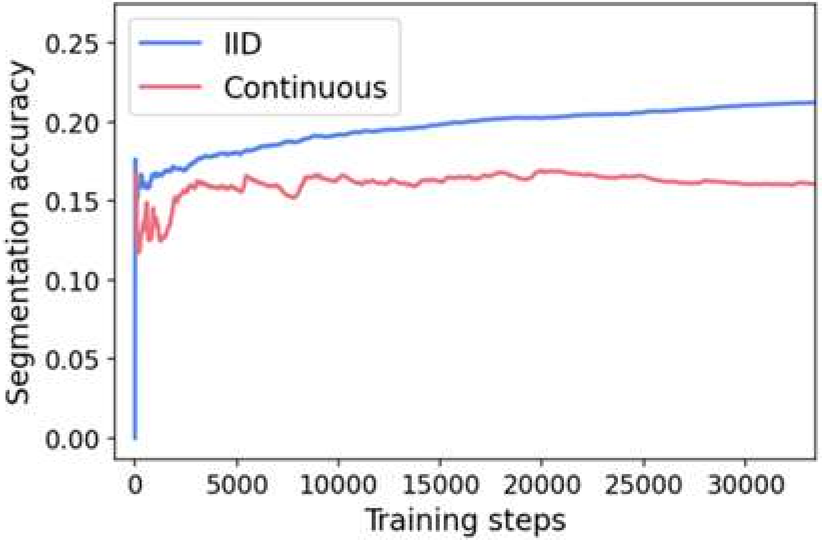
Dima Damen
HVU - CVPRW'25

Current learning paradigms…

Han et al (2024). Learning from One Continuous Video Stream. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

Current learning paradigms…





Han et al (2024). Learning from One Continuous Video Stream. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

This year…

Learning from Streaming Video with Orthogonal Gradients

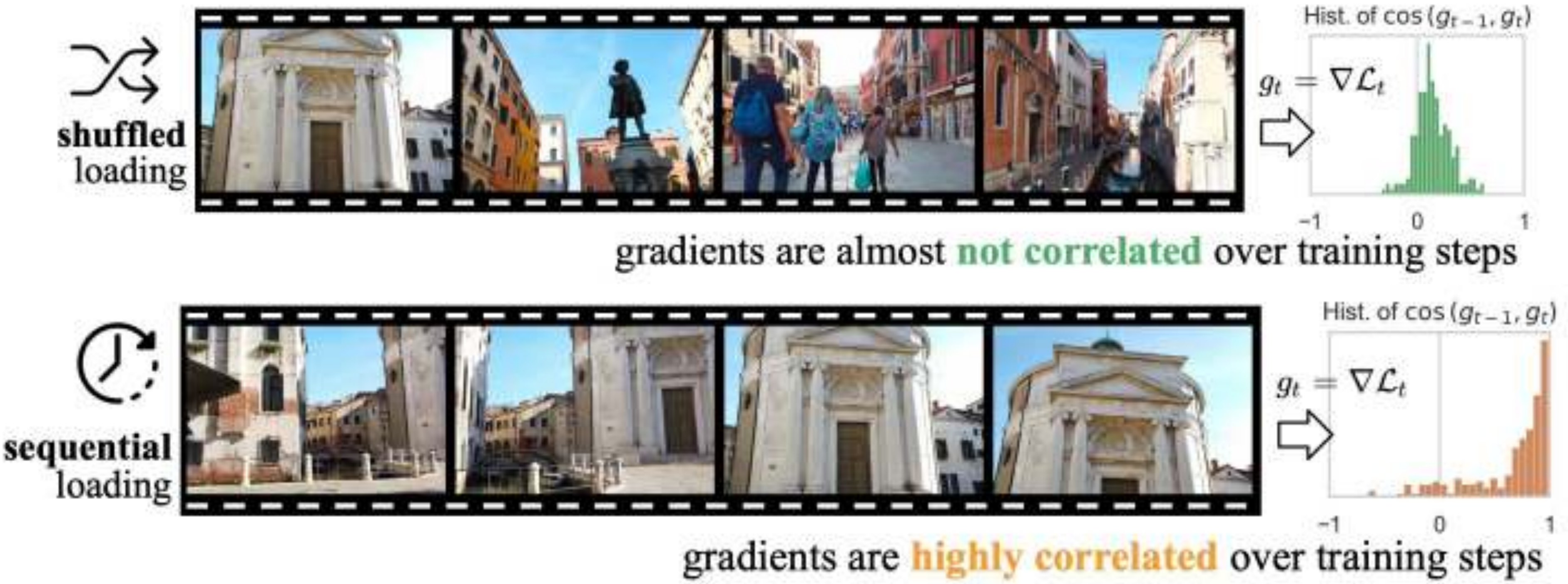
Tengda Han°, Dilara Gokay°, Joseph Heyward°, Chuhan Zhang°
Daniel Zoran°, Viorica Pătrăucean°, João Carreira°, Dima Damen°†, Andrew Zisserman°‡
°Google DeepMind, †University of Bristol, ‡University of Oxford

Han et al (2025). Learning from Streaming Video with Orthogonal Gradients. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

Learning from Streaming Videos with Orthogonal Gradients

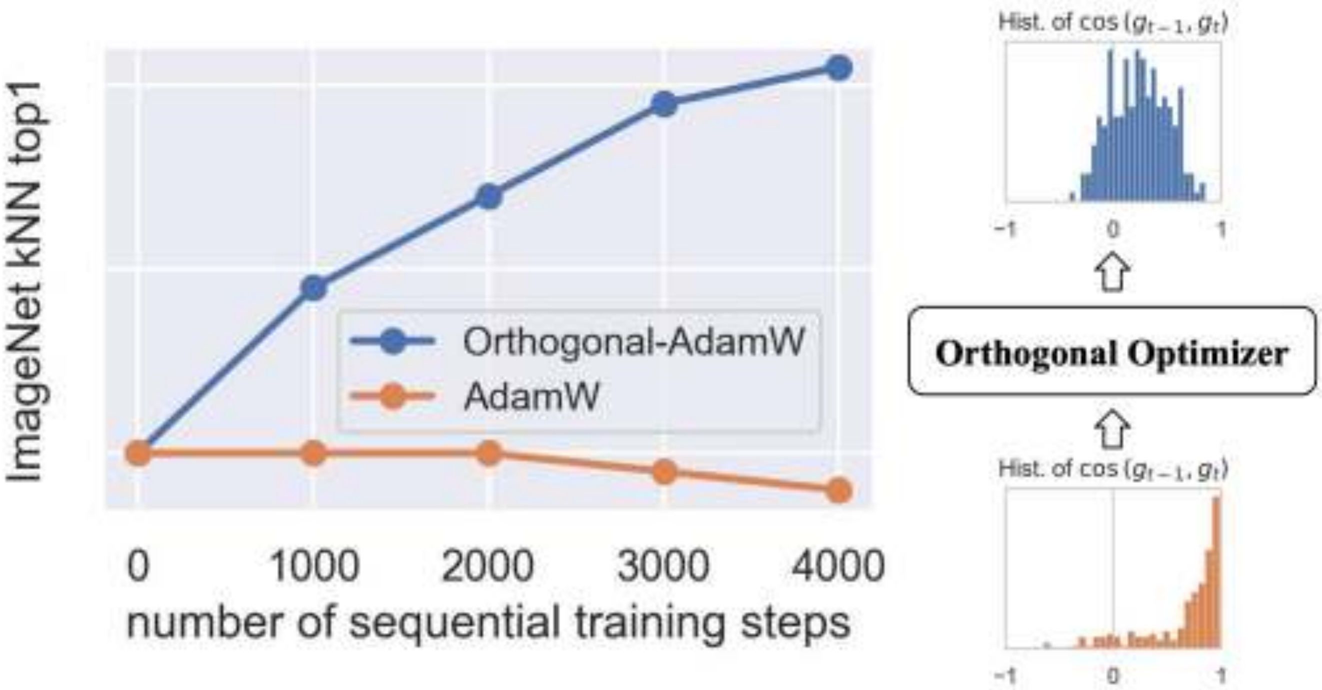
Han et al (2025). Learning from Streaming Video with Orthogonal Gradients. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

Learning from Streaming Videos with Orthogonal Gradients



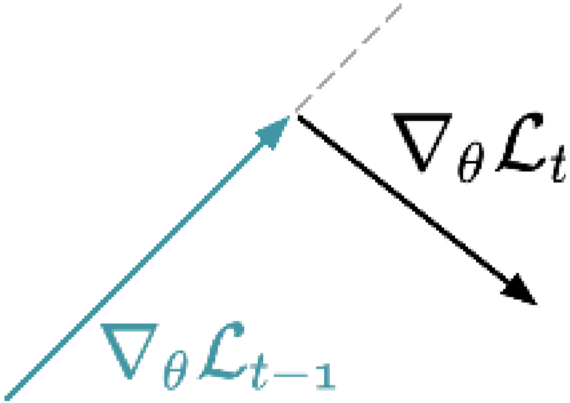
Han et al (2025). Learning from Streaming Video with Orthogonal Gradients. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

Learning from Streaming Videos with Orthogonal Gradients



Han et al (2025). Learning from Streaming Video with Orthogonal Gradients. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

Learning from Streaming Videos with Orthogonal Gradients



(a)

Han et al (2025). Learning from Streaming Video with Orthogonal Gradients. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

Learning from Streaming Videos with Orthogonal Gradients

Algorithm 2 AdamW

Require: Learning rate $\eta > 0$, weight decay coefficient $\lambda > 0$, decay rates $\beta_1, \beta_2 \in [0, 1)$, small constant $\epsilon > 0$, initial parameters θ_0, number of iterations T

1: Initialize first moment vector $m_0 = 0$, , and second moment vector $v_0 = 0$
2: **for** $t = 1$ to T **do**
3: Sample a mini-batch of data $\mathcal{B}_t$ from the training set
4: Compute the gradient: $g_t = \nabla_\theta \mathcal{L}(\theta_{t-1}; \mathcal{B}_t)$
8: Update biased first moment estimate: $m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$
9: Update biased second moment estimate: $v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$
10: Compute bias-corrected first moment: $\hat{m}_t = \frac{m_t}{1-\beta_1^t}$
11: Compute bias-corrected second moment: $\hat{v}_t = \frac{v_t}{1-\beta_2^t}$
12: Apply weight decay: $\theta_{t-1} = \theta_{t-1} - \eta\lambda\theta_{t-1}$
13: Update parameters: $\theta_t = \theta_{t-1} - \eta\frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon}$
14: **end for**

Han et al (2025). Learning from Streaming Video with Orthogonal Gradients. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

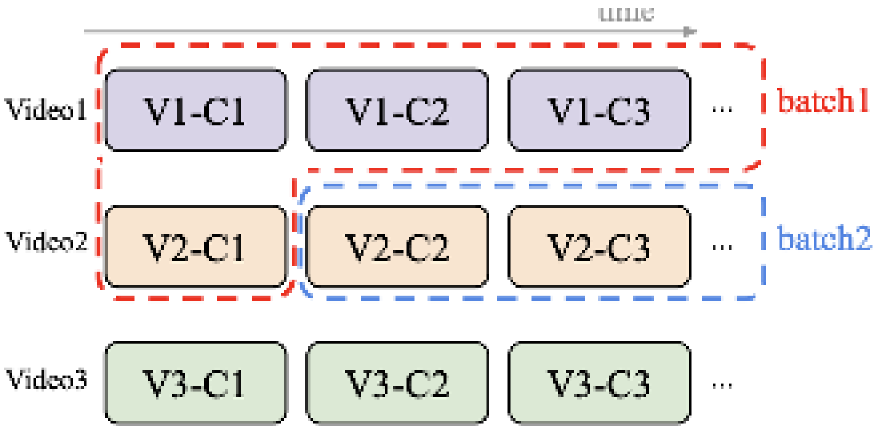
Learning from Streaming Videos with Orthogonal Gradients

initialization	pretraining dataset: WT_{venice}		downstream ImageNet	
	pretraining method	optimizer	linear probe top1	kNN top1
$DINO_{ImageNet}$	-	-	-	74.4
$DINO_{ImageNet}$	DoRA sequential (batch-along-time)	AdamW	6.1	1.8
$DINO_{ImageNet}$	DoRA sequential (batch-along-time)	Orthogonal-AdamW	**64.5**	**51.8**

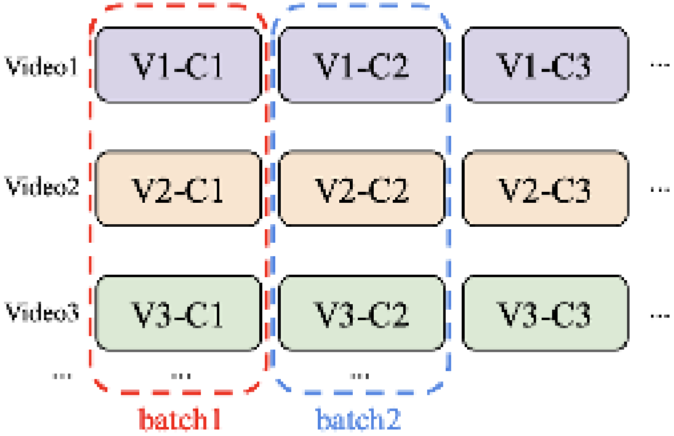
Venkataramanan et al (2024). Is ImageNet worth 1 video? learning strong image encoders from 1 long unlabelled video. ICLR

Han et al (2025). Learning from Streaming Video with Orthogonal Gradients. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

Learning from Streaming Videos with Orthogonal Gradients



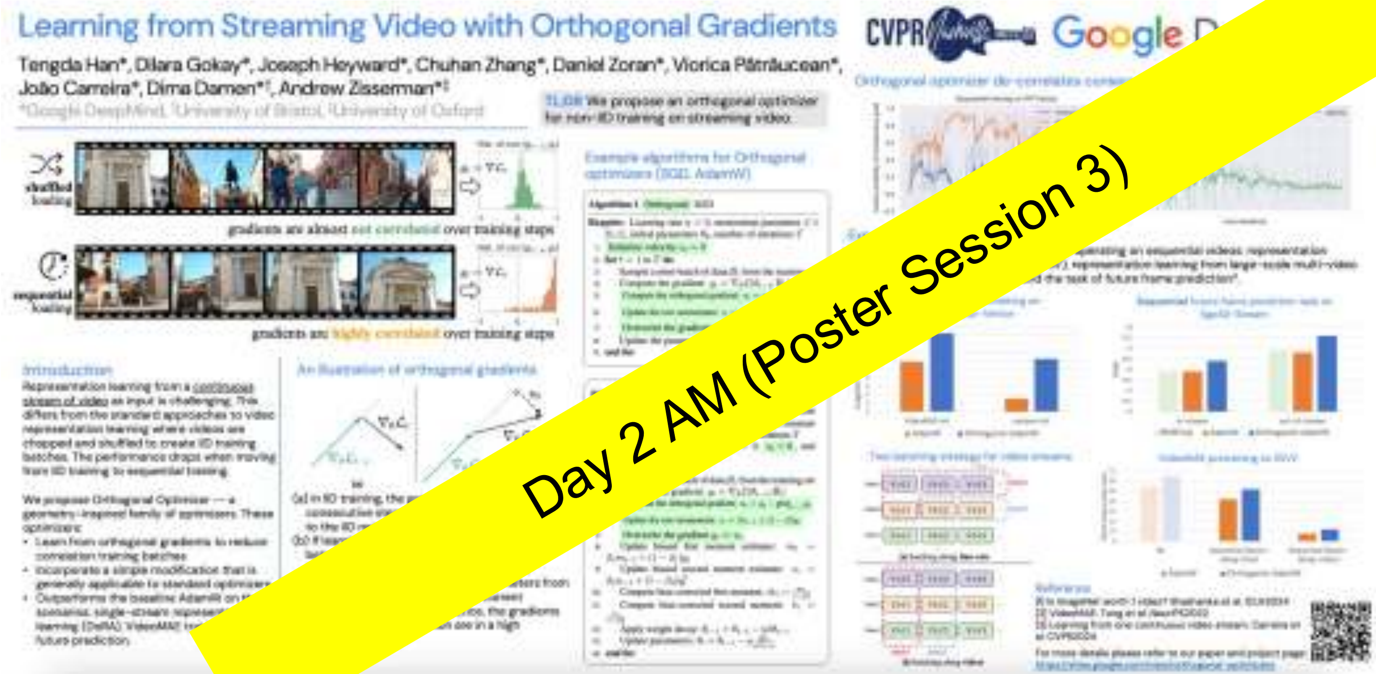
(a) batching along time axis



(b) batching along videos

Han et al (2025). Learning from Streaming Video with Orthogonal Gradients. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

Learning from Streaming Videos with Orthogonal Gradients



Day 2 AM (Poster Session 3)

Han et al (2025). Learning from Streaming Video with Orthogonal Gradients. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

In this talk…

Unique Captioning

Visual Instructions

Learning from
Continuous Streams

Out of Sight,
Not Out of Mind

HD-EPIC

HD-EPIC: A Highly-Detailed Egocentric Video
Dataset

CVPR Nashville JUNE 11-15, 2025

ShowHowTo: Generating Scene-Conditioned Step-by-Step Visual Instructions

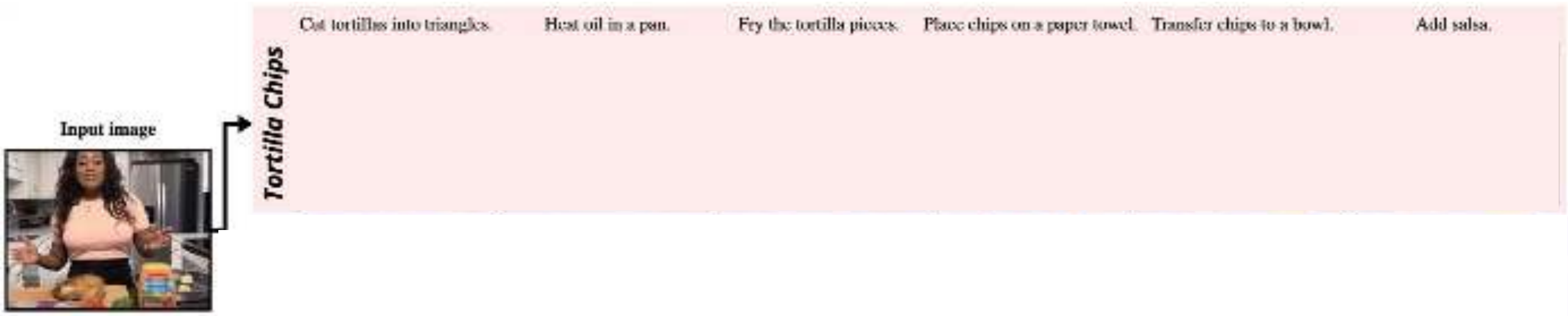
Tomáš Souček[1] Prajwal Gatti[2] Michael Wray[2] Ivan Laptev[3] Dima Damen[2] Josef Sivic[1]

[1]CIIRC CTU [2]University of Bristol [3]MBZUAI

Soucek et al (2025). ShowHowTo: Generating Scene-Conditioned Step-by-Step Visual Instructions. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

ShowHowTo

with: Tomas Soucek Michael Wray
Prajwal Gatti Ivan Laptev
Josef Sivic



ShowHowTo

with: Tomas Soucek Michael Wray
Prajwal Gatti Ivan Laptev
Josef Sivic

Soucek et al (2025). ShowHowTo: Generating Scene-Conditioned Step-by-Step Visual Instructions. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

ShowHowTo

with: Tomas Soucek Michael Wray
Prajwal Gatti Ivan Laptev
Josef Sivic

Soucek et al (2025). ShowHowTo: Generating Scene-Conditioned Step-by-Step Visual Instructions. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

ShowHowTo

with: Tomas Soucek Michael Wray
Prajwal Gatti Ivan Laptev
Josef Sivic

Soucek et al (2025). ShowHowTo: Generating Scene-Conditioned Step-by-Step Visual Instructions. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

ShowHowTo

with: Tomas Soucek Michael Wray
Prajwal Gatti Ivan Laptev
Josef Sivic

Soucek et al (2025). ShowHowTo: Generating Scene-Conditioned Step-by-Step Visual Instructions. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

ShowHowTo

with: Tomas Soucek Michael Wray
Prajwal Gatti Ivan Laptev
Josef Sivic

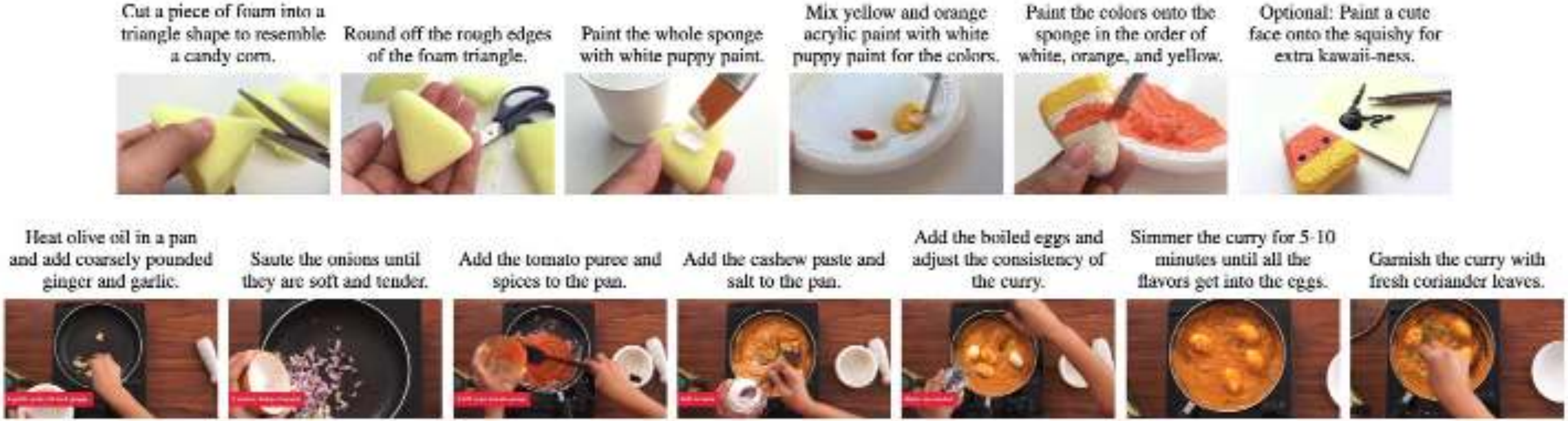
Processing 1M instructional videos in HowTo100M leads to…

- A large-scale dataset (578K sequences, 4.5M steps)
- Task diversity (25K+ HowTo tasks)
- Ability to scale further (no manual annotation required)

Soucek et al (2025). ShowHowTo: Generating Scene-Conditioned Step-by-Step Visual Instructions. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

ShowHowTo

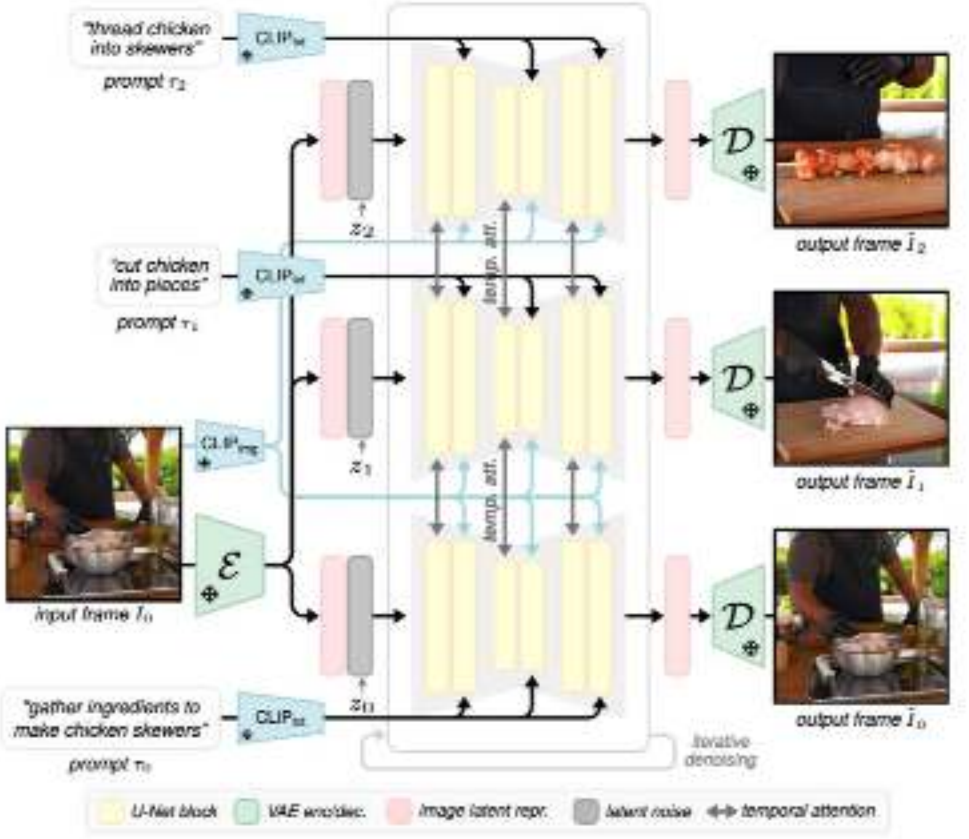
with: Tomas Soucek Michael Wray
Prajwal Gatti Ivan Laptev
Josef Sivic



Soucek et al (2025). ShowHowTo: Generating Scene-Conditioned Step-by-Step Visual Instructions. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

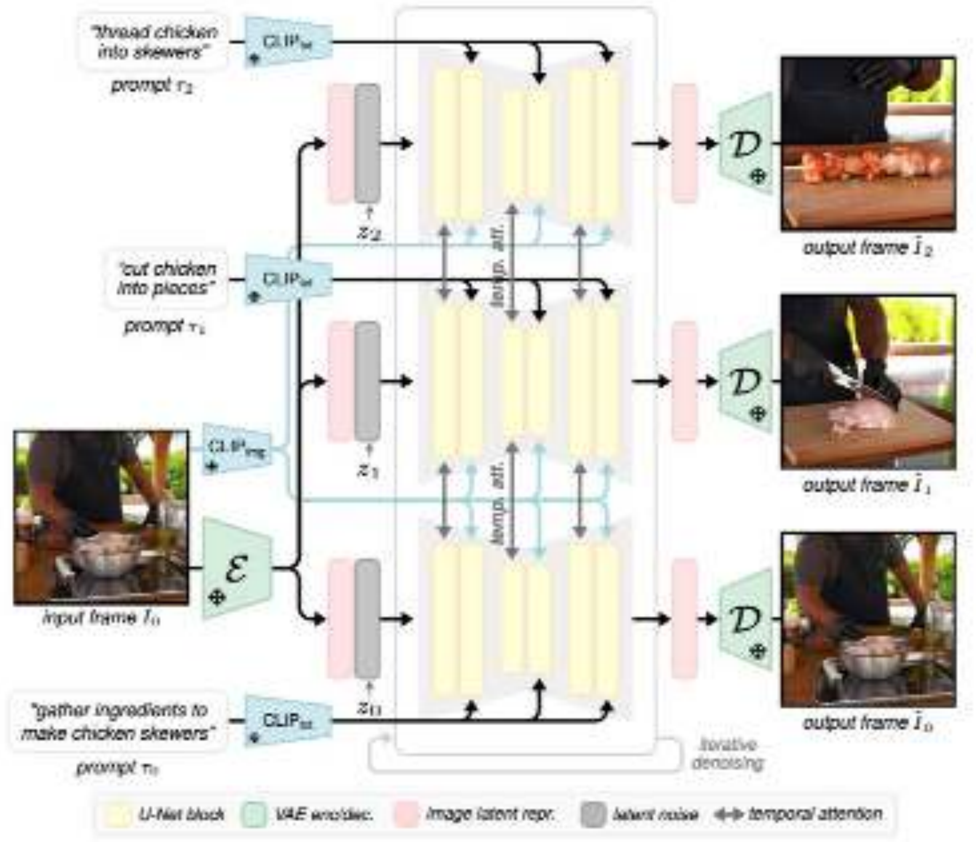
ShowHowTo

with: Tomas Soucek Michael Wray
Prajwal Gatti Ivan Laptev
Josef Sivic



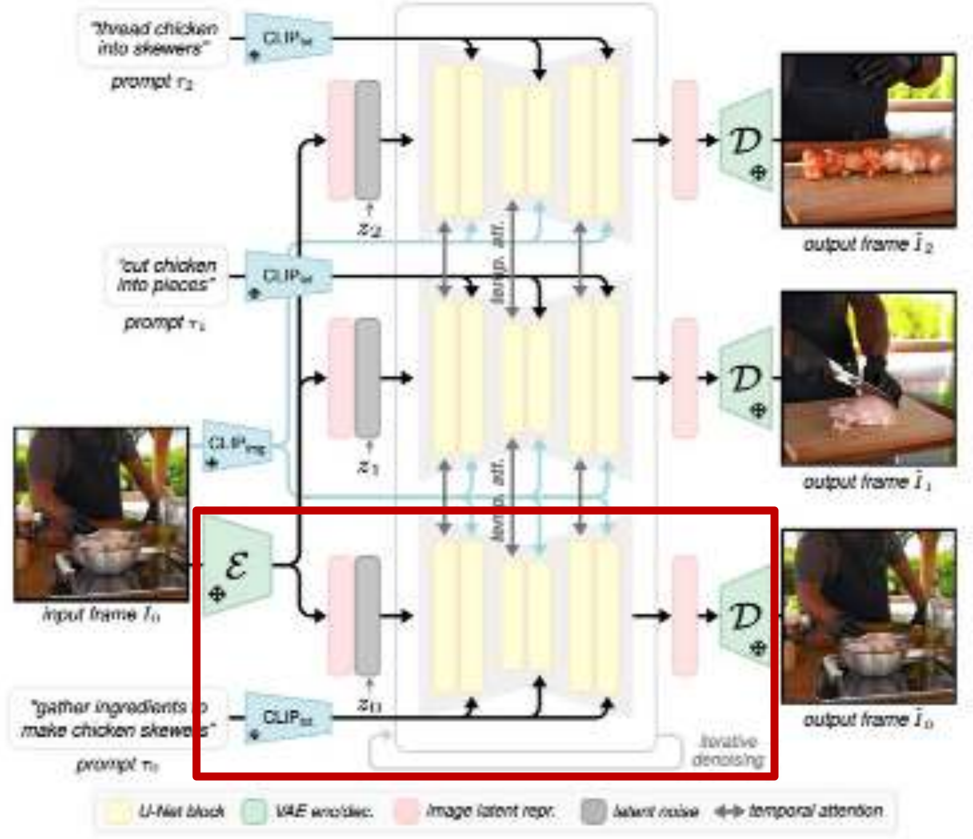
ShowHowTo

with: Tomas Soucek Michael Wray
Prajwal Gatti Ivan Laptev
Josef Sivic



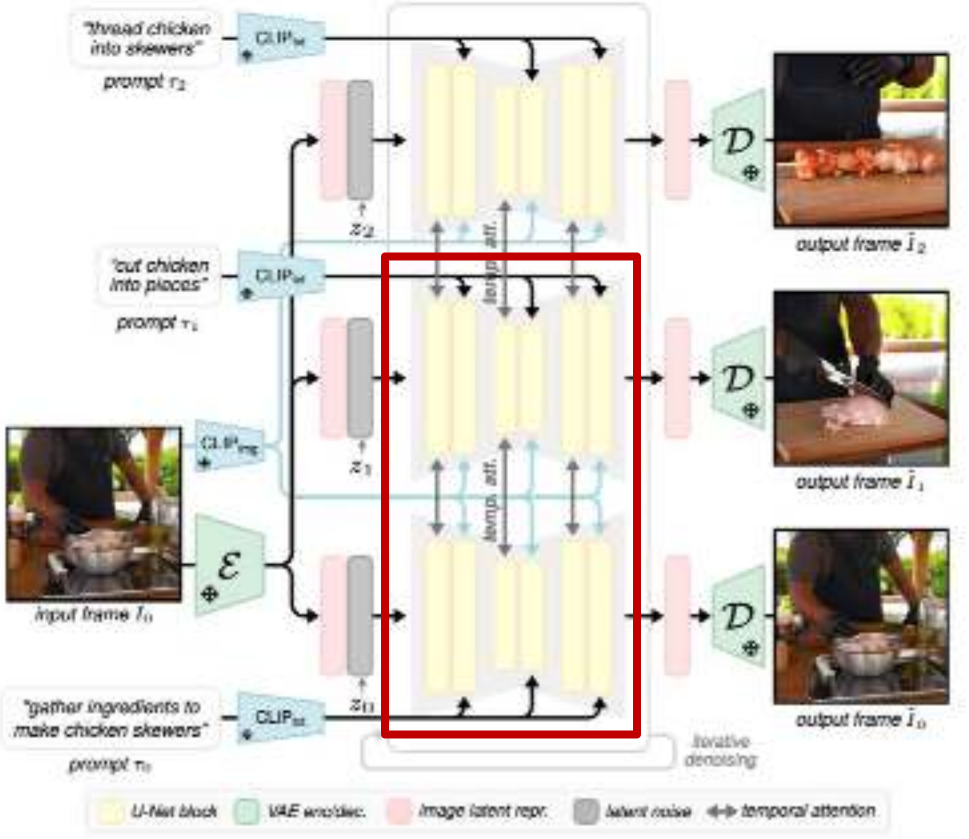
➢ Benefits from a pretrained video generation model (DynamiCrafter)

with: Tomas Soucek Michael Wray
Prajwal Gatti Ivan Laptev
Josef Sivic



- Benefits from a pretrained video generation model (DynamiCrafter)

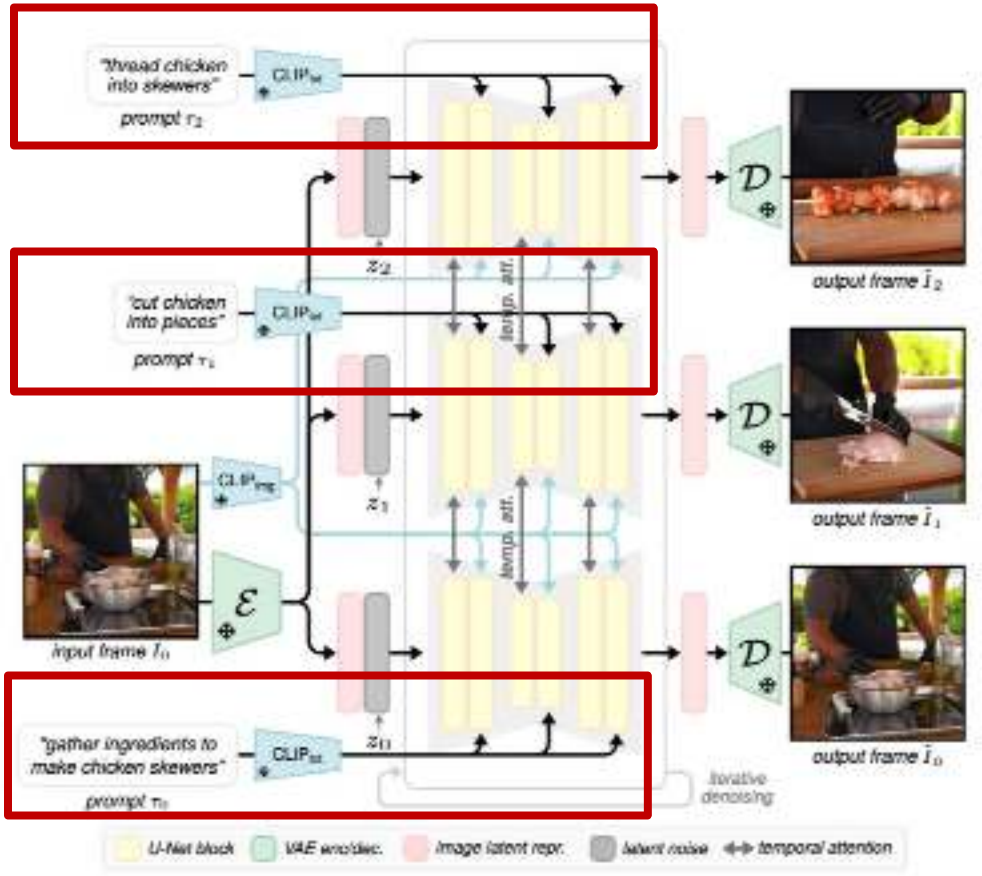
with: Tomas Soucek Michael Wray
Prajwal Gatti Ivan Laptev
Josef Sivic



- ➢ Benefits from a pretrained video generation model (DynamiCrafter)

ShowHowTo

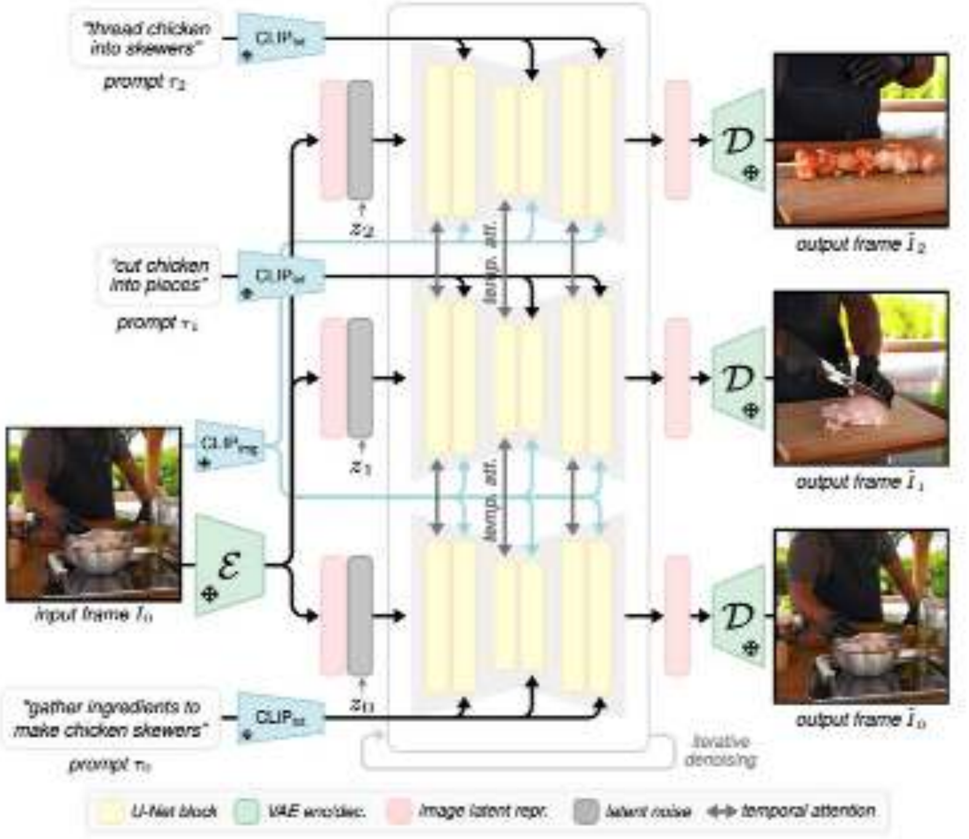
with: Tomas Soucek Michael Wray
Prajwal Gatti Ivan Laptev
Josef Sivic



- ➢ Benefits from a pretrained video generation model (DynamiCrafter)
- ➢ Per-frame text conditioning

ShowHowTo

with: Tomas Soucek Michael Wray
Prajwal Gatti Ivan Laptev
Josef Sivic



- ➢ Benefits from a pretrained video generation model (DynamiCrafter)
- ➢ Per-frame text conditioning
- ➢ Handles variable sequence-length generations

ShowHowTo

with: Tomas Soucek Michael Wray
Prajwal Gatti Ivan Laptev
Josef Sivic

Method	Step Faithf.	Scene Consist.	Task Faithf.	Overall
(a) InstructPix2Pix [12]	0.25	0.17	0.25	0.22
(b) AURORA [35]	0.25	0.33	0.24	0.27
(c) GenHowTo [53]	0.49	0.13	0.27	0.29
(d) Phung *et al.* [45]	0.36	0.03	0.38	0.26
(e) StackedDiffusion [41]	0.43	0.02	**0.42**	0.29
(f) **ShowHowTo**	**0.52**	**0.34**	**0.42**	**0.43**
(g) *Random*	0.19	0.00	0.01	0.07
(h) Stable Diffusion [48]†	0.70	0.03	0.44	0.39
(i) *Copy of the input image*	0.19	0.62	0.39	0.40
(j) *Source sequences*	0.50	1.00	0.56	0.69

Soucek et al (2025). ShowHowTo: Generating Scene-Conditioned Step-by-Step Visual Instructions. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

ShowHowTo

with: Tomas Soucek Michael Wray
Prajwal Gatti Ivan Laptev
Josef Sivic

Soucek et al (2025). ShowHowTo: Generating Scene-Conditioned Step-by-Step Visual Instructions. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

ShowHowTo

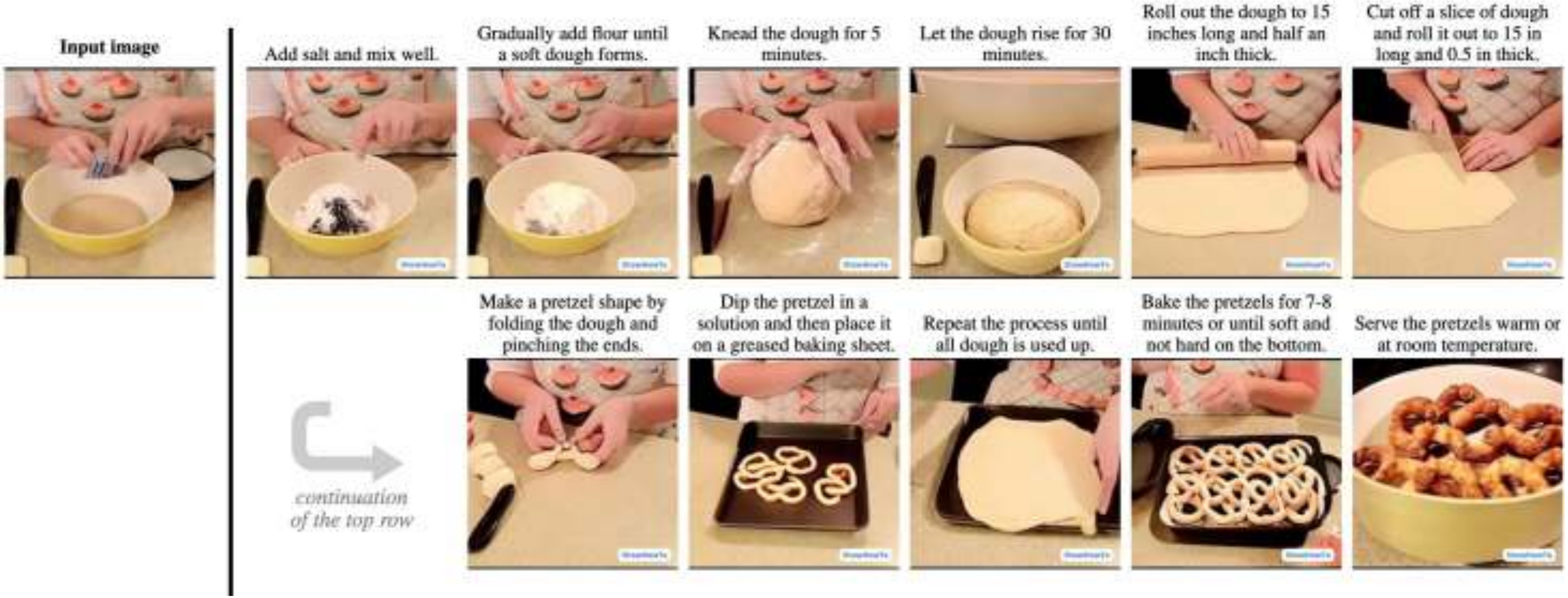
with: Tomas Soucek   Michael Wray
Prajwal Gatti   Ivan Laptev
Josef Sivic

University of BRISTOL

Soucek et al (2025). ShowHowTo: Generating Scene-Conditioned Step-by-Step Visual Instructions. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

ShowHowTo

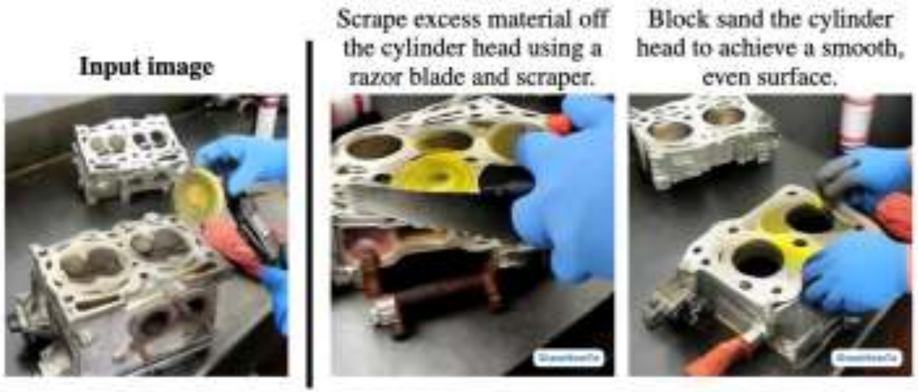
with: Tomas Soucek Michael Wray
Prajwal Gatti Ivan Laptev
Josef Sivic



Soucek et al (2025). ShowHowTo: Generating Scene-Conditioned Step-by-Step Visual Instructions. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

ShowHowTo

with: Tomas Soucek Michael Wray
Prajwal Gatti Ivan Laptev
Josef Sivic



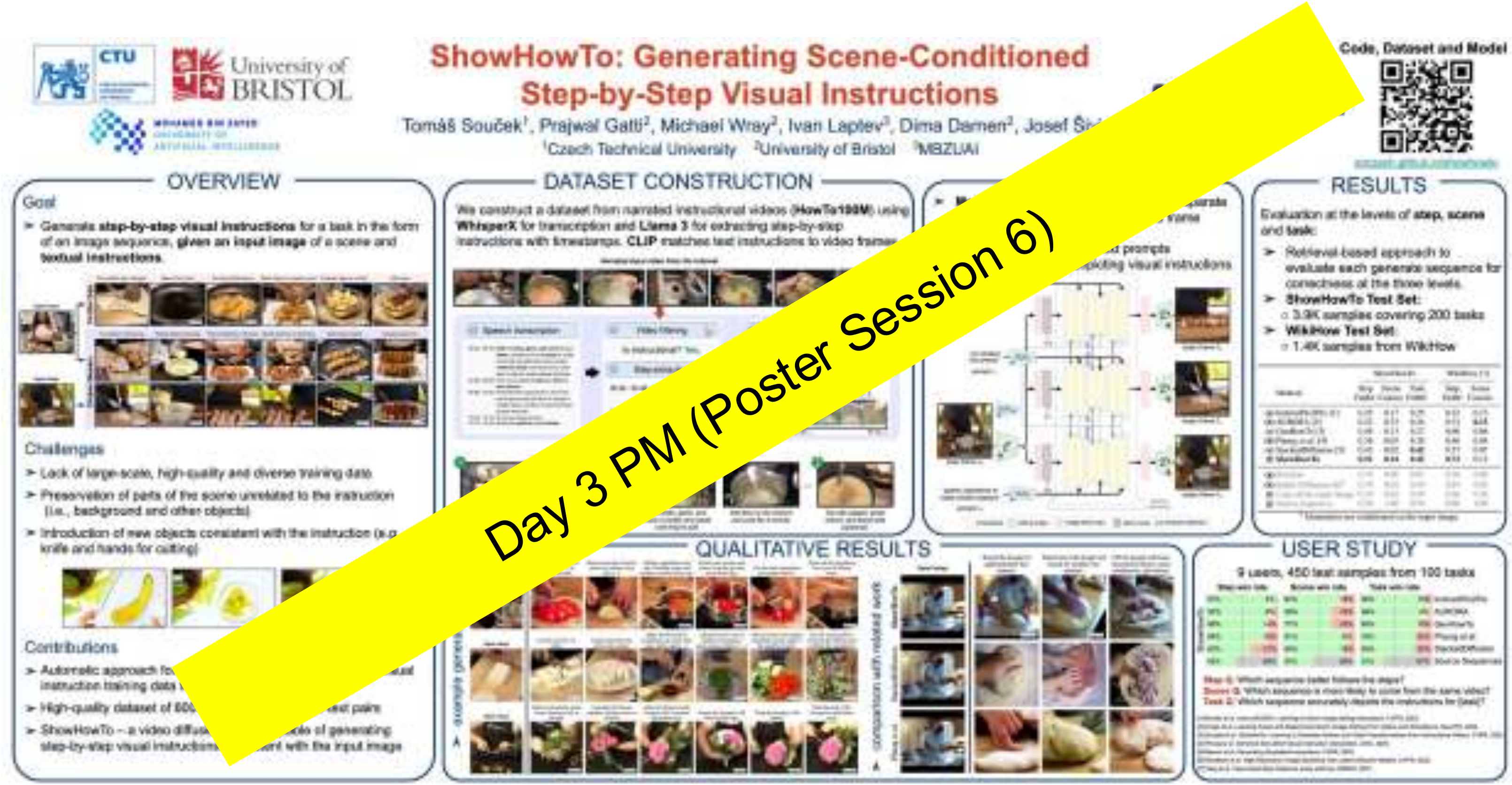
Can struggle with rare objects or tools

Can fail to update object states
E.g., Raw → Cooked → Raw

Soucek et al (2025). ShowHowTo: Generating Scene-Conditioned Step-by-Step Visual Instructions. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

ShowHowTo

with: Tomas Soucek Michael Wray
Prajwal Gatti Ivan Laptev
Josef Sivic



Soucek et al (2025). ShowHowTo: Generating Scene-Conditioned Step-by-Step Visual Instructions. IEEE/CVF Computer Vision and Pattern Recognition (CVPR)

In this talk…

Unique Captioning

Visual Instructions

Learning from
Continuous Streams

Out of Sight,
Not Out of Mind

HD-EPIC

HD-EPIC: A Highly-Detailed Egocentric Video
Dataset

EPIC-KITCHENS

Out of Sight, not Out of Mind

with: Chiara Plizzari Shubham Goel
Toby Perrett Angjoo Kanazawa

Spatial Cognition from Egocentric Video:

Out of Sight, Not Out of Mind

Chiara Plizzari Shubham Goel Toby Perrett Jacob Chalk
Angjoo Kanazawa Dima Damen

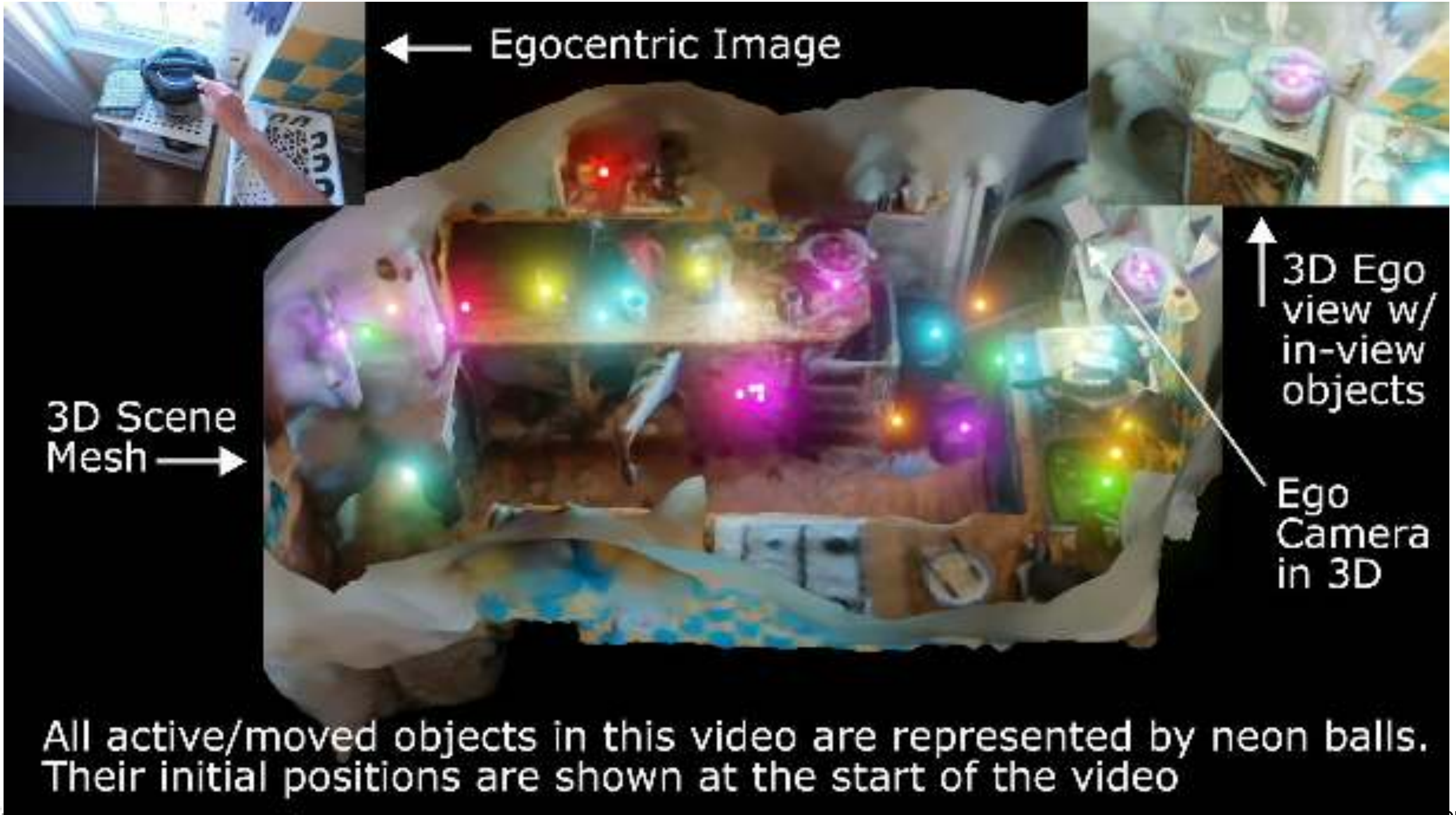
http://dimadamen.github.io/OSNOM

Politecnico di Torino
Berkeley UNIVERSITY OF CALIFORNIA
University of BRISTOL

Plizzari et al (2025). Spatial Cognition from Egocentric Video: Out of Sight, Not Out of Mind. 3DV

Out of Sight, not Out of Mind

with: Chiara Plizzari Shubham Goel
Toby Perrett Angjoo Kanazawa



Plizzari et al (2025). Spatial Cognition from Egocentric Video: Out of Sight, Not Out of Mind. 3DV

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

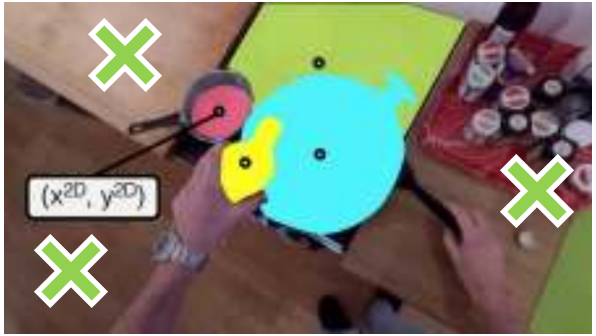
Out of Sight, not Out of Mind

with: Chiara Plizzari
Toby Perrett
Shubham Goel
Angjoo Kanazawa

Lift

Match

Keep



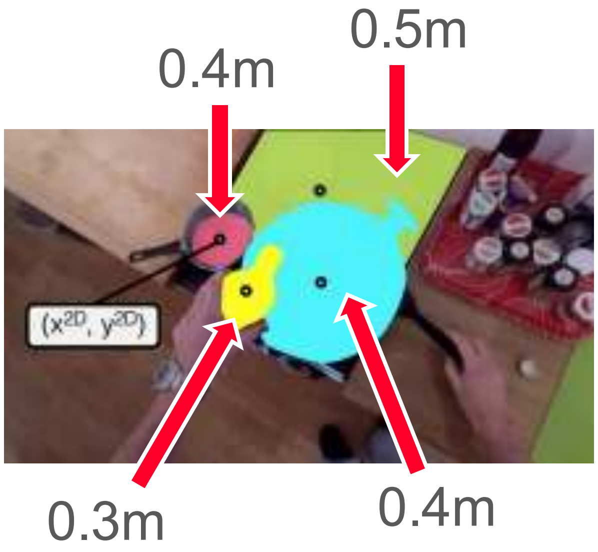
0.0 … 1.0

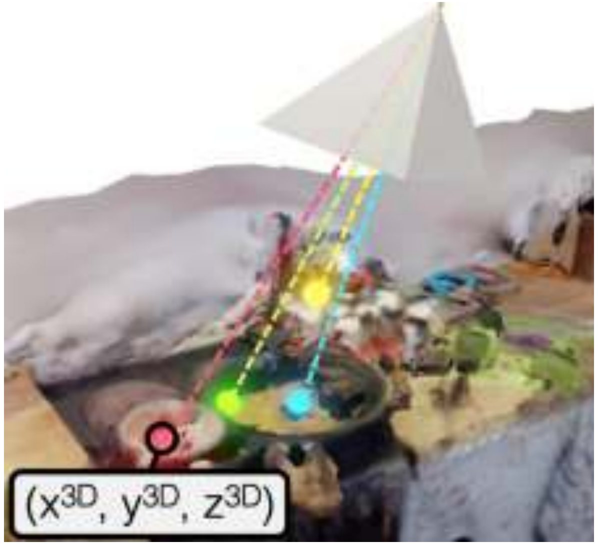
0.3m … 1.8m

Out of Sight, not Out of Mind

with: Chiara Plizzari
Toby Perrett
Shubham Goel
Angjoo Kanazawa

Lift → Match → Keep

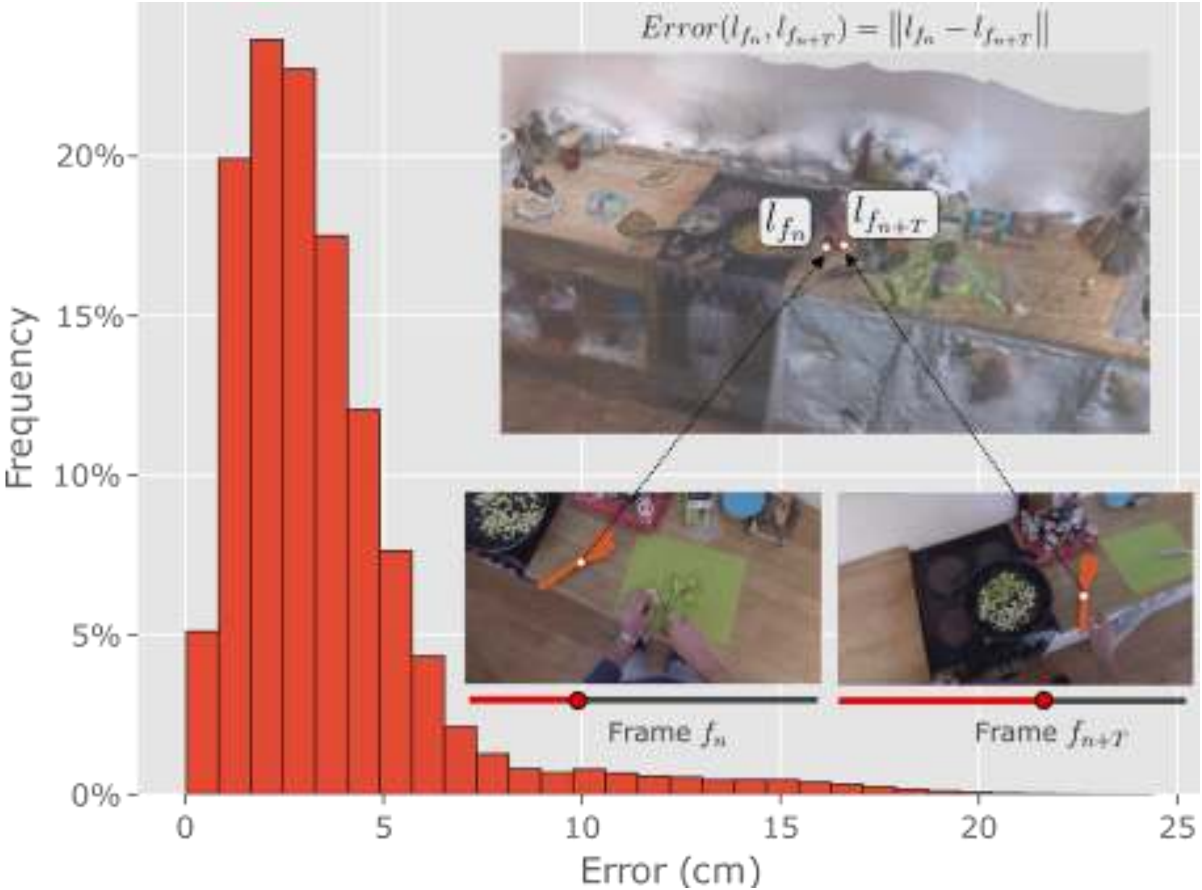




Plizzari et al (2025). Spatial Cognition from Egocentric Video: Out of Sight, Not Out of Mind. 3DV

Out of Sight, not Out of Mind

with: Chiara Plizzari
Toby Perrett
Shubham Goel
Angjoo Kanazawa



Plizzari et al (2025). Spatial Cognition from Egocentric Video: Out of Sight, Not Out of Mind. 3DV

with: Chiara Plizzari
Toby Perrett
Shubham Goel
Angjoo Kanazawa

Lift → Match → Keep

Instead of tracking in 2D, we track in 3D, using combination of appearance and location distances

Plizzari et al (2025). Spatial Cognition from Egocentric Video: Out of Sight, Not Out of Mind. 3DV

Out of Sight, not Out of Mind

with: Chiara Plizzari
Toby Perrett
Shubham Goel
Angjoo Kanazawa

After we Lift, Match and Keep (LMK), we can reason about an object's visibility and position

- In-View vs Out-of-View
- In-Sight vs Out-of-Sight (Occluded)
- Within-Reach vs Out-of-Reach (defining the camera wearer's near space)

University of BRISTOL

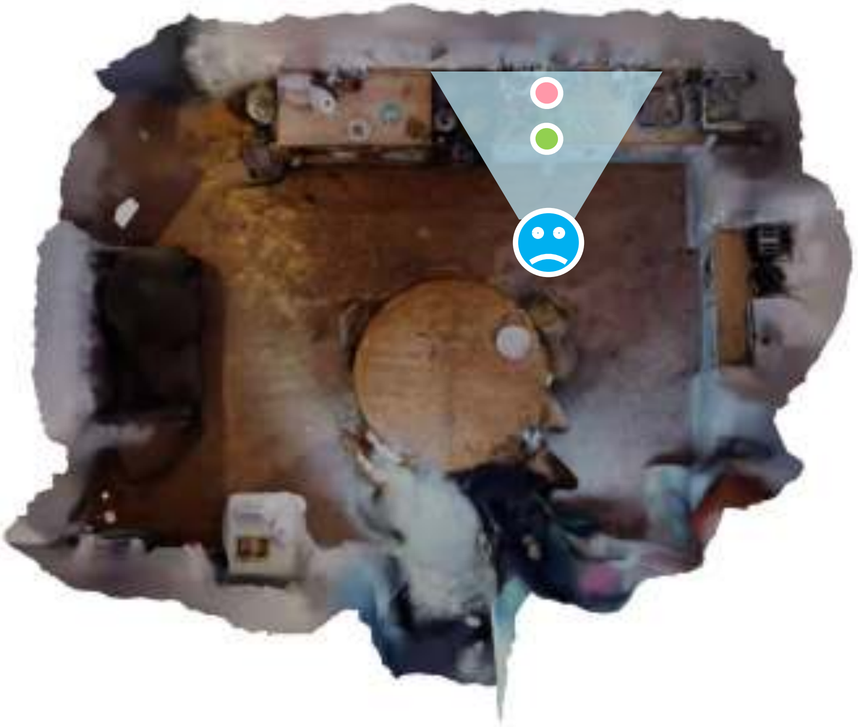
Plizzari et al (2025). Spatial Cognition from Egocentric Video: Out of Sight, Not Out of Mind. 3DV

Out of Sight, not Out of Mind

with: Chiara Plizzari Toby Perrett Shubham Goel Angjoo Kanazawa

After we Lift, Match and Keep (LMK), we can reason about an object's visibility and position

- In-View vs Out-of-View
- In-Sight vs Out-of-Sight (Occluded)
- Within-Reach vs Out-of-Reach (defining the camera wearer's near space)

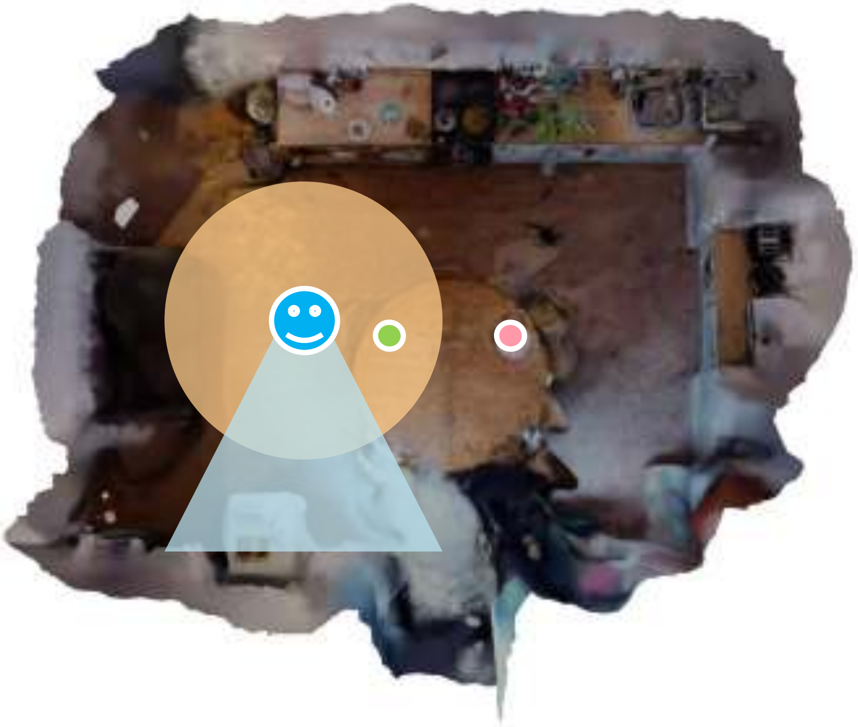


Out of Sight, not Out of Mind

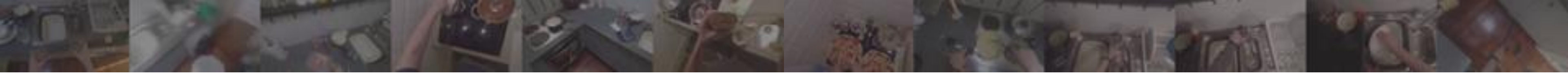
with: Chiara Plizzari
Toby Perrett
Shubham Goel
Angjoo Kanazawa

After we Lift, Match and Keep (LMK), we can reason about an object's visibility and position

- In-View vs Out-of-View
- In-Sight vs Out-of-Sight (Occluded)
- Within-Reach vs Out-of-Reach (defining the camera wearer's near space)



Plizzari et al (2025). Spatial Cognition from Egocentric Video: Out of Sight, Not Out of Mind. 3DV



Spatial Cognition from Egocentric Video:

Out of Sight, Not Out of Mind

Ground-Truth??

Chiara Plizzari   Shubham   Toby Perrett   Jacob Chalk

Angi... a   Dima Damen

http://dimadamen.github.io/OSNOM

Politecnico di Torino   Berkeley UNIVERSITY OF CALIFORNIA   University of BRISTOL

Plizzari et al (2025). Spatial Cognition from Egocentric Video: Out of Sight, Not Out of Mind. 3DV

In this talk…

Unique Captioning

Visual Instructions

Learning from
Continuous Streams

Out of Sight,
Not Out of Mind

HD-EPIC

HD-EPIC: A Highly-Detailed Egocentric Video Dataset

HD-EPIC

HD-EPIC: A Highly-Detailed Egocentric Video Dataset

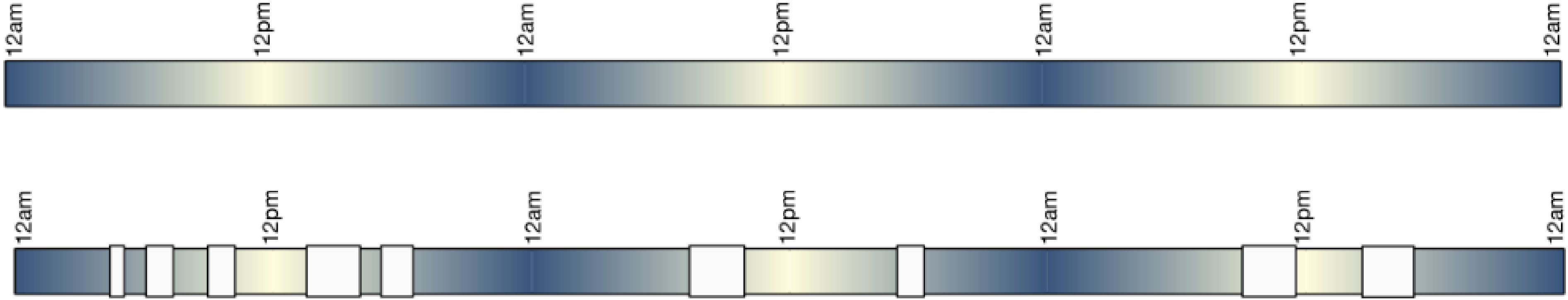
University of Bristol
Universiteit Leiden
Singapore Management University
University of Bath

Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha
Omar Emara, Sam Pollard, Kranti Parida
Kaiting Liu, Prajwal Gatti, Siddhant Bansal
Kevin Flanagan, Jacob Chalk, Zhifan Zhu
Rhodri Guerrier, Fahd Abdelazim, Bin Zhu
Davide Moltisanti, Michael Wray, Hazel Doughty
Dima Damen

Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR

HD-EPIC

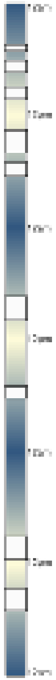
HD-EPIC



Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

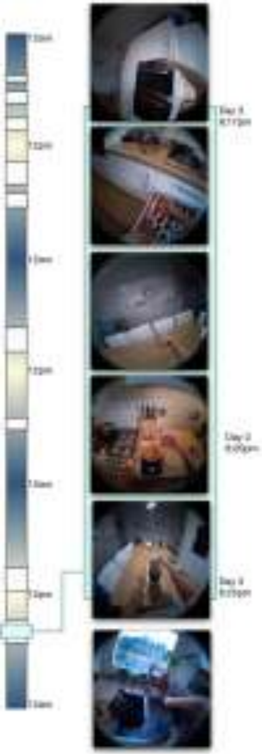
HD-EPIC

HD-EPIC



Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

HD-EPIC



Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

HD-EPIC

ay 3
:17pm

Recipe: Southwestern Salad

1: Preheat the oven to 400F

2: Wash and peel the sweet potatoes and chop into bite-sized pieces. Put the sweet potatoes in a bowl and add the olive oil, cumin, and chili powder. Pour onto tray and roast for 10 mins.

3: Pulse all the dressing ingredients in a food processor until mostly smooth.

Recipe and nutrition

HD-EPIC

Cacio e Pepe (modified)

Ingredients:

200 g ~~400g~~ of pasta of your choice → penne
(we recommend bucatini)

1 ~~2~~ tablespoon of black peppercorn

30 g ~~200g~~ of freshly grated ~~pecorino~~ parmigiano cheese

+25g of slightly salted butter

Steps:

1. Toast the peppercorns until fragrant in a dry frying pan over medium heat, about 2 minutes. Keep them moving to prevent them from burning. ~~Once toasted, roughly crush.~~ → step 2

2. Cook your choice of pasta in a large pot of generously salted boiling water ~~for around 4-6 minutes~~, or until al dente. → step 1

3. While the pasta cooks, add freshly grated cheese and crushed black peppercorns to a large serving bowl. on very low heat Gradually add a cup of the boiling cooking water constantly mixing to obtain a silky, smooth sauce that's able to completely coat the pasta. → step 3

University of BRISTOL

Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

Dima Damen
HVU - CVPRW'25

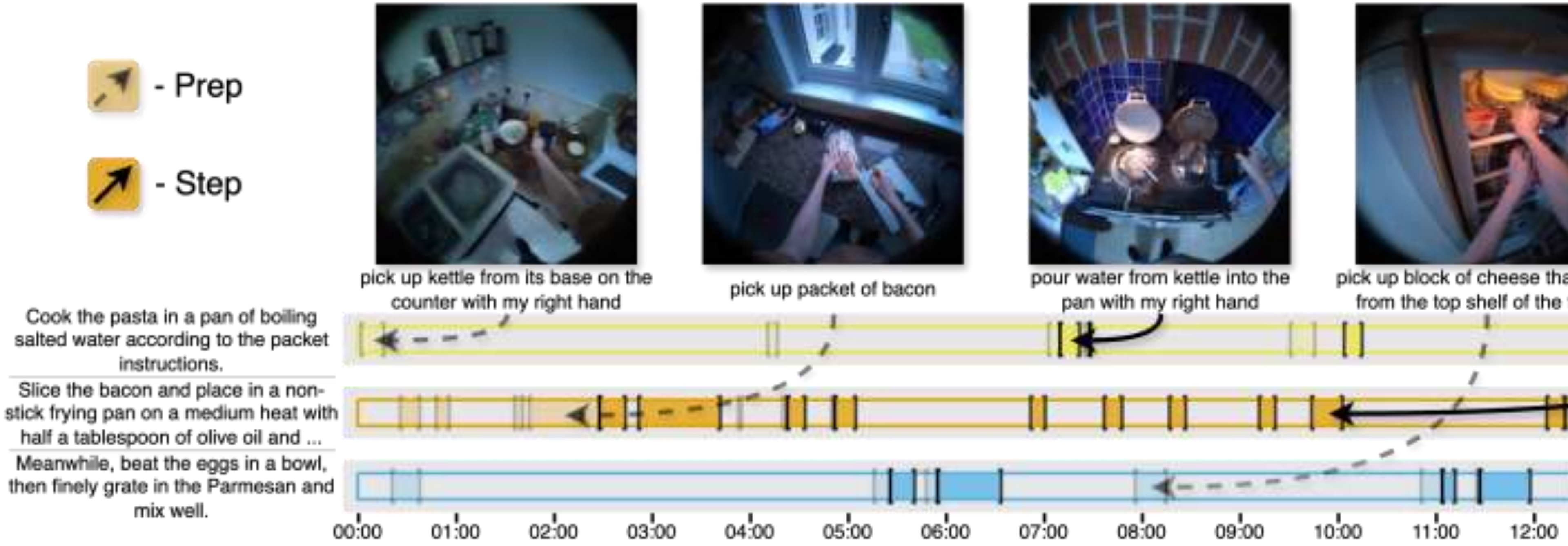
HD-EPIC

- The **prep** of a corresponding **step** is defined as
all essential actions the participant takes to get ready to execute a given step.

- For example, the **step** ‘chop tomato’:
  - **Prep:** retrieve tomato from storage, wash tomato, retrieve a knife and chopping board.

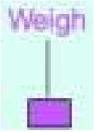
- the **step** ‘add chopped onions and stir’:
  - **Prep:** retrieve tomato from storage, wash tomato, retrieve a knife and chopping board, **and chop the onions.**

HD-EPIC


- Prep
- Step
pick up kettle from its base on the counter with my right hand
pick up packet of bacon
pour water from kettle into the pan with my right hand
pick up block of cheese tha
from the top shelf of the
Cook the pasta in a pan of boiling salted water according to the packet instructions.
Slice the bacon and place in a non-stick frying pan on a medium heat with half a tablespoon of olive oil and ...
Meanwhile, beat the eggs in a bowl, then finely grate in the Parmesan and mix well.
00:00
01:00
02:00
03:00
04:00
05:00
06:00
07:00
08:00
09:00
10:00
11:00
12:00

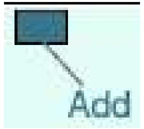
HD-EPIC

"P01_R03_I01": {
"name": "penne pasta",
"amount": 125,
"amount_unit": "g",
"calories": 445,
"fat": 1.9,
"carbs": 90,
"protein": 15,

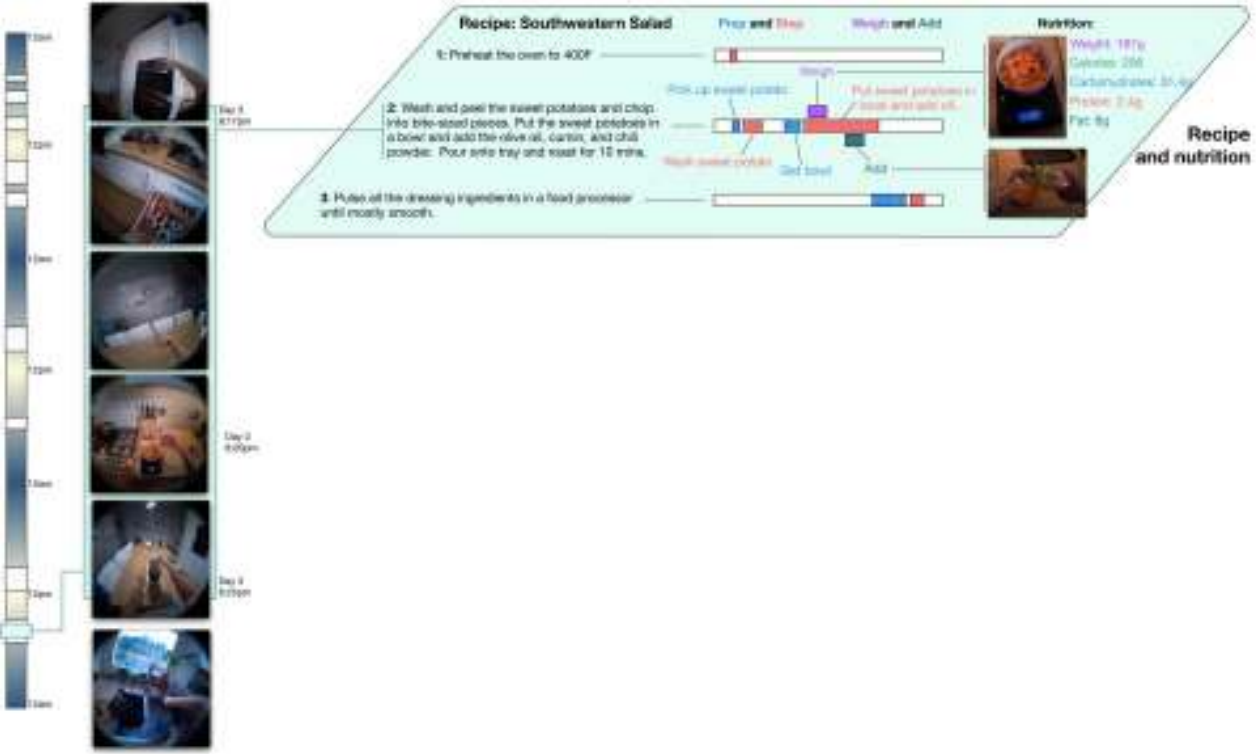


HD-EPIC

"P01_R03_I01": {
"name": "penne pasta",
"amount": 125,
"amount_unit": "g",
"calories": 445,
"fat": 1.9,
"carbs": 90,
"protein": 15,

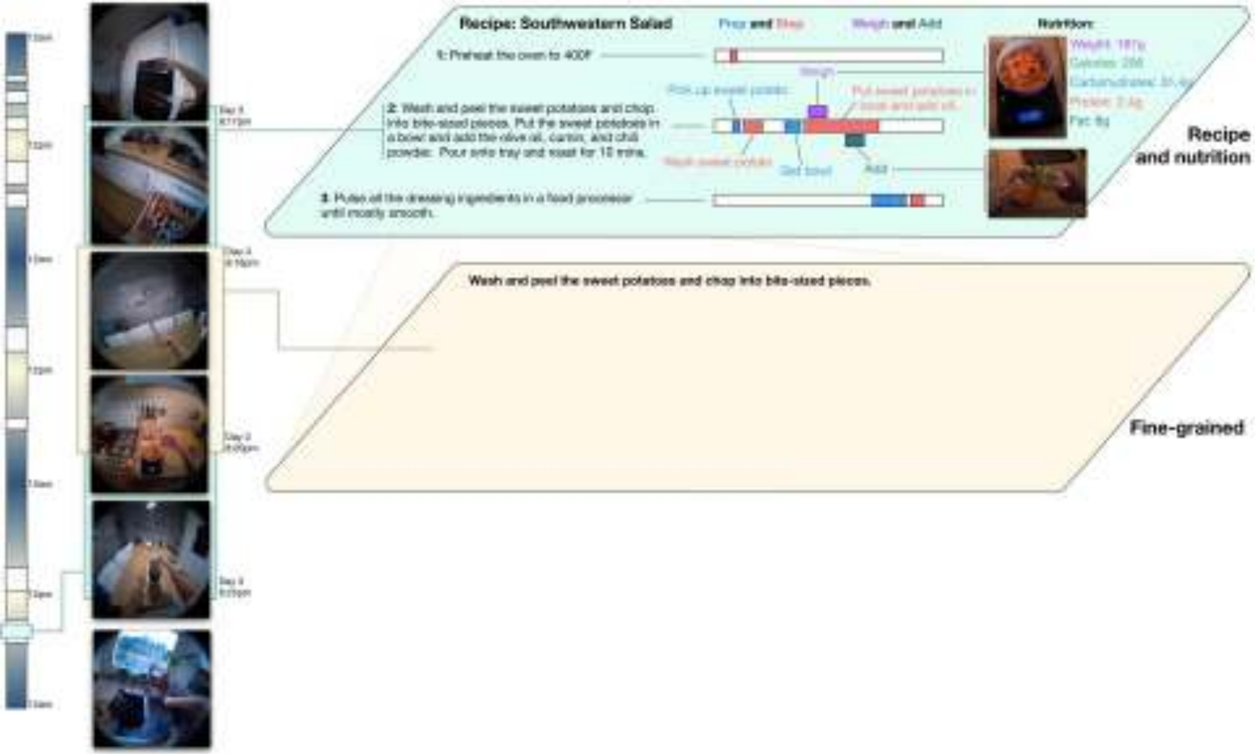


HD-EPIC



Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

HD-EPIC



Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

HD-EPIC

Wash and peel the sweet potatoes and chop into bite-sized pieces.

Narrations

Reach the back of the counter, pick up a packet of avocados and a sweet potato and slide them forward to the front of the kitchen island.

Fine-grained

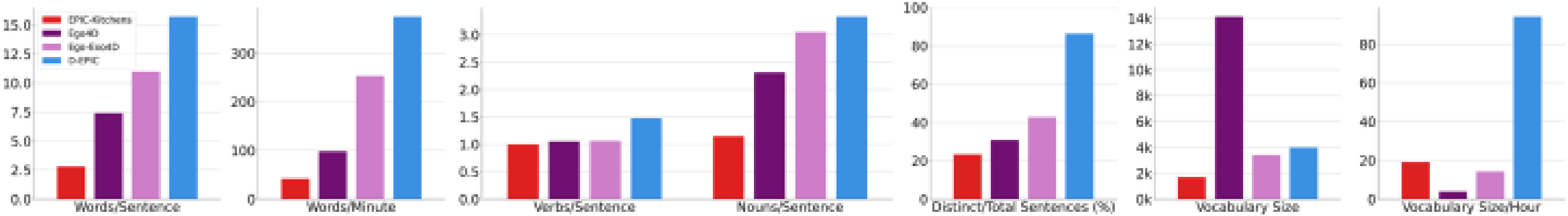
University of BRISTOL

Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

Highly-Detailed Narrations

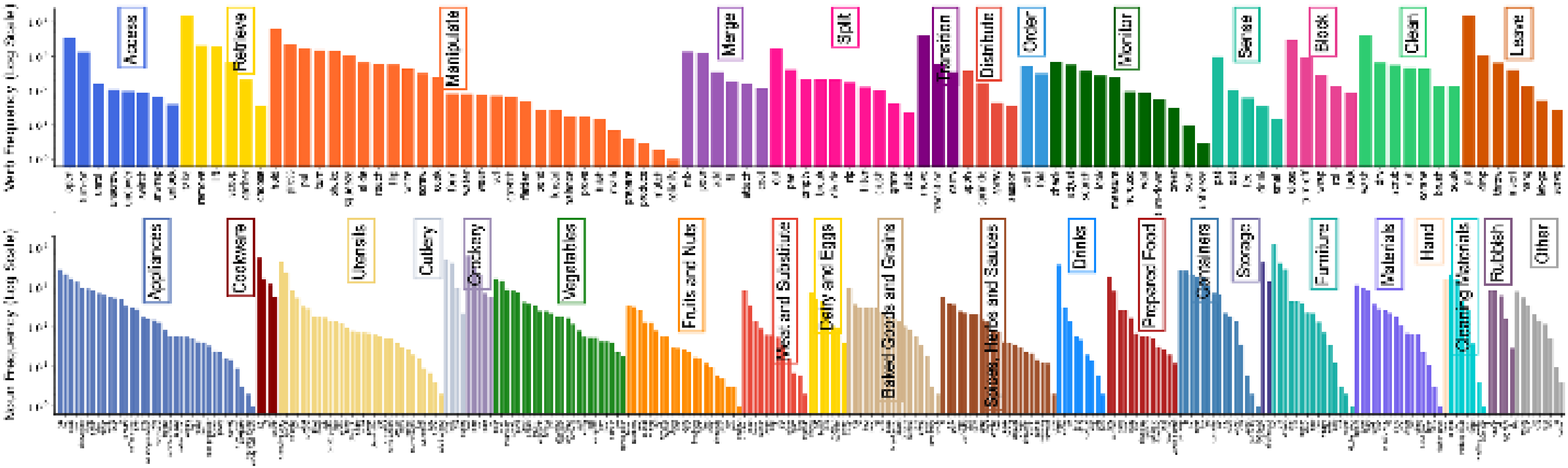
HD-EPIC

- 59,454 fine-grained actions, with a mean duration of 2.0s (±3.4s).

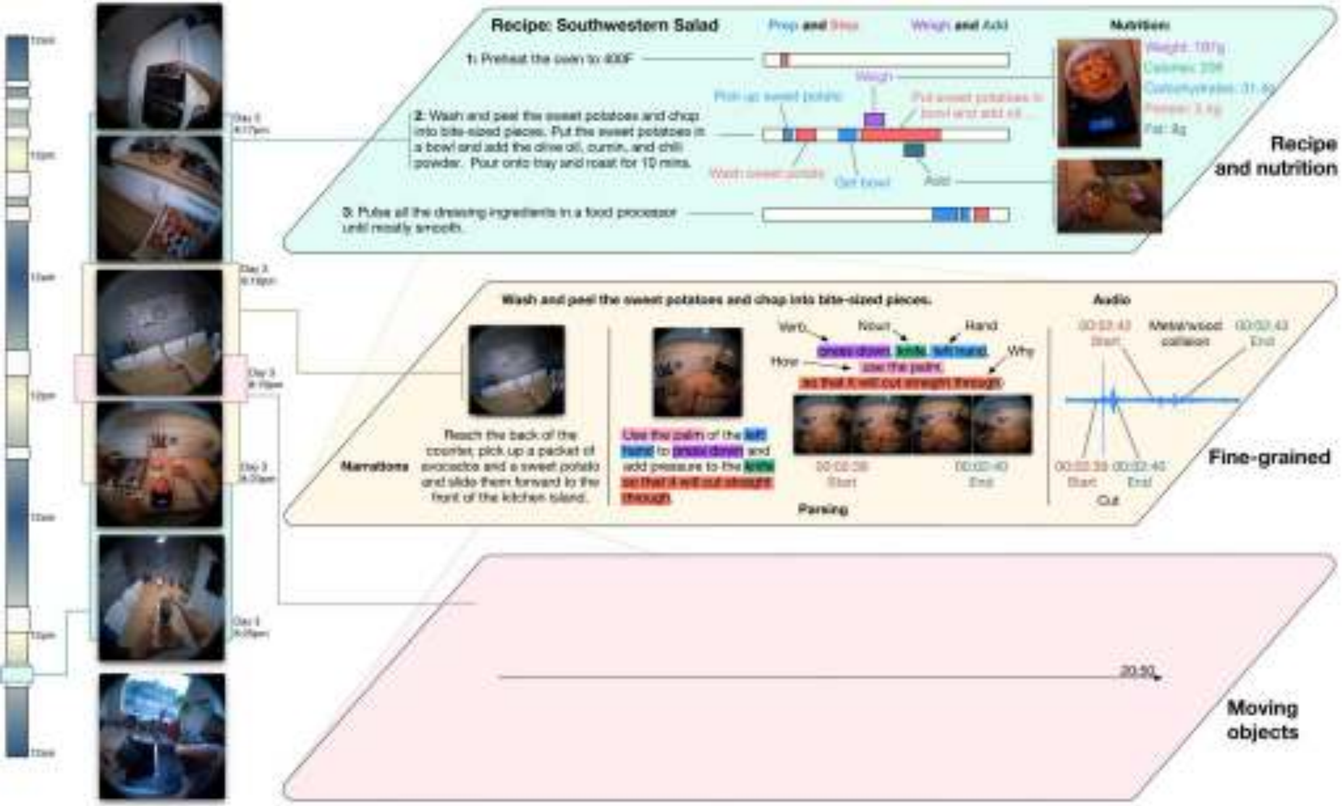


HD-EPIC

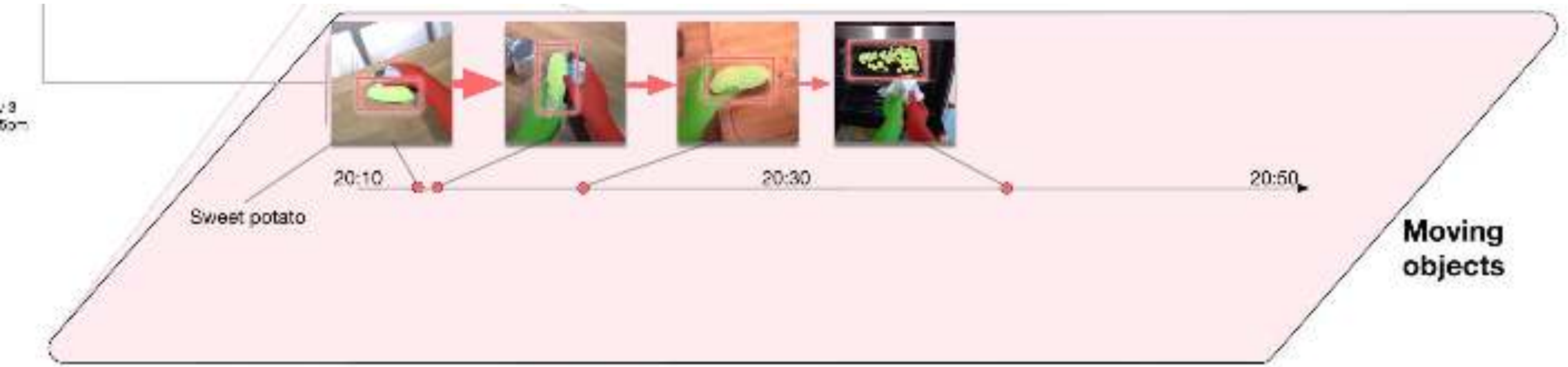
- 59,454 fine-grained actions, with a mean duration of 2.0s (±3.4s).



HD-EPIC



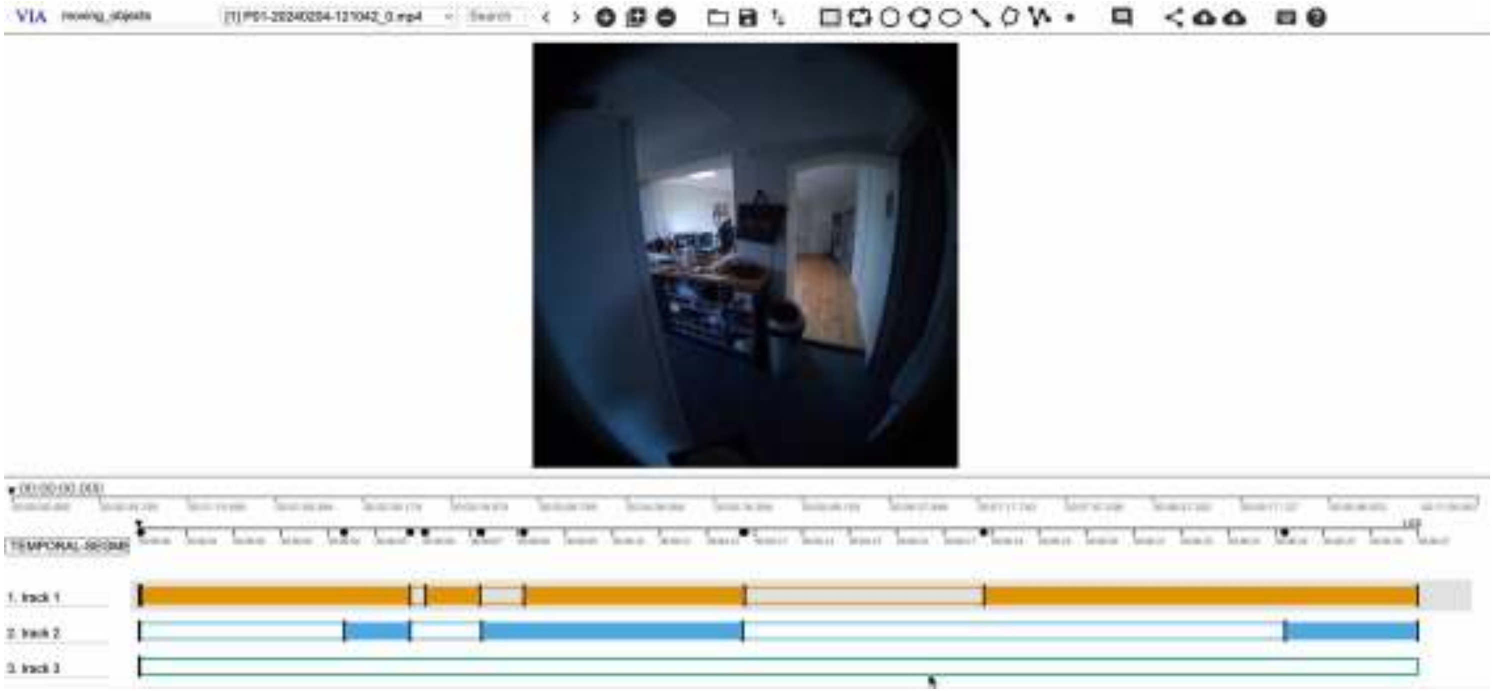
Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.



Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

HD-EPIC

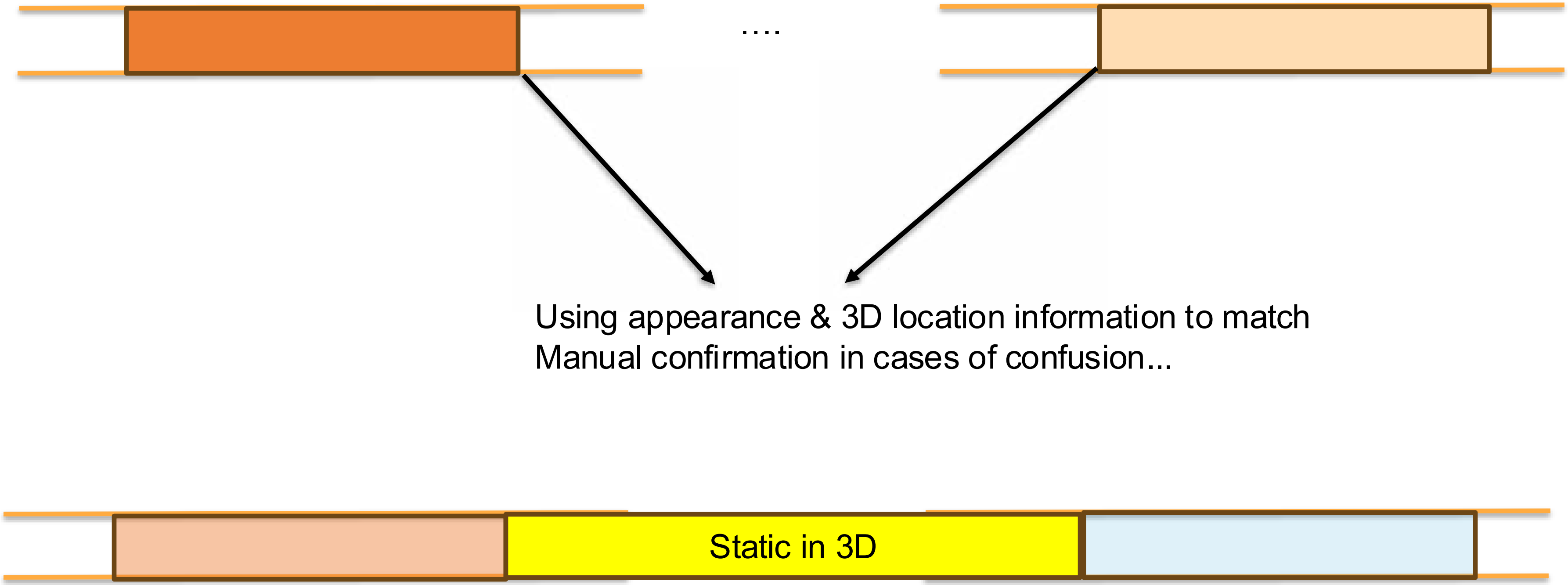
- How to minimize the annotations for tracking objects…



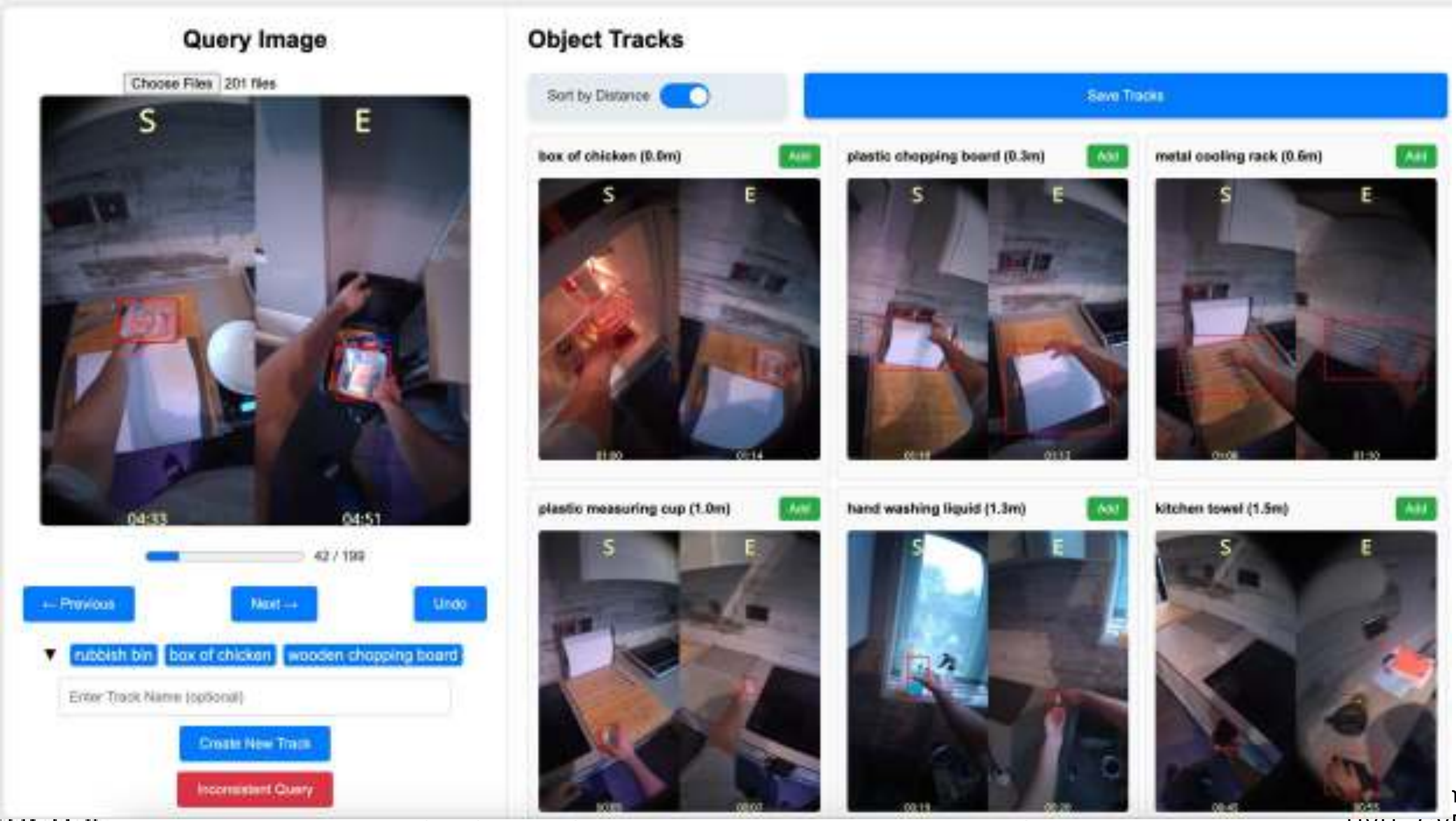
Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

HD-EPIC

- How to minimize the annotations for tracking objects…

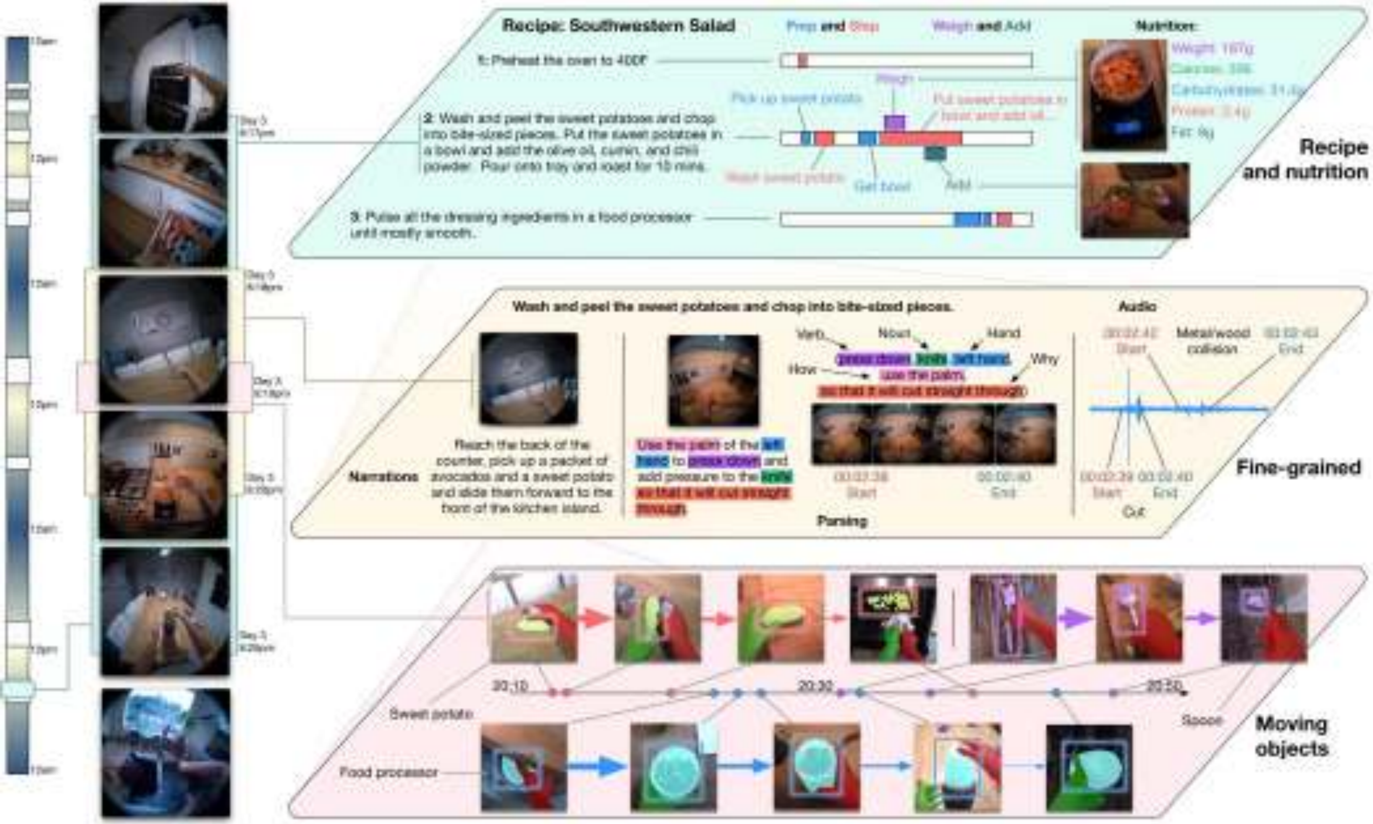


HD-EPIC



Query Image
Choose Files 201 files
S
E
04:33
04:51
42 / 199
← Previous
Next →
Undo
rubbish bin
box of chicken
wooden chopping board
Enter Track Name (optional)
Create New Track
Inconsistent Query
Object Tracks
Sort by Distance
Save Tracks
box of chicken (0.0m)
Add
plastic chopping board (0.3m)
Add
metal cooling rack (0.6m)
Add
plastic measuring cup (1.0m)
Add
hand washing liquid (1.3m)
Add
kitchen towel (1.5m)
Add

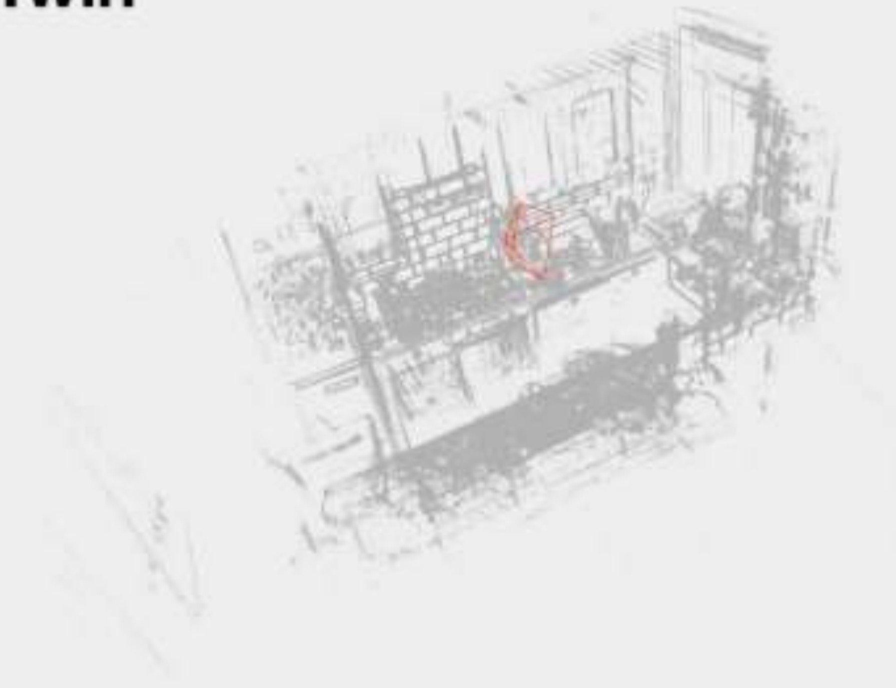
HD-EPIC



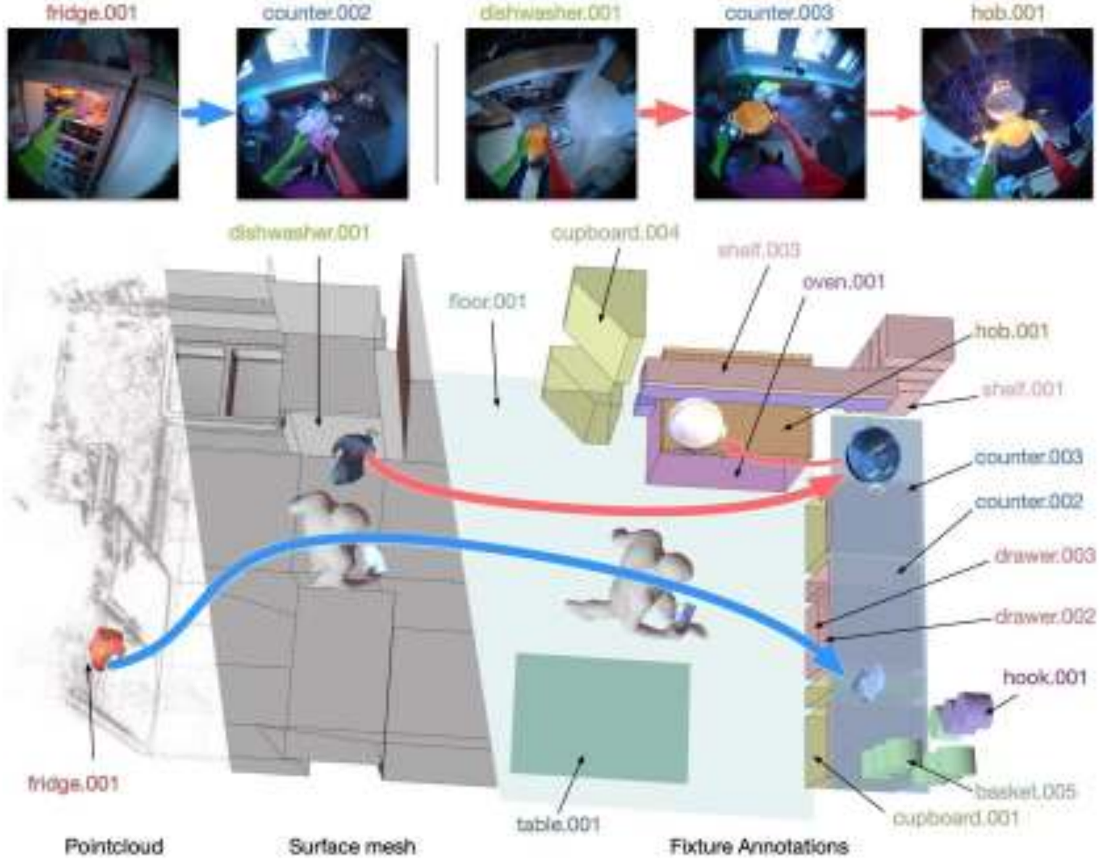
Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

Dima Damen
HVU - CVPRW'25

Digital Twin

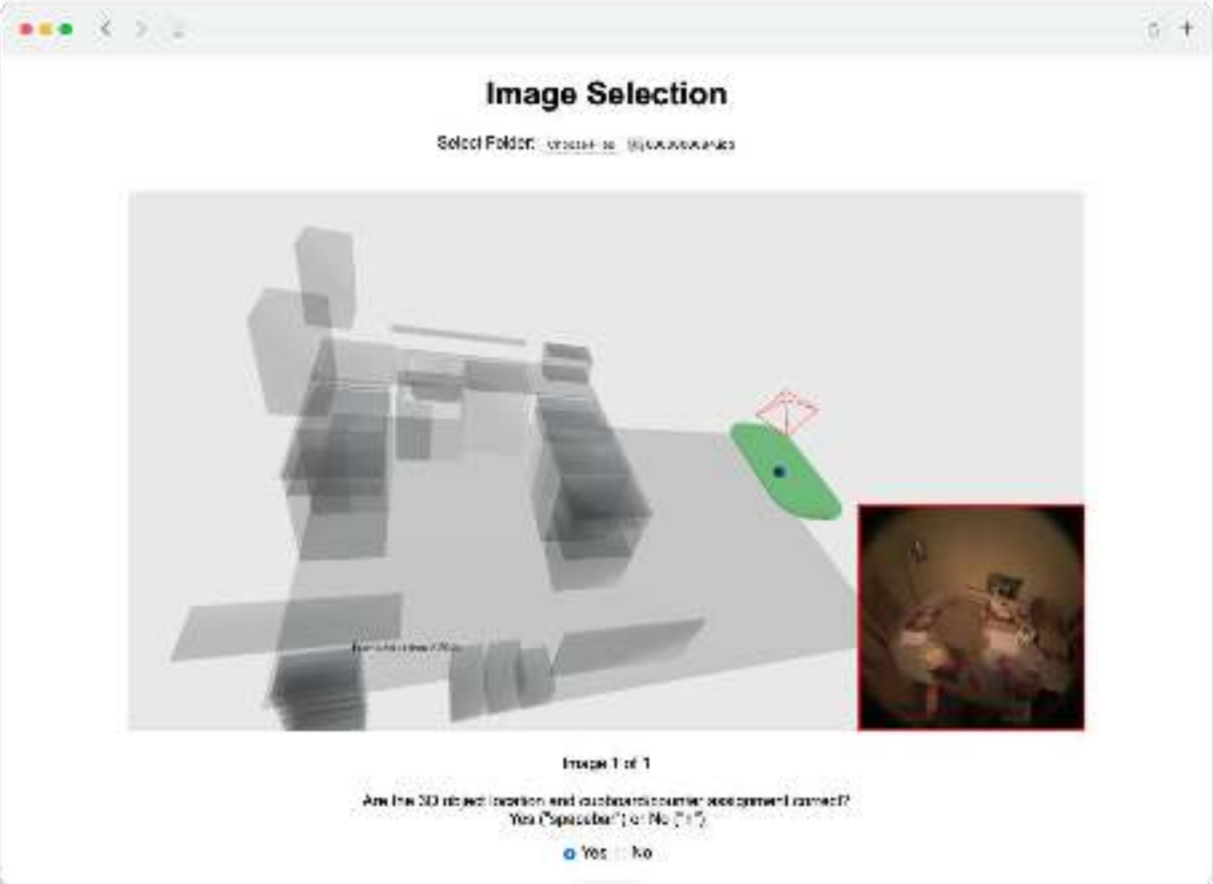


HD-EPIC



Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

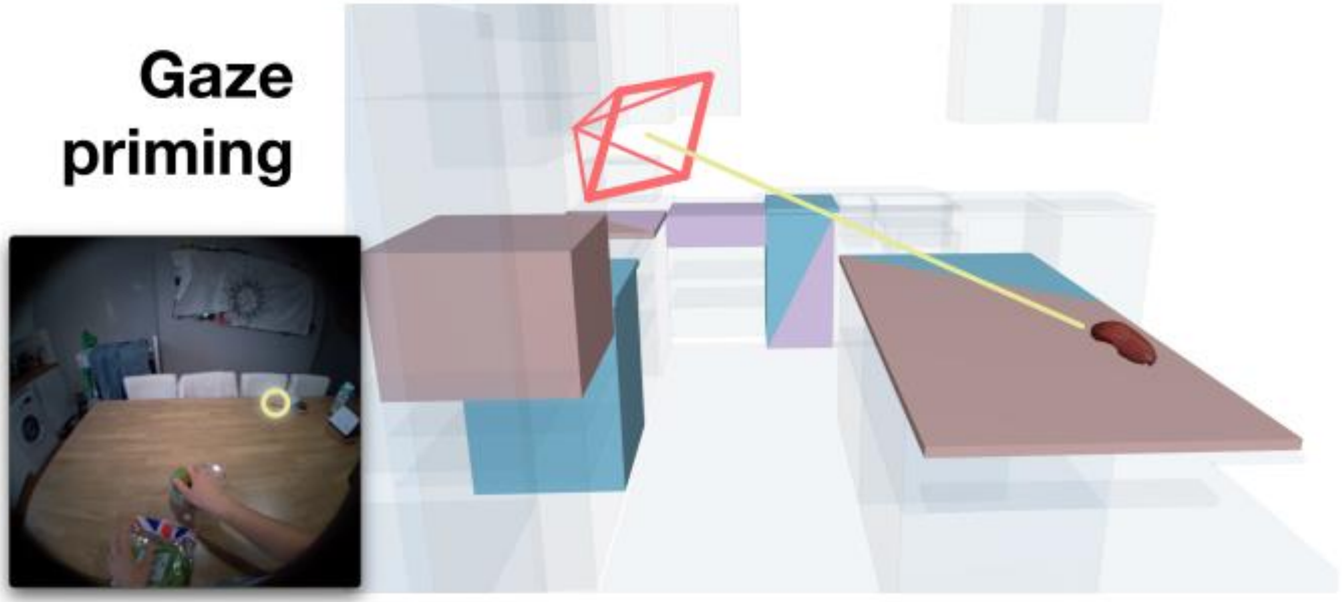
HD-EPIC

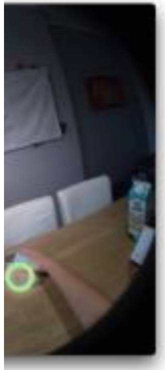


Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

HD-EPIC

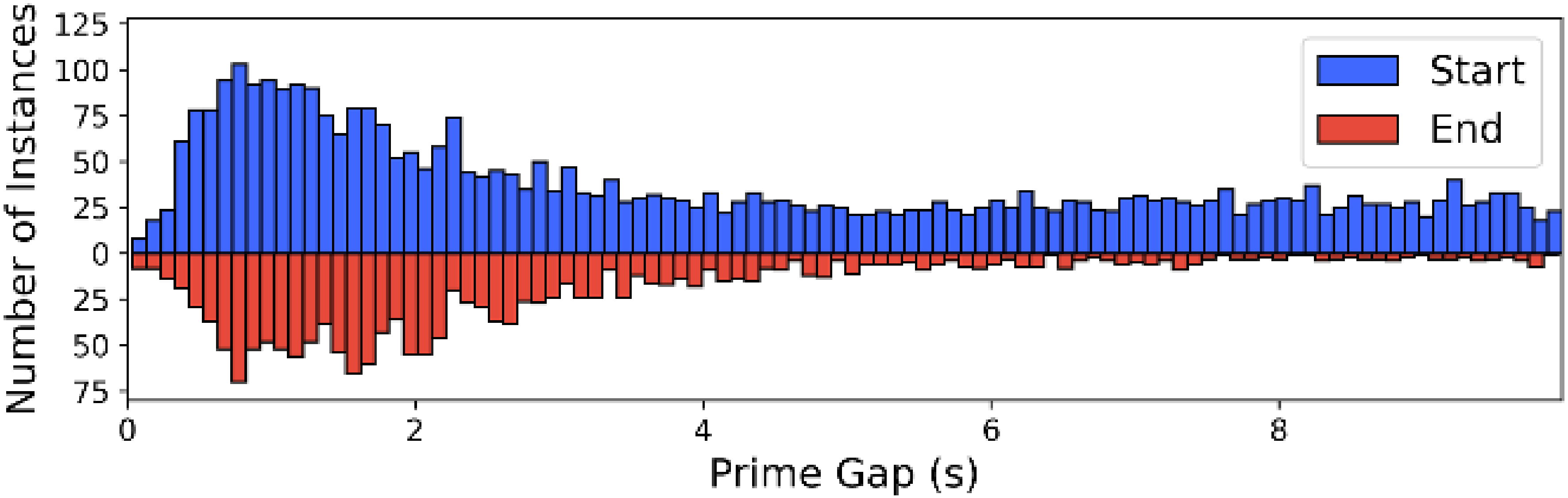
Gaze priming





Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

HD-EPIC

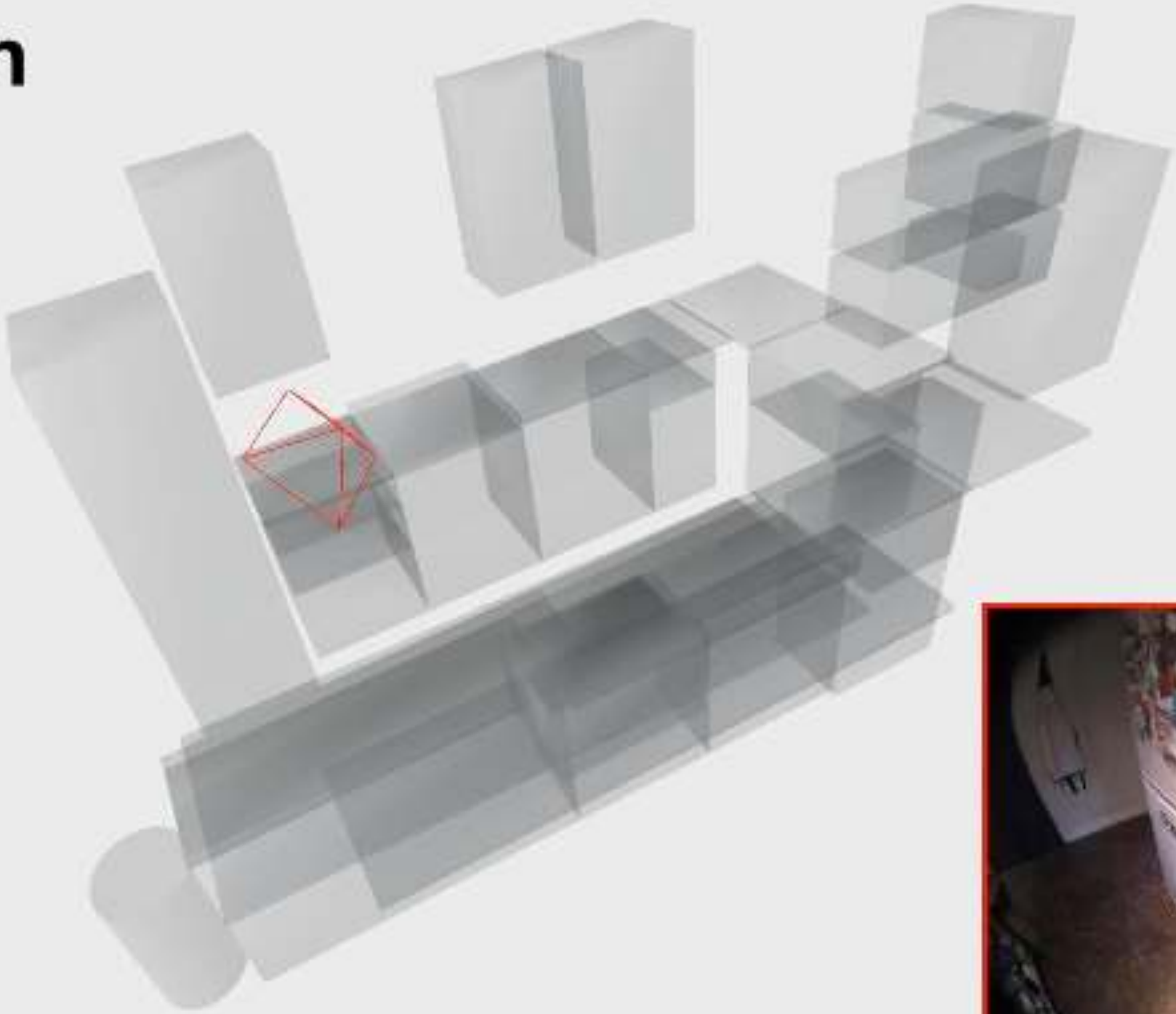


Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

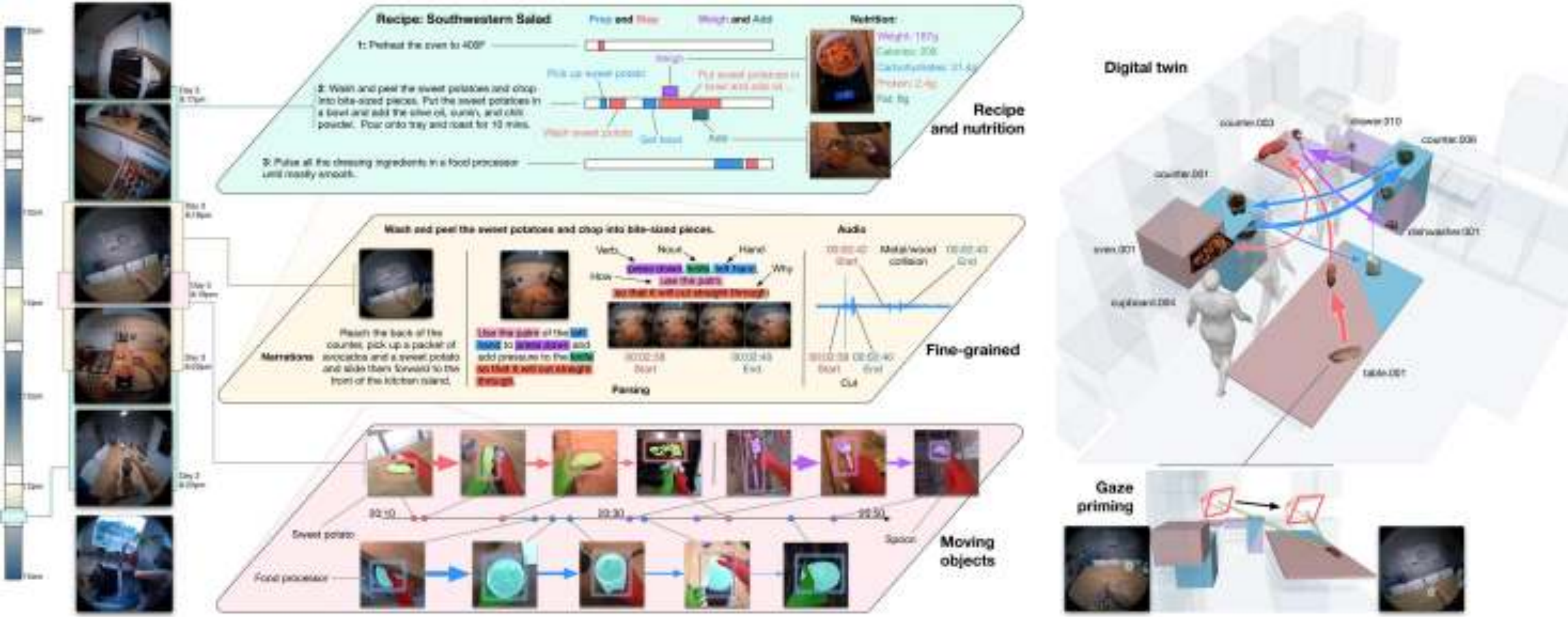
Digital Twin

Fixtures

Open drawer



HD-EPIC



Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

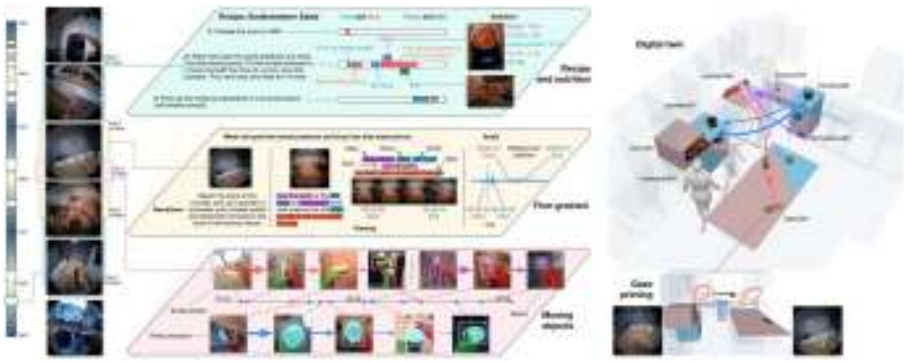
HD-EPIC

Annotation Type	Total annotations	Annotations/min
Narrations	59,454	24.0
Parsing (Verbs + Nouns + Hands + How + Why)	303,968	122.7
Recipes (Preps + Steps)	4,052	1.6
Sound	50,968	20.6
Action boundaries	59,454	24.0
Object Motion (Pick up + Put down + Fixtures + Bboxes + Masks)	153,480	62.0
Object Itinerary	4,881	2.0
Object Priming (Starts + Ends)	18,264	7.4
Total		263.2

Table A3. HD-EPIC annotations per minute

Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

HD-EPIC



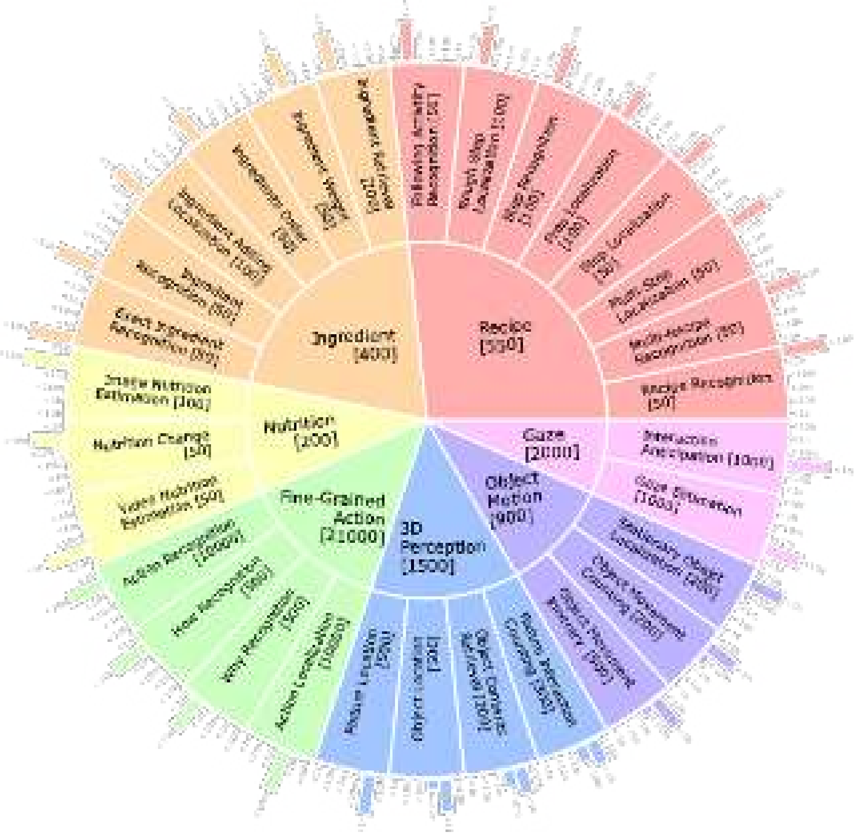
Sec 1: Highly-Detailed Dataset

Sec 2: HD-EPIC VQA Benchmark

Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

HD-EPIC

1. Recipe. Questions on temporally localising, retrieving, or recognising recipes and their steps.
2. Ingredient. Questions on the ingredients used, their weight, their adding time and order.
3. Nutrition. Questions on nutrition of ingredients and nutritional changes as ingredients are added to recipes.
4. Fine-grained action. What, how, and why of actions and their temporal localisation.
5. 3D perception. Questions that require the understanding of relative positions of objects in the 3D scene.
6. Object motion. Questions on where, when and how many times objects are moved across long videos.
7. Gaze. Questions on estimating the fixation on large landmarks and anticipating future object interactions.



HD-EPIC

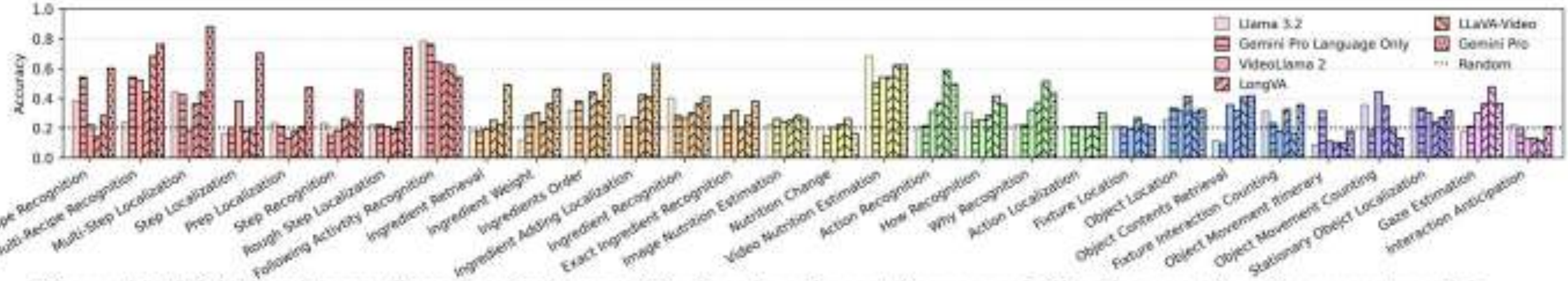


Figure 11. **VQA Results per Question Prototype**. Our benchmark contains many challenging questions for current models.

Model	Recipe	Ingredient	Nutrition	Action	3D	Motion	Gaze	Avg.
Blind - Language Only								
Llama 3.2	33.5	25.0	36.7	23.3	22.3	25.5	19.5	26.5
Gemini Pro	38.0	26.8	30.0	22.1	21.5	27.7	20.5	26.7
Video-Language								
VideoLlama 2	30.8	25.7	32.7	27.2	25.7	28.5	21.2	27.4
LongVA	29.6	30.8	33.7	30.7	32.9	22.7	24.5	29.3
LLaVA-Video	36.3	33.5	38.7	43.0	27.3	18.9	29.3	32.4
Gemini Pro	64.3	48.6	34.7	39.6	32.5	20.8	28.7	38.5
Sample Human Baseline	*96.7*	*96.7*	*85.0*	*92.5*	*93.8*	*92.7*	*75.0*	*90.3*

University of BRISTOL

Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

HD-EPIC

Which of these sentences best describe the action(s) in the video? [00:03:56 - 00:04:03]

- **A.** Wash the cutting board using the sponge in right hand, then, rotate the cutting board so that the back side can be washed
- **B.** With sponge in right hand, clean cutting board while holding board steady with left hand, then with left hand put cutting board under water to clean from soap
- **C.** With left hand, grab cutting board from dish rack, then, with both hands put cutting board down on kitchen counter
- **D.** With my left hand, pick up cutting board, then, with both hands, run the cutting board under water to clean
- **E.** Pick up cutting board from drying rack using right hand, then, dry the cutting board using tea towel in left hand whilst flipping and rotating the cutting board with right hand

HD-EPIC

What is the best description for how the person carried out the action pick up bowl of coconut milk in this video segment? [00:18:44 - 00:18:46]

- **A.** Using both hands holding the bowl from bowl rim.
- **B.** By holding both sides using the oven gloves.
- **C.** using the right hand and lift the large white bowl up.
- **D.** using left hand and removing the fork used to stir it using right hand.
- **E.** using both hands from the kitchen top above the dishwasher.

Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

HD-EPIC

How many times did I open the item at bounding box (165, 452, 1408, 1408) in 00:00:57?

A. 3 **B.** 1 **C.** 4 **D.** 5 **E.** 2

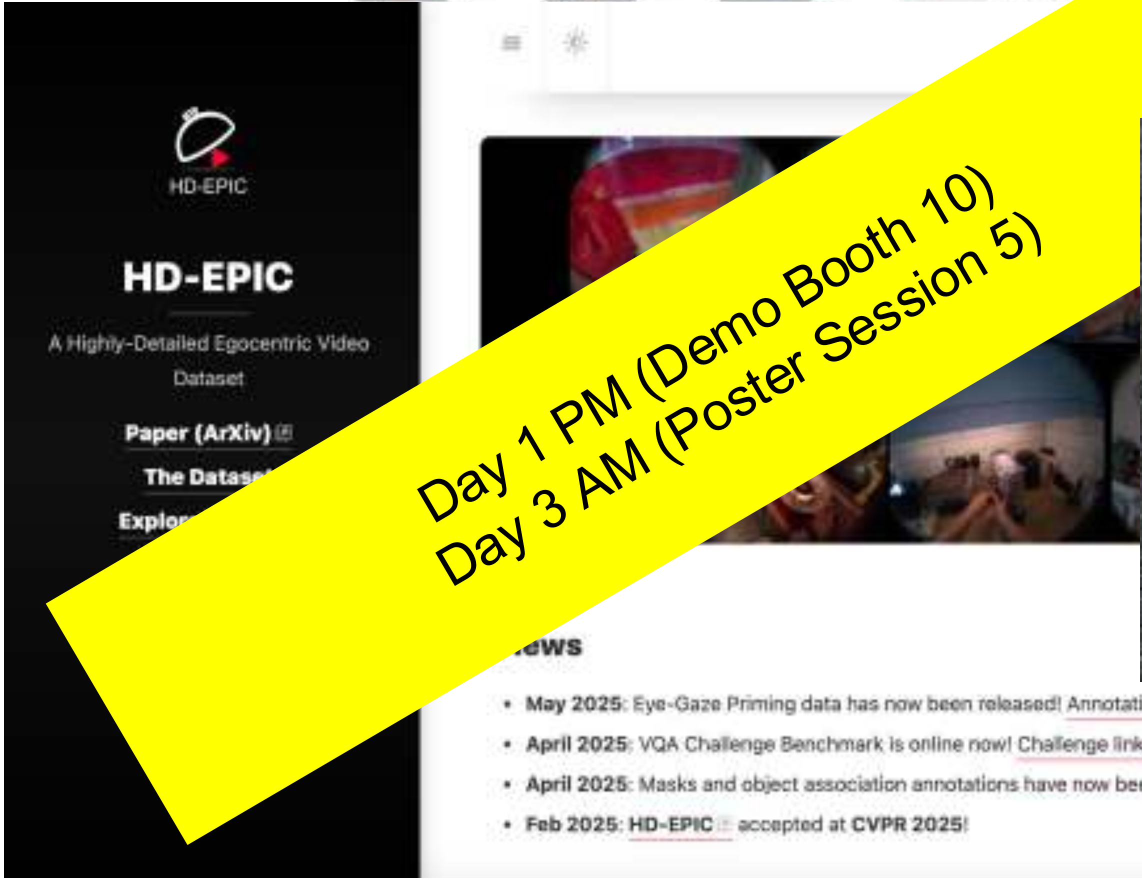
HD-EPIC

University of Bristol

Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.

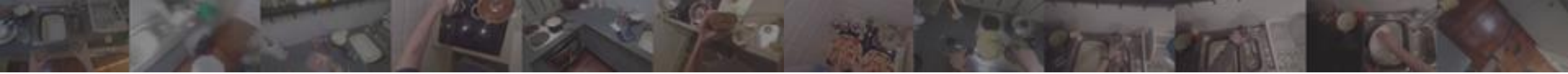
Dima Damen
HVU - CVPRW'25

HD-EPIC



Day 1 PM (Demo Booth 10)
Day 3 AM (Poster Session 5)

Perrett et al (2025). HD-EPIC: A Highly-Detailed Egocentric Video Dataset. CVPR.



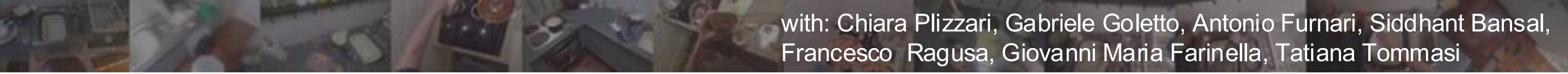
An Outlook into the Future of Egocentric Vision

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

Politecnico di Torino

University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA



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

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

How did we do this?

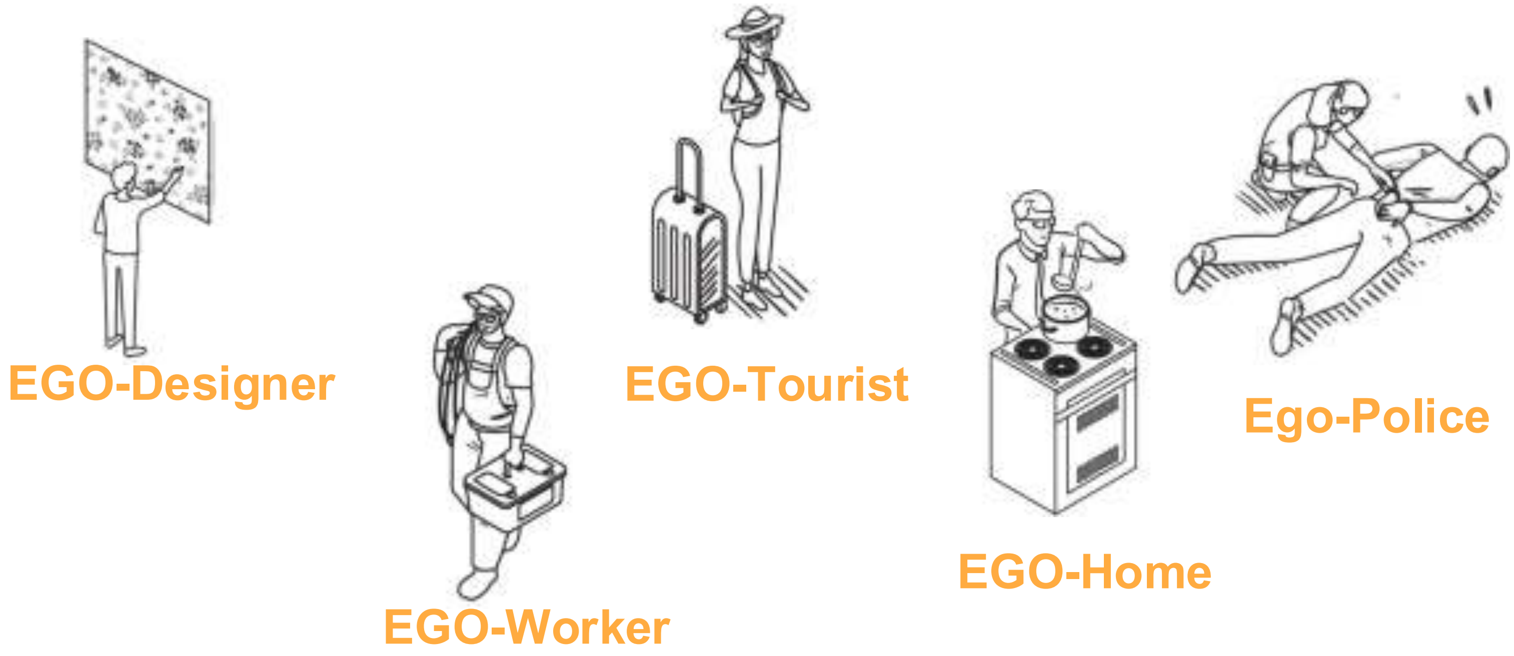
University of BRISTOL

Dima Damen
HVU - CVPRW'25

C Plizzari et al (2025). An Outlook into the Future of Egocentric Vision. IJCV

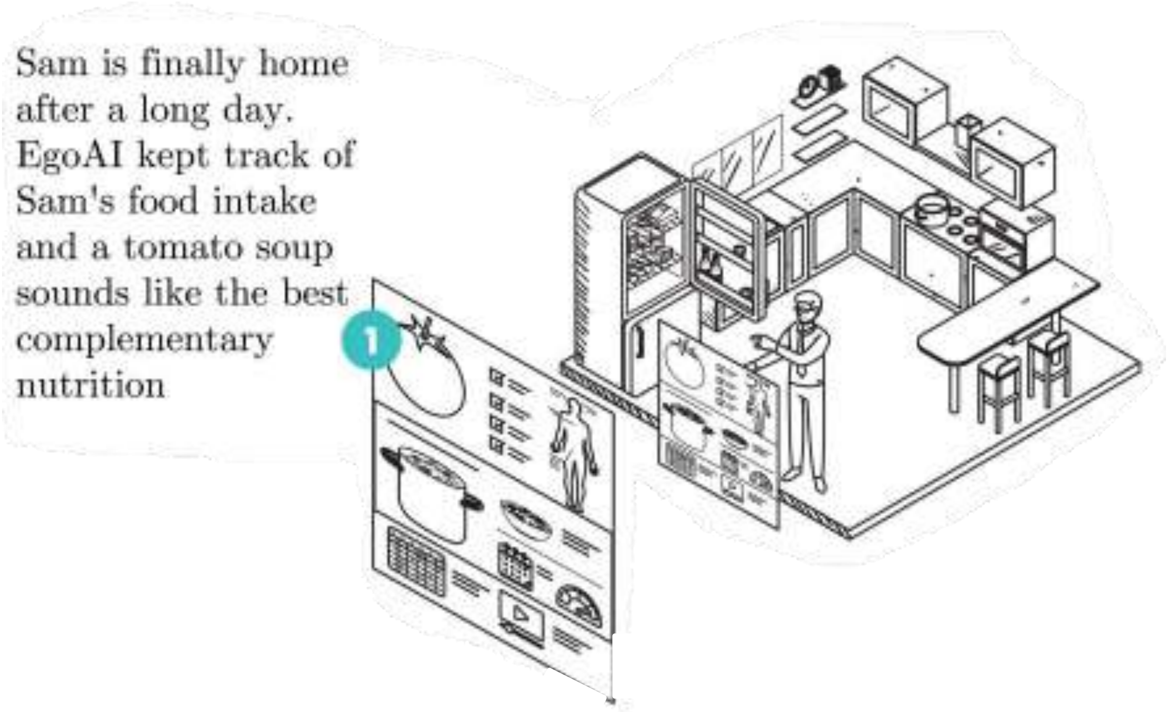
with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi

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



EGO-Home

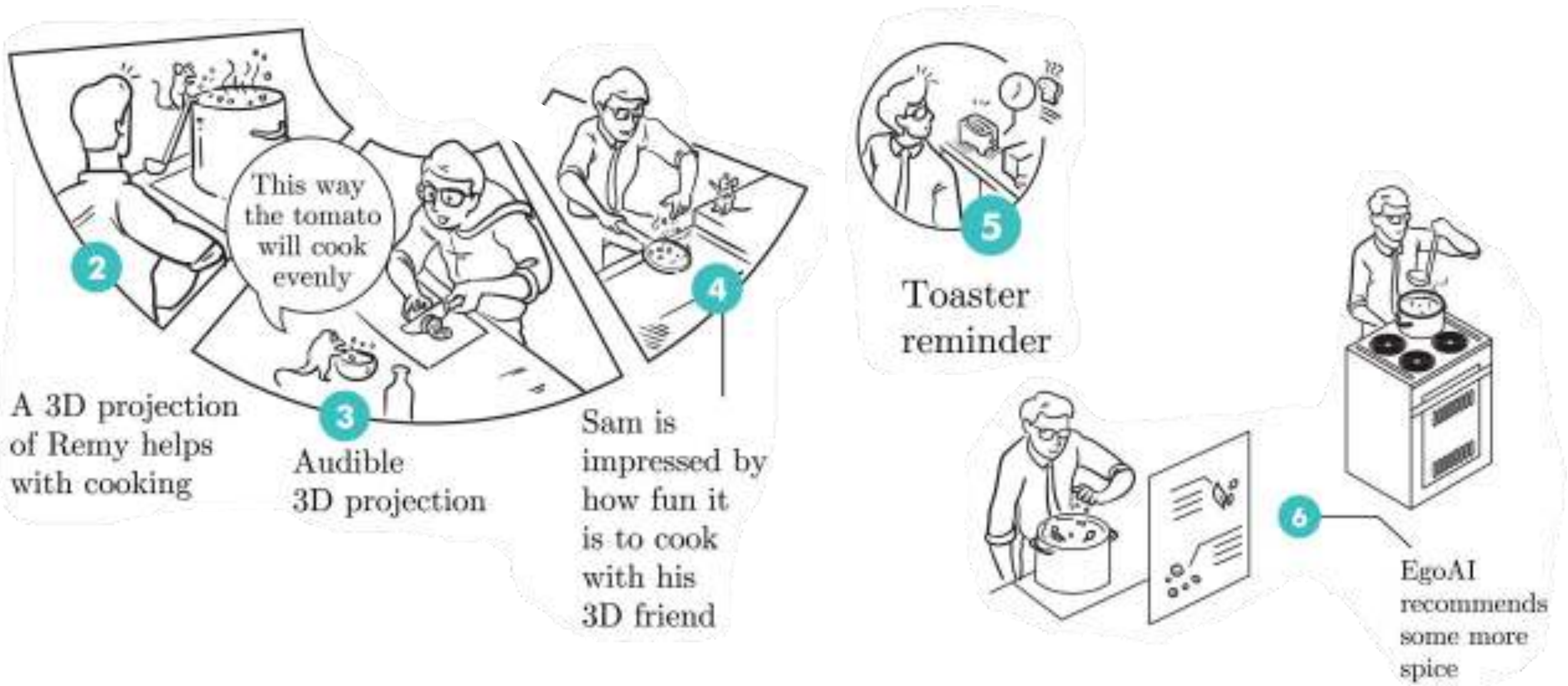
with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi



C Plizzari et al (2025). An Outlook into the Future of Egocentric Vision. IJCV

EGO-Home

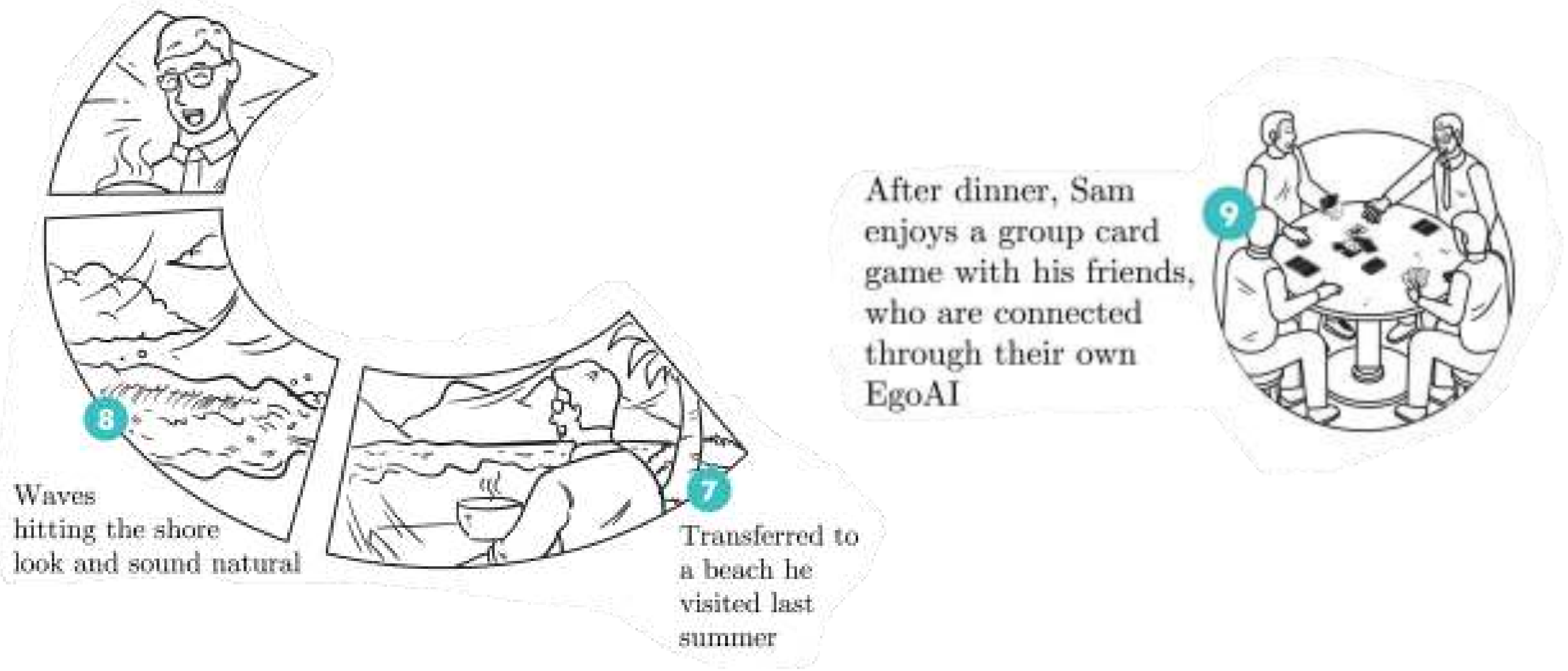
with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi



C Plizzari et al (2025). An Outlook into the Future of Egocentric Vision. IJCV

EGO-Home

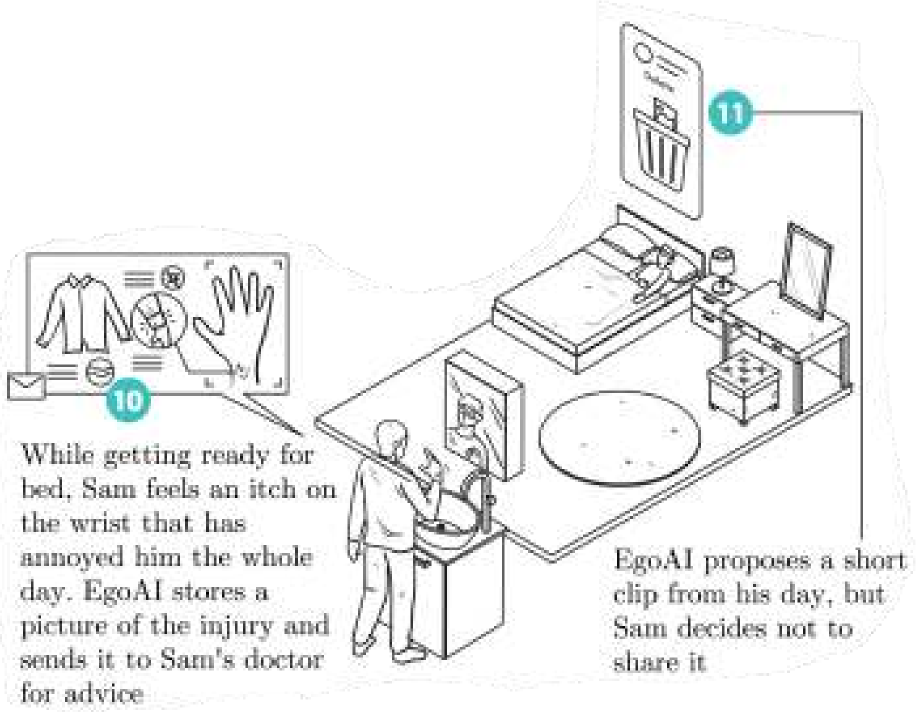
with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi



C Plizzari et al (2025). An Outlook into the Future of Egocentric Vision. IJCV

EGO-Home

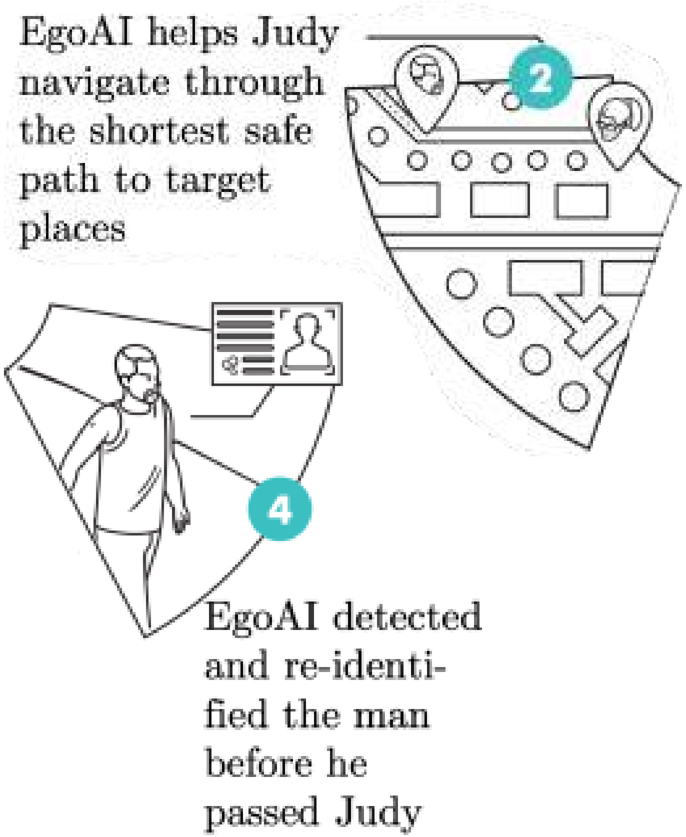
with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi



C Plizzari et al (2025). An Outlook into the Future of Egocentric Vision. IJCV

From Stories to Tasks

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



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

In this talk…

Unique Captioning

Visual Instructions

Learning from
Continuous Streams

Out of Sight,
Not Out of Mind

HD-EPIC

HD-EPIC: A Highly-Detailed Egocentric Video Dataset

University of BRISTOL

The Team

Thank you

For further info, datasets, code, publications…

http://dimadamen.github.io

@dimadamen

@dimadamen.bsky.social

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

Q&A

University of BRISTOL