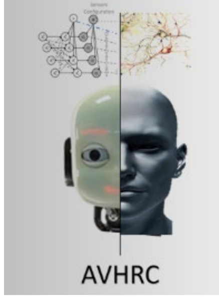
SAPIENZA
UNIVERSITÀ DI ROMA
ALCOR Lab



Perception, activities and sustained attention when robots help humans

AVHRC

Fiora Pirri

Summary

- We shall first present experiments in real world with a high performing robot.

  - Here attention is pursued by searching the environment, which builds the local robot knowledge.

  - We will see the beauty and limitations of the experiments.

- Then we will talk about a new attention model for activity recognition, in particular we will talk about *sustained attention*, which we applied to the Kinetics 700 dataset.

Preamble

Deep learning so far, has shown to be able to manage excellently low level features, and low-level salient features selection, for a huge amount of tasks.

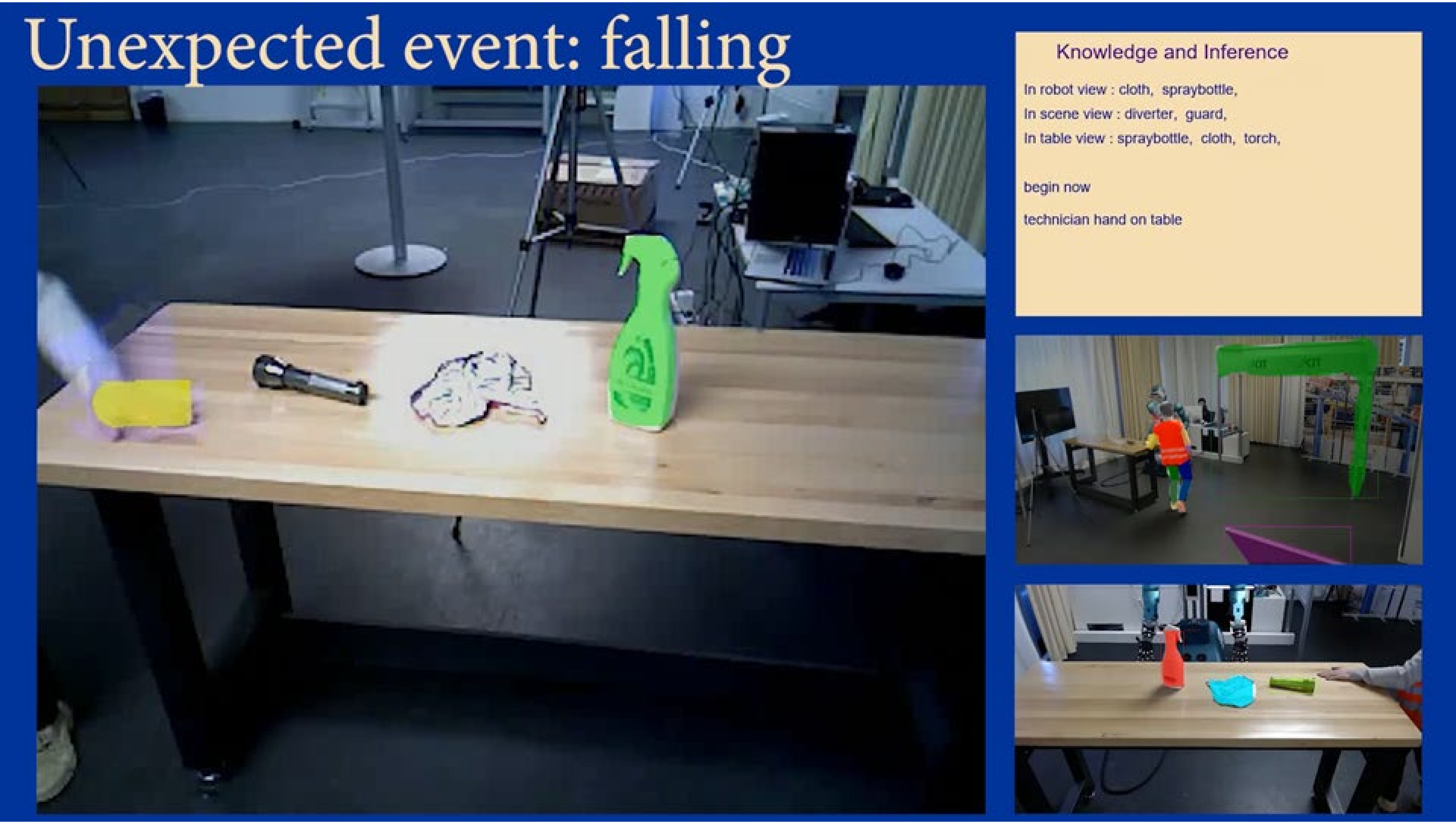
Many approaches to action and activity recognition have explored spatio-temporal attention with 3D kernel based networks such as I3D and R(2+1)D.
Yet the features considered are low level and saliency based.

In most collaboration tasks – in particular helping humans – a robot has to keep focused on a task for a long time.

There are cognitive functions requiring to be activated by an attention network:
-- attending long term tasks
-- making preferences
-- searching
-- identifying relations

These processes require sustained attention, a form of attention inferring relations in time and space, to keep attention focused on the task.

VIDEO



Unexpected event: falling

Knowledge and Inference

In robot view : cloth, spraybottle,

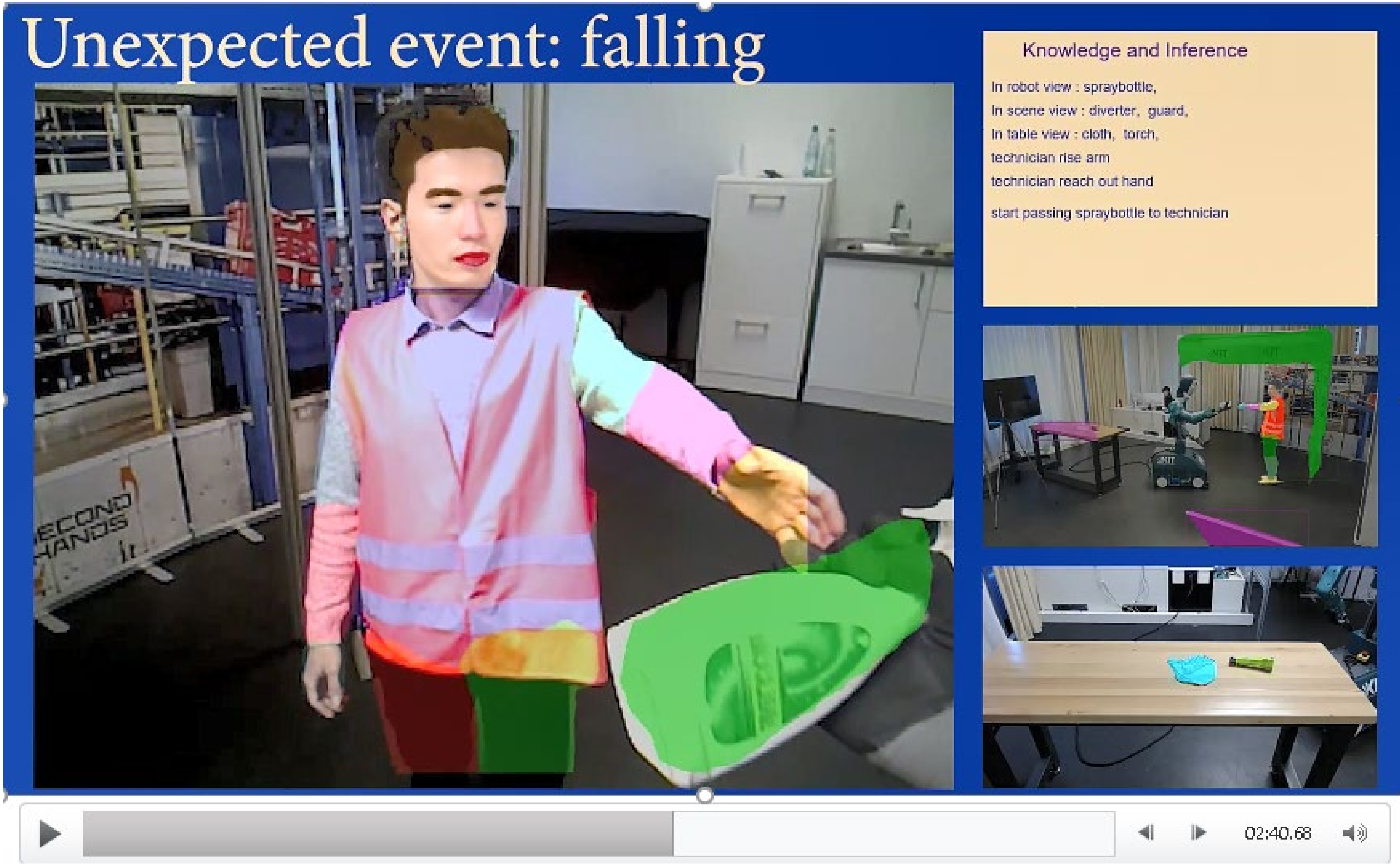
In scene view : diverter, guard,

In table view : spraybottle, cloth, torch,

begin now

technician hand on table

VIDEO



Unexpected event: falling

Knowledge and Inference

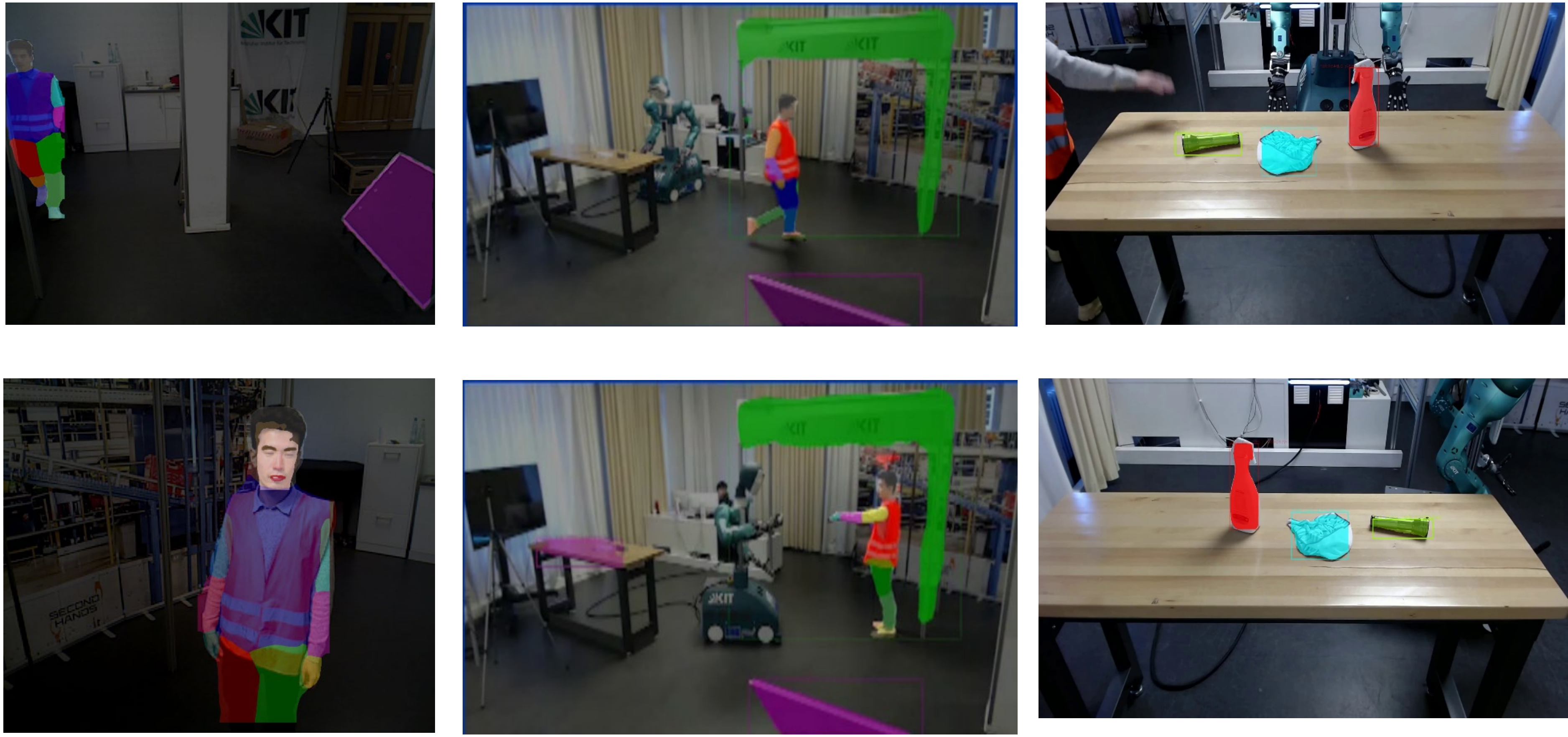
In robot view : spraybottle,
In scene view : diverter, guard,
In table view : cloth, torch,
technician rise arm
technician reach out hand

start passing spraybottle to technician

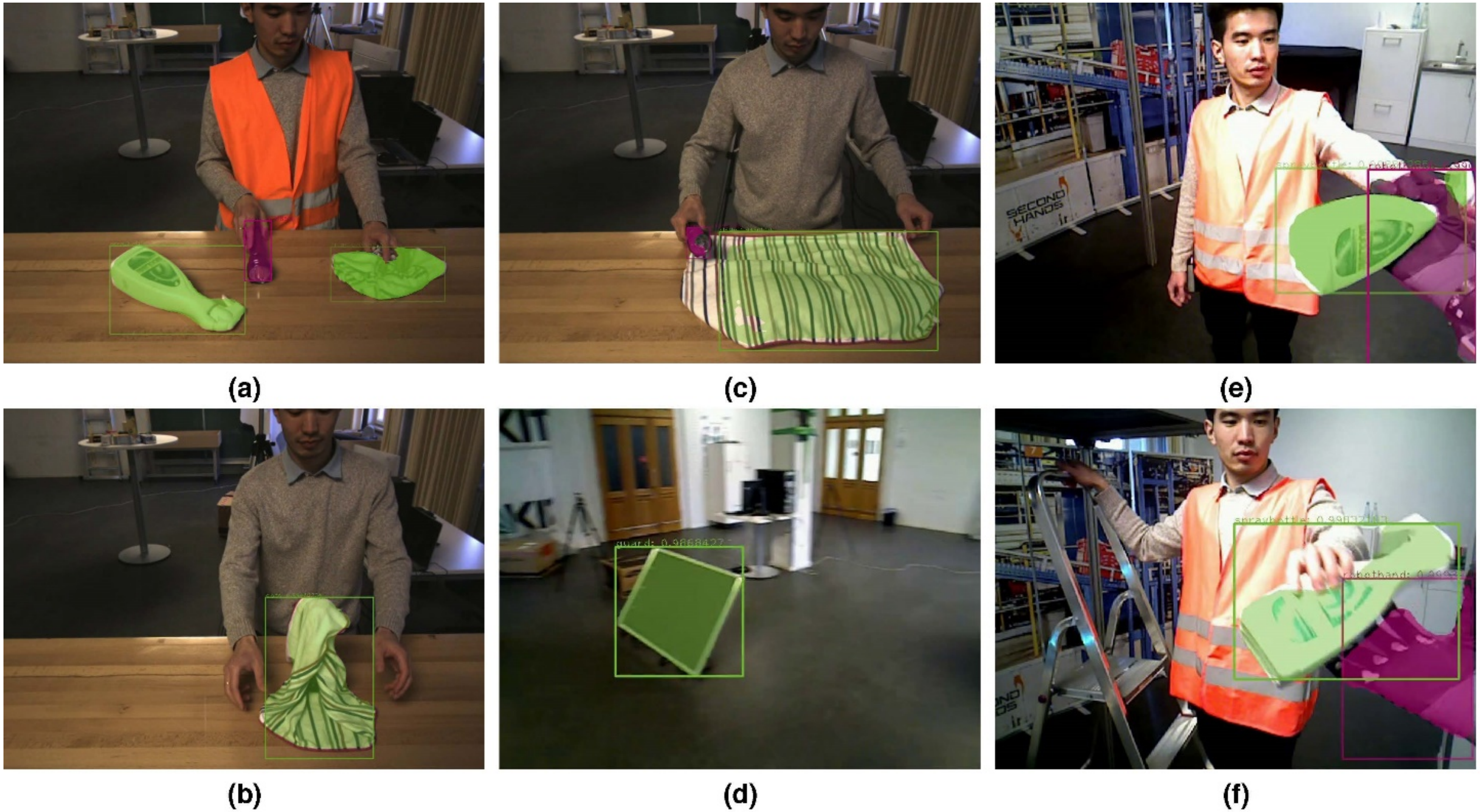
Model of inference and knowledge

Requirements:

3 views: scene, table, robot

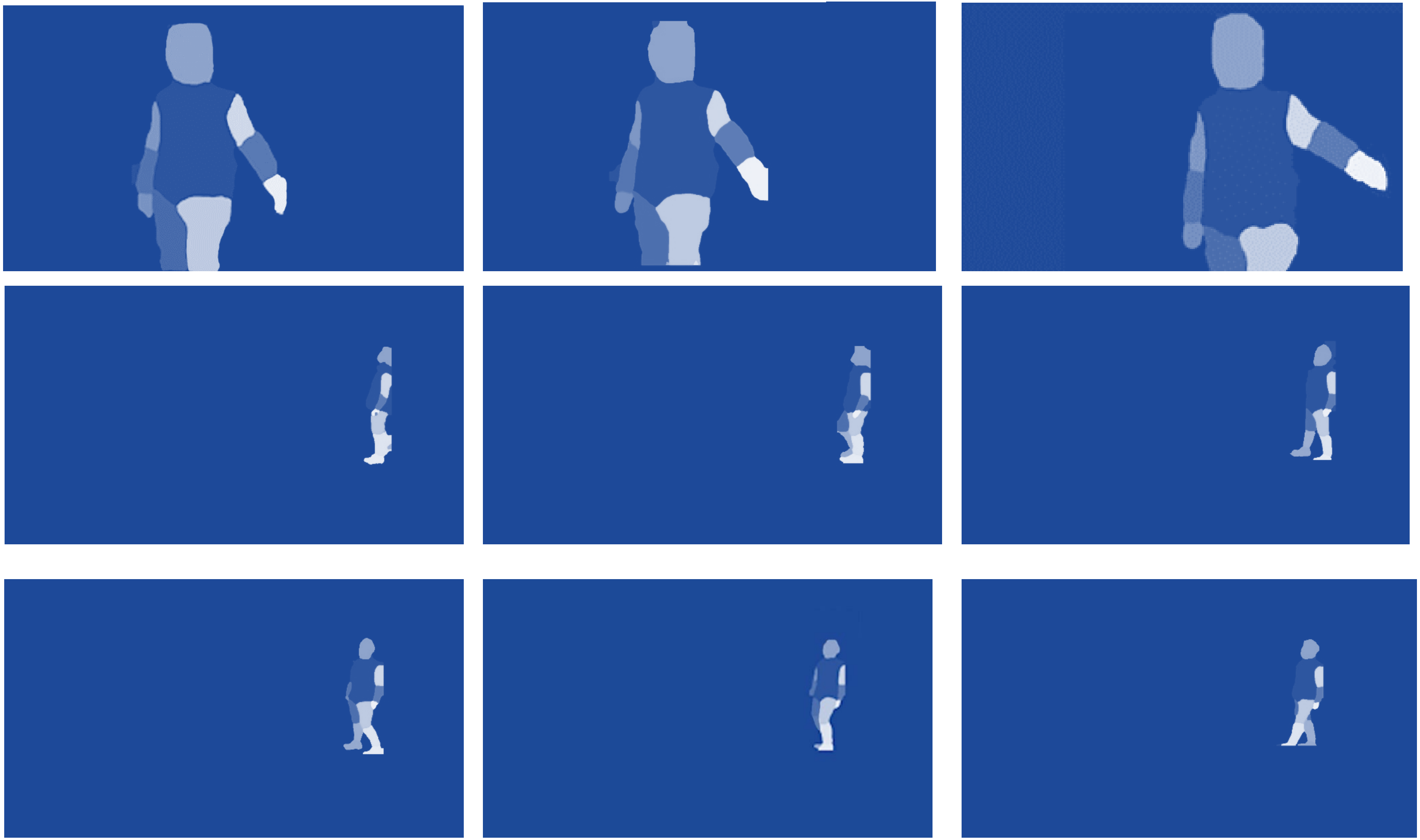


Combine the three views to obtain relations, locations, and basic knowledge.

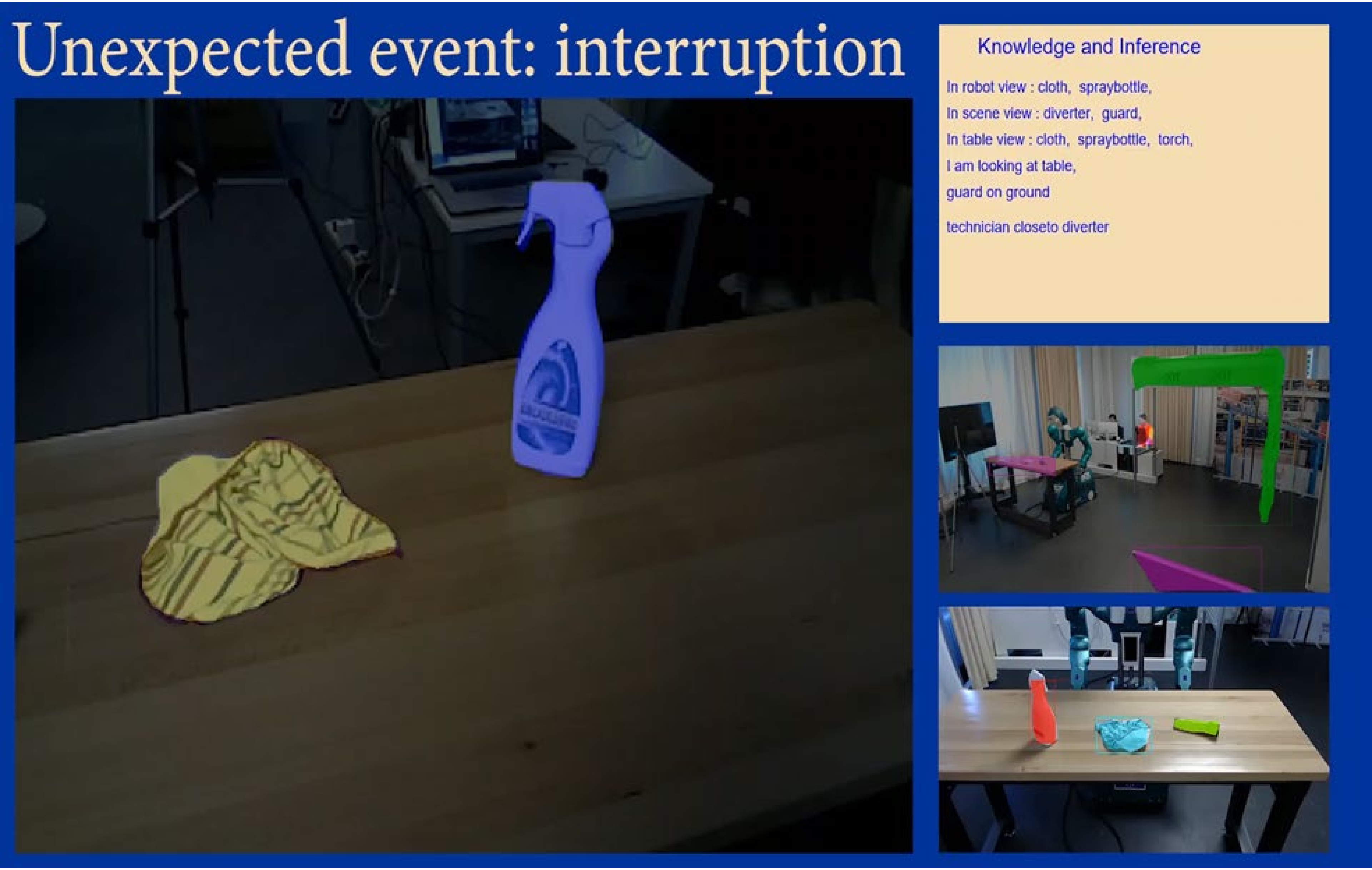


(a) (c) (e)

(b) (d) (f)



VIDEO



Unexpected event: interruption

Knowledge and Inference

In robot view : cloth, spraybottle,

In scene view : diverter, guard,

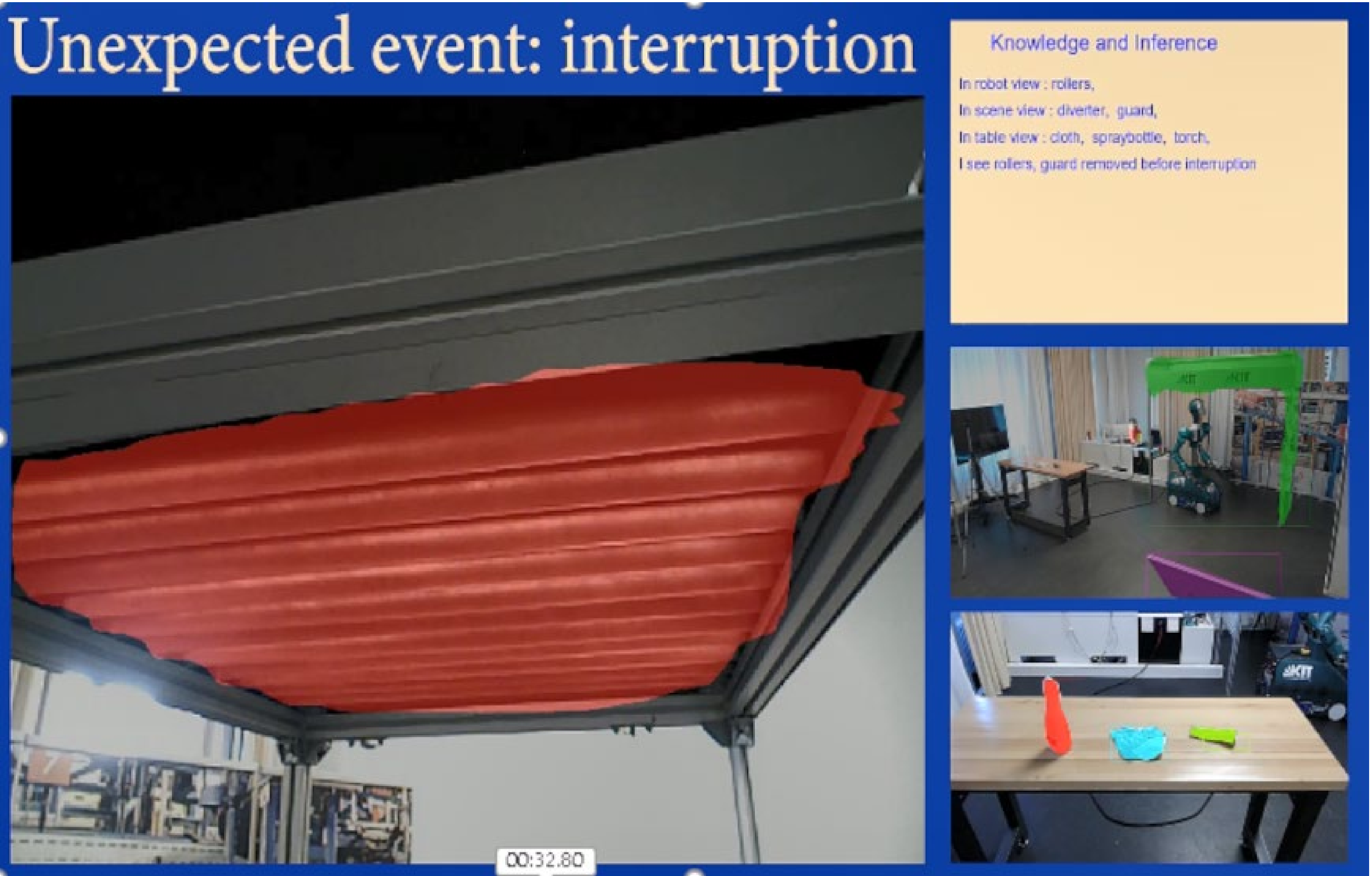
In table view : cloth, spraybottle, torch,

I am looking at table,

guard on ground

technician closeto diverter

VIDEO



Unexpected event: interruption

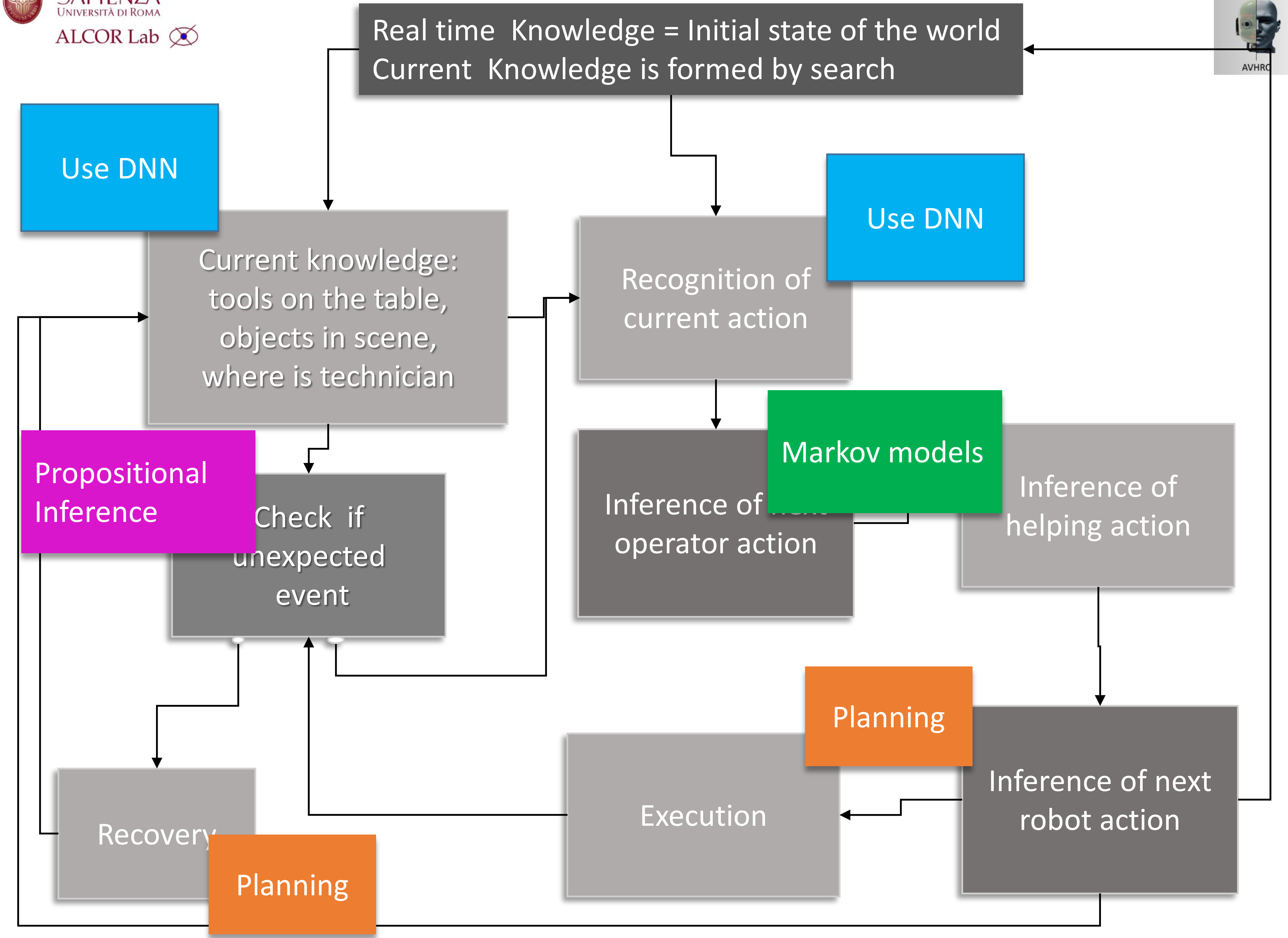
Knowledge and Inference

In robot view : rollers,

In scene view : diverter, guard,

In table view : cloth, spraybottle, torch,

I see rollers, guard removed before interruption



Real time Knowledge = Initial state of the world
Current Knowledge is formed by search

Use DNN

Current knowledge: tools on the table, objects in scene, where is technician

Recognition of current action

Use DNN

Propositional Inference

Check if unexpected event

Inference of next operator action

Markov models

Inference of helping action

Planning

Inference of next robot action

Execution

Recovery

Planning

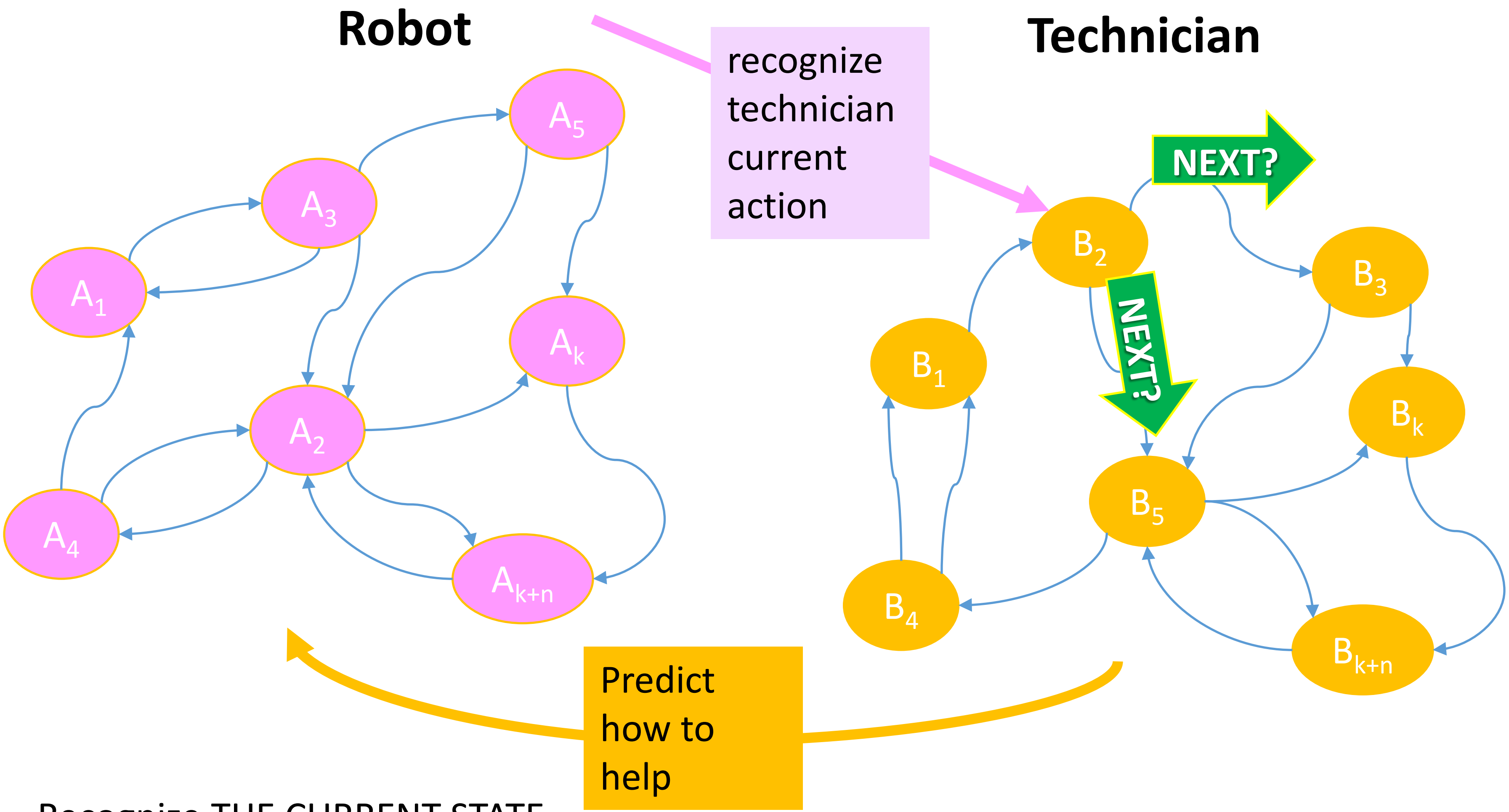
Inference and knowledge

The model is based on a combination of processes that are learned off line and the online application of them is triggered purely by visual perception.

The processes are: current action recognition, next action forecasting, searching, objects continuity, choice of help action

The action recognition model tell us about the current action, and the object recognition model provides an accurate recognition of items in the scene.

The next actions form the hidden states of the scene configuration.



Robot

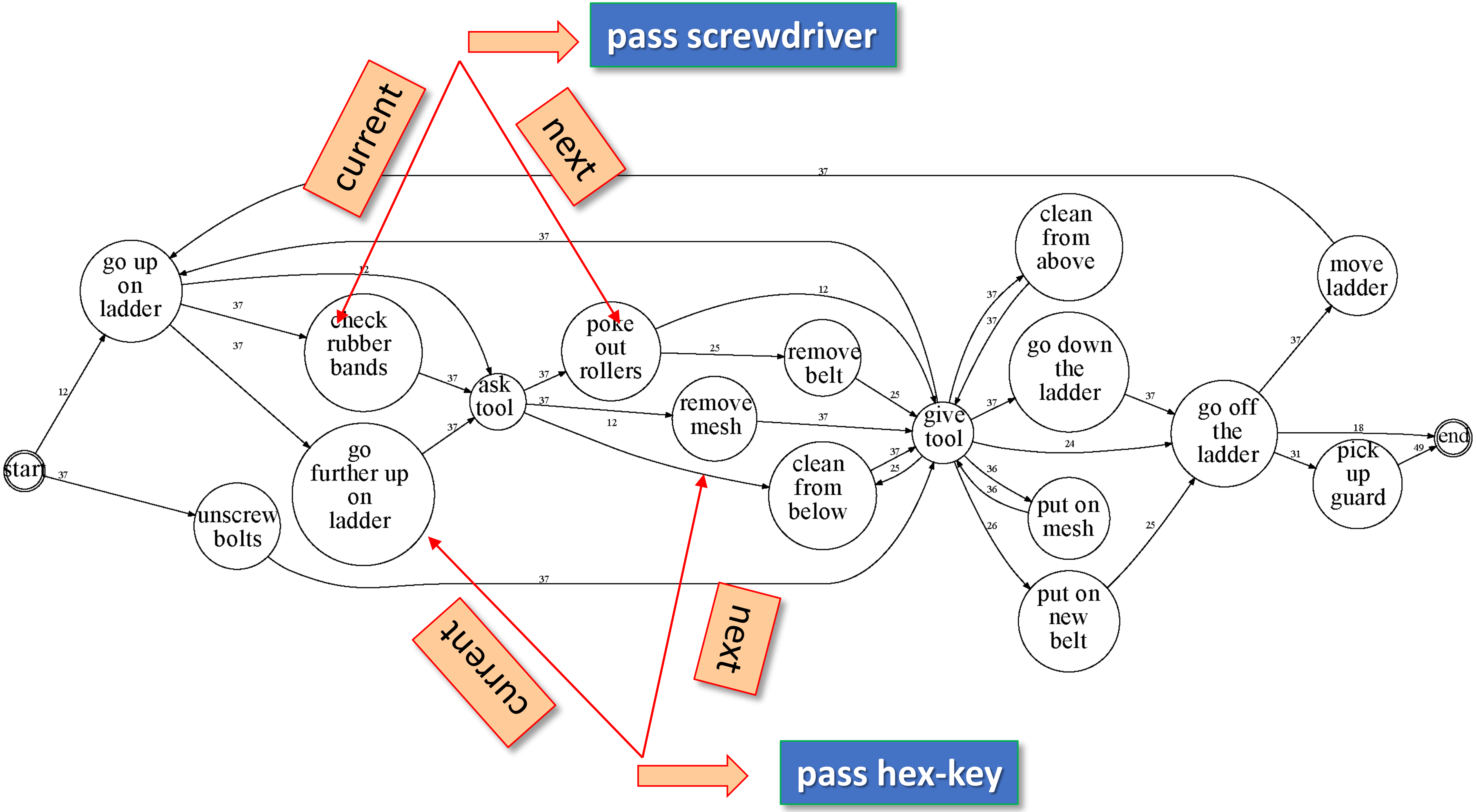
Technician

recognize technician current action

Predict how to help

Recognize THE CURRENT STATE
Recognize current operator action
Predicts next operation action
Infer Helping action

Learning trajectory patterns



pass screwdriver

current

next

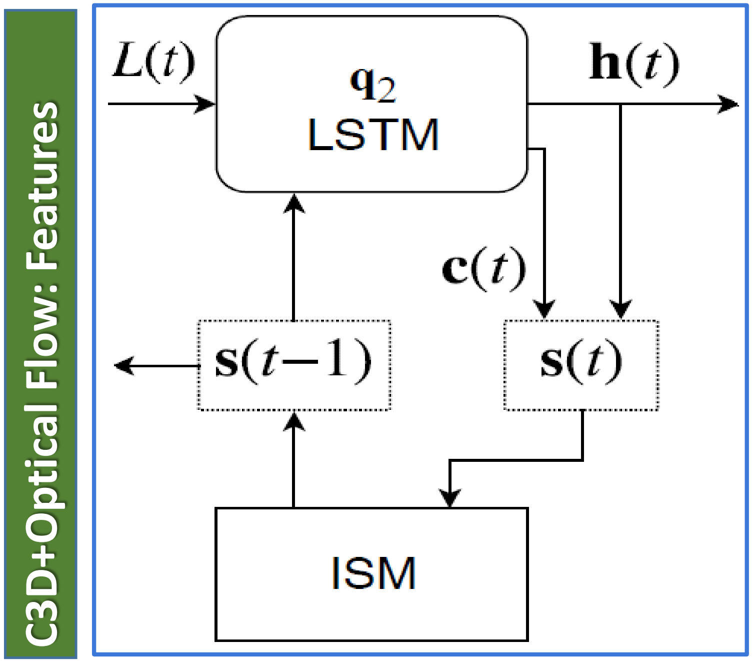
current

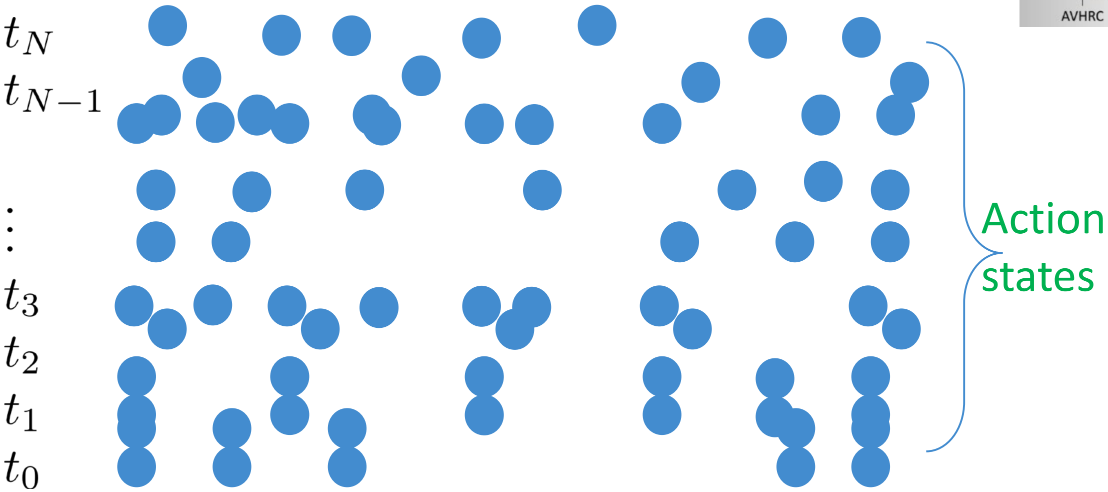
next

pass hex-key

Forecasting: previous method

$$\mathbf{s}(t) = (\mathbf{h}(t)^\top, \mathbf{c}(t)^\top)^\top, t \in [0, T]$$





When a sequence of action is given in a succession it is considered an activity.
In order not to have limit in the dimension of the sequence the INTERNAL MEMORY updates the specific action state each time a frame of the action is observed (in training)

$$v = \begin{Bmatrix} s(0) \\ . \\ . \\ . \\ s(T) \end{Bmatrix}$$

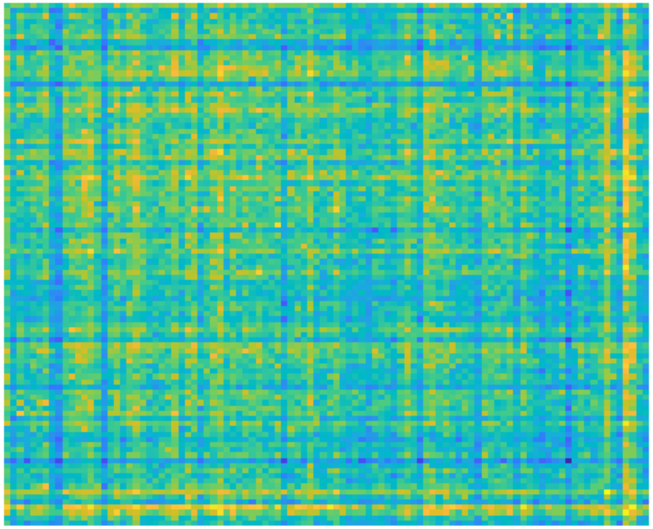
$$s_t = \begin{cases} \emptyset & \text{if the stack is empty} \\ \varphi(s_{t-1}) & \text{if the previous state is in the stack} \end{cases}$$

Management of the backpropagation is similar to the truncated temporal backpropagation

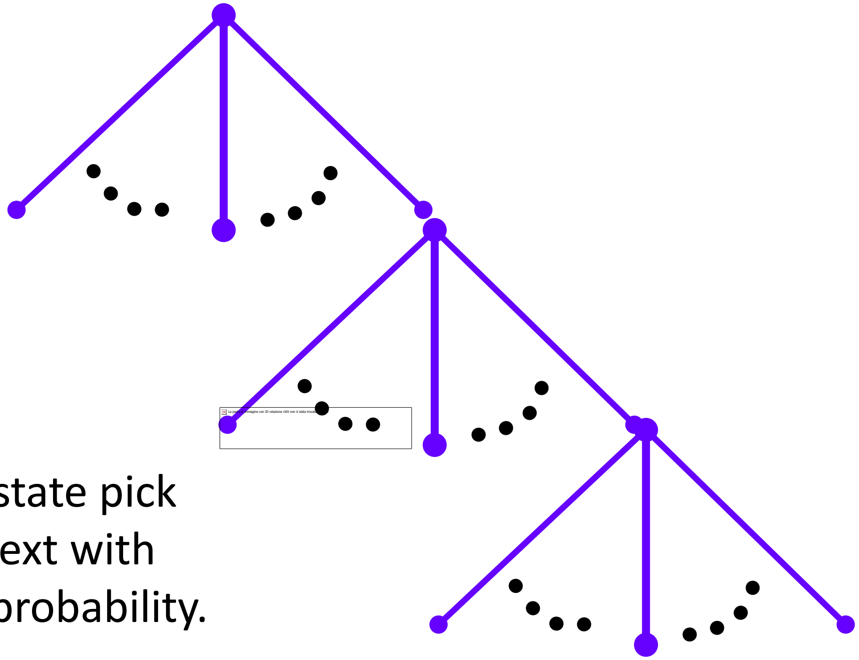
In most methods the probability of moving from one state to another is defined by a stochastic transition matrix,

$$\sum_i row_i = 1$$

$$\tau_{i,j} = p(x_{t+1} = j \mid x_t = i) = \frac{n_{i,j}}{\sum_j n_{i,j}}$$

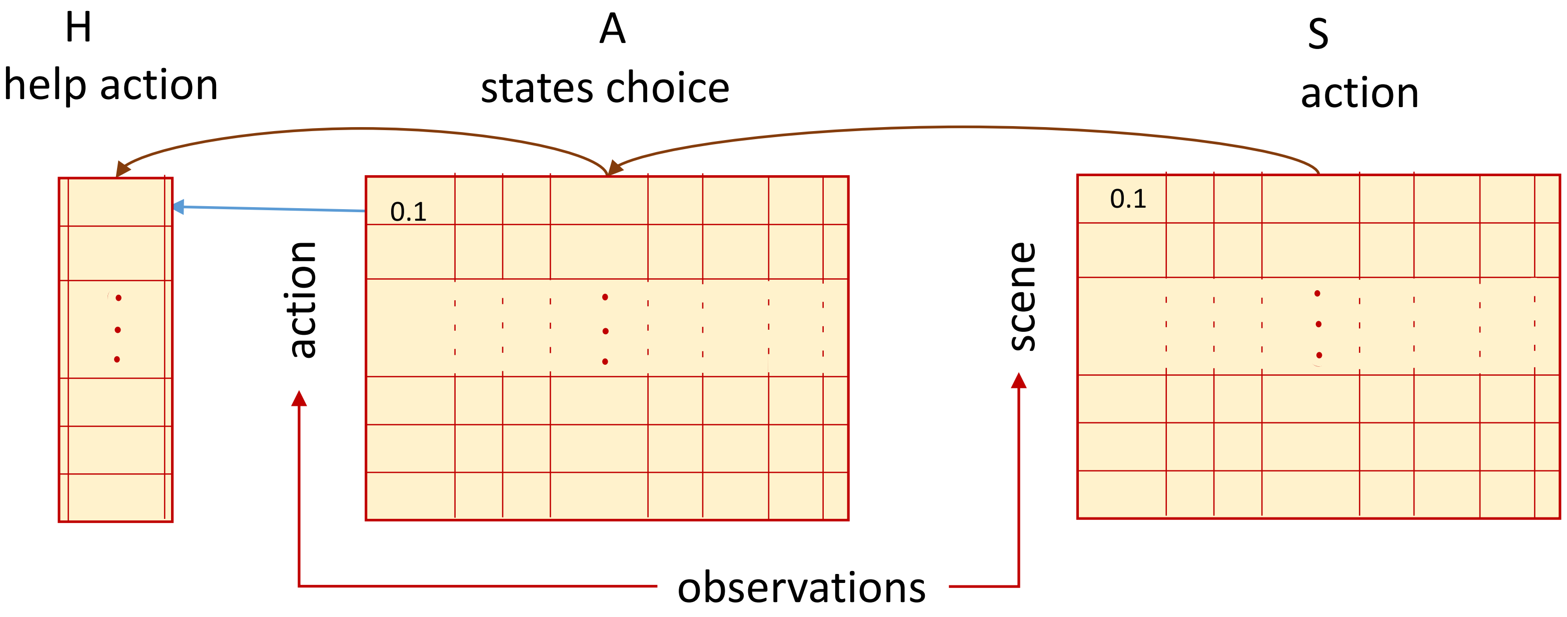


At each state pick up the next with highest probability.



$$p(x_0 = a_0, \ldots x_n = a_n) = p(x_0 = a_0) \prod_j p(x_{i+1} = a_{i+1} \mid x_i = a_i)$$

Model of inference and knowledge:



A scene state explains an action
Current action anticipates next possible actions
Next possible action induces a possible request of help.

Initial states computed during search: π_A, π_T

Search requires attention

$$(A_k, S_t, H_m)$$

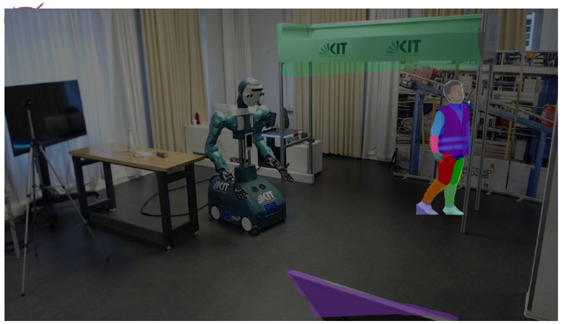
$$\pi_A = P(a_0 = A_{current}) \leftarrow \text{from AR}$$

$$\pi_T = P(b_0 = S_{current}) \leftarrow \text{from SR}$$

$$P(a_{j+1} = A_k) = \arg\max_k \{P(S_t \mid A_k)\}_{0\le k\le n}$$

$$P(h_{j+1} = H_m) = \arg\max_m \{P(H_m \mid A_k)\}_{0\le m\le r}$$

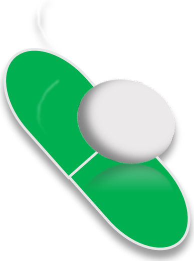
Example



move-under-diverter

Inspecting

Clean from below



Current state:
on_table: cloth, torch, knife, hex-key, spray-bottle, screwdriver,
on floor: guard
under-diverter: technician

inference: bring spray-bottle or torch to technician
(since guard has been removed because on floor)

Search

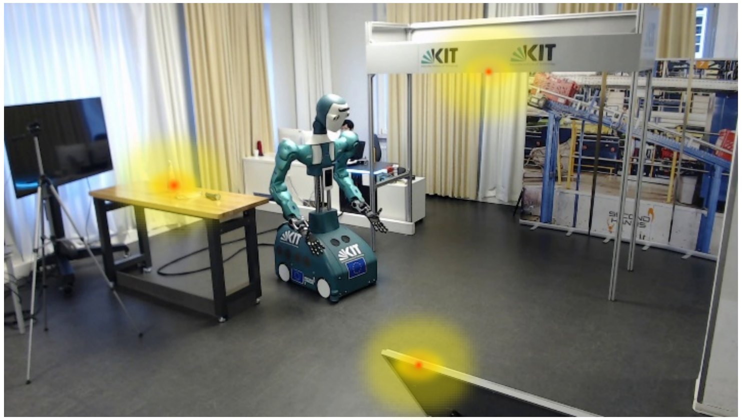
Search requires attention and ability to prioritize where to look at.

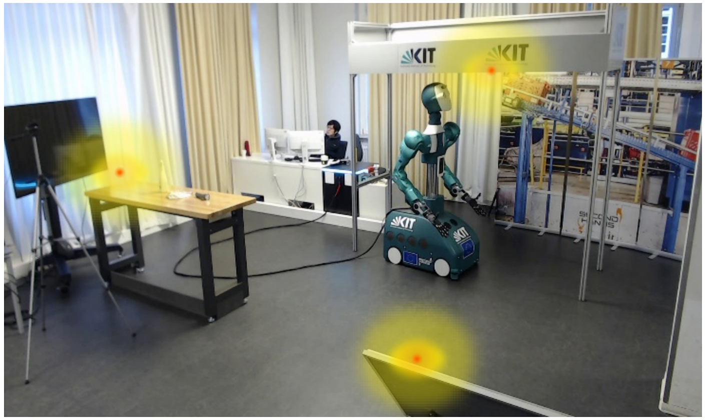
Fixed objects possible locations are known a priori (where could be the guard on the floor, where is the table, where are the rollers..).

Attention is directed towards known locations. needed by the robot to build its knowledge about the current state of the world.
During search all objects are accumulated in the memory, up to when the initial state is complete for both table and scene configuration.

Two cases:
- The set built from scene coincide with standard initial situation with probability π_A
- The set displays some unexpected initial situation.

This last case is *class 1 of unexpected events*.





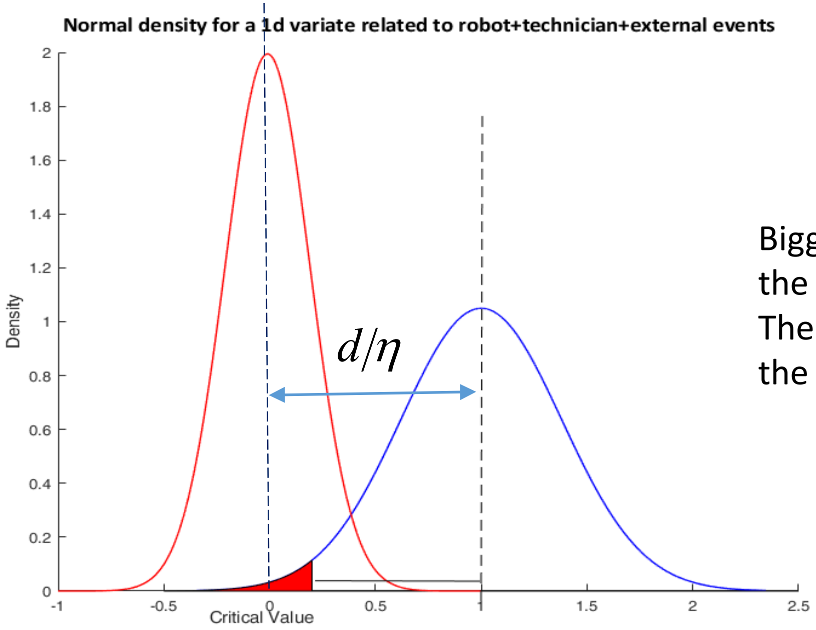
The two initial probability distributions π_A and π_T are computed as soon as the knowledge is filled and updated with new information up to the point is considered complete.

First simple view

ALL unexpected events are part of the sampling space of the events we consider.

{passing a tool, inspecting, dropping an object, loosing a tool, climbing the ladder, fall down the ladder, cleaning from below, cleaning from above,..}

In particular, unexpected events can be eventually observed: they can be observed like the actions of the technician or of the robot, in a specified limited space.



Bigger the distance d, more skewed/rare is the exception.
The multivariate mixture component of the event will be very tiny

$$d\,|\,\eta$$

Distance with prior η

Descriptive example, it would be multivariate

Steps forward

Limitations:

1. There is a quite limited number of actions, and exceptions have been defined to be in a language of known actions.

2. Interpretation of the scene amounts to few events occurring in almost known locations

3. The environment is mostly static.

Yet, understanding a scene requires a long and varied training to refine perceptual differentiation:

- a scene from the robot vantage point could be dark, affected by varying light, over/under exposed, at a low resolution, motion blurred. Events could be hundreds.
- to face the great diversity of events and unexpected events, training over a huge amount of actions – also with egocentric view – is paramount for understanding scenes.

What is attention when the task is prolonged over time and requires focus?

Kinetics 700

- More than 650K videos, millions of frames, 700 types of activities.
- Accuracy on Kinetics is not saturated like for other datasets.
- There are 20 classes with mAP less than 20%, tested under I3D.
- Kinetics is a double challenge:
  - The size of the dataset
  - The perceptual hardness, also for humans.

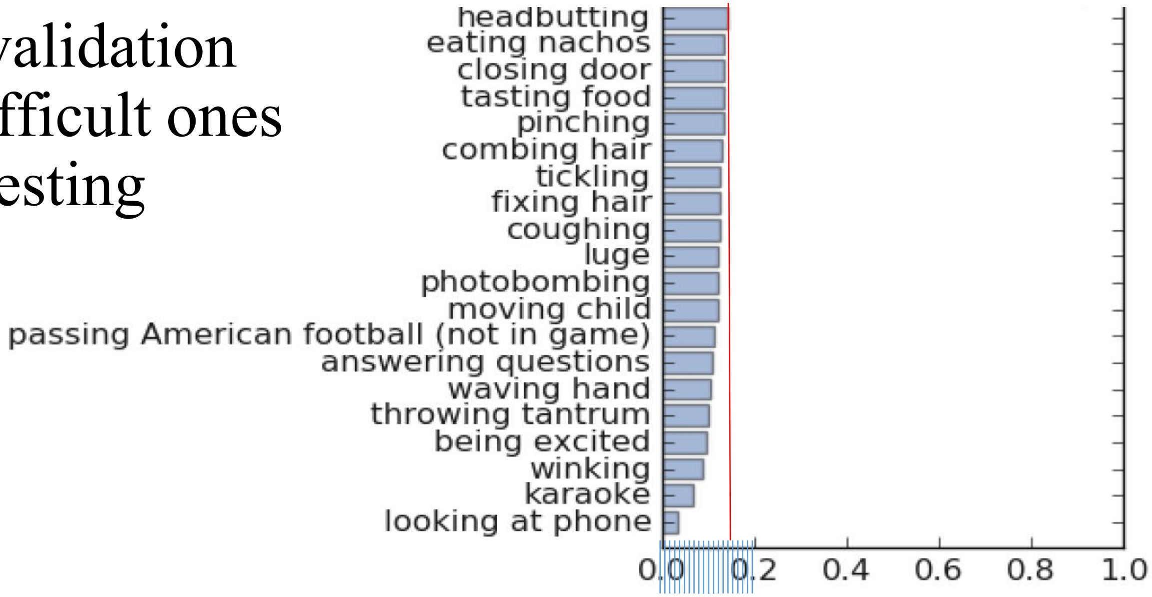
Kinetics 700, the main challenges:

Resources for computing: in Carreira & Zisserman 2019 they trained and tested the dataset with I3D, with 64 GPUs.

64 GPUs require the energy of a village. (Do not think CINECA has 64 GPUs)

Videos (300 frames for 10sec of video)
700 Classes,
545 318 Videos for training

105 000 videos for validation
20 classes are the difficult ones
These are very interesting



The interesting classes

Videos in these classes stimulate relevant cognitive factors which are not learned by the spatial-temporal models defined so far.

For example, there might be no objects to trigger the discovery of actions compatible with the object itself.

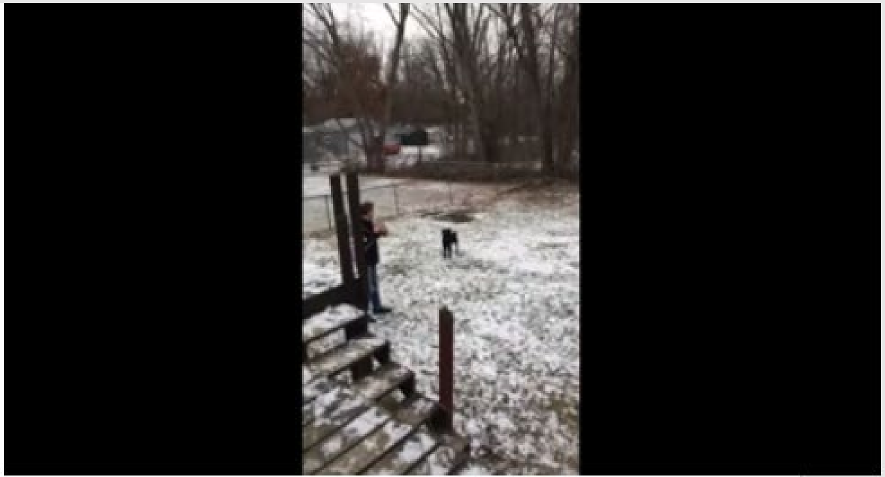
There are videos with hieratic subjects, impeding to prime the action according to pose or gestures.

Vary in the semantic content, in the recording conditions (locus and day-night), in the appearance properties, such as resolution, aspect ratio, contrast, brightness and camera motion.
In fact, Kinetics collects amateur videos, (in so differing from professional videos as in most datasets).

There are videos without people except for very few frames, and videos that jump from one subject to another, in so requiring a relational semantics along multiple tracks, as noted in (Ji et al. 2019).

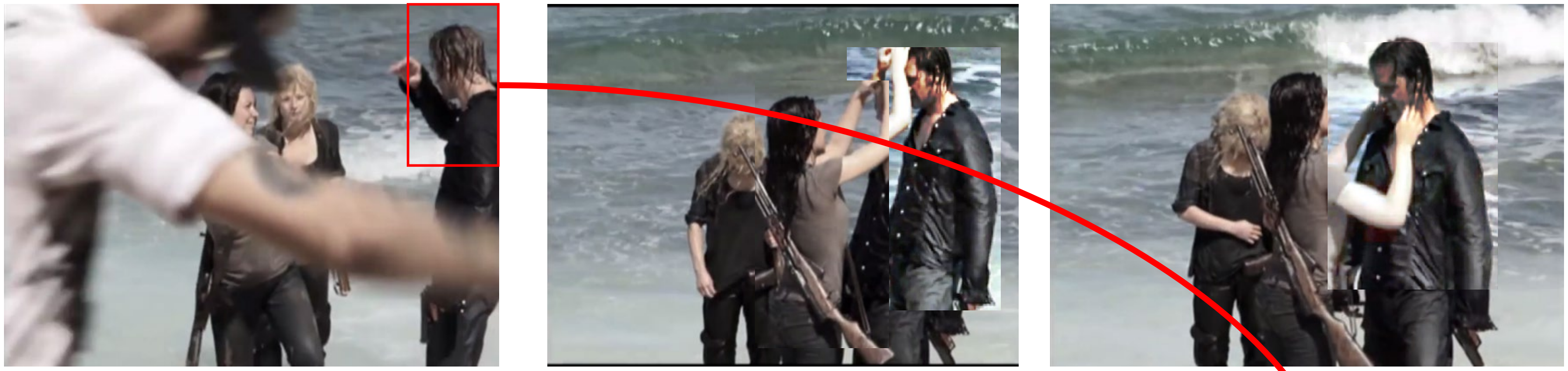
The typical saccadic mask of humans when rapidly turning their head is in these classes displayed by the camera random motion, which translates into significant motion blur. In many videos, the subjects scroll in windows where it is not possible to evaluate the direction of motion (the aperture problem, see also (Feichtenhofer et al. 2019)).

There are videos with very high variations in contrast and brightness, which require adaptive vision.

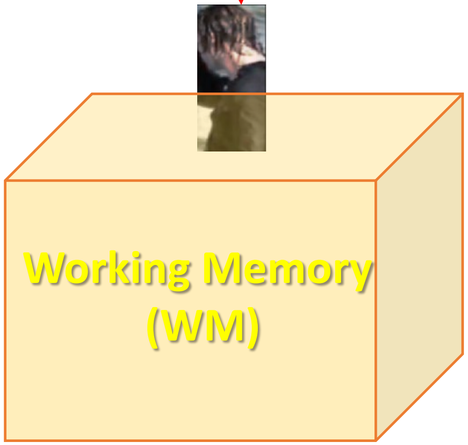


bielgame
Made with

Sustained attention



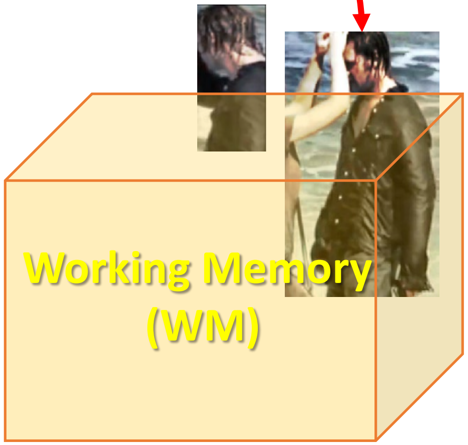
Sustained attention is the human ability to activate an attention network to understand relations amid snapshots and use the working memory to drive the gaze where to look at.



Sustained attention uses the *WM* to prioritize hot regions to build spatio-temporal relations.

Sustained attention

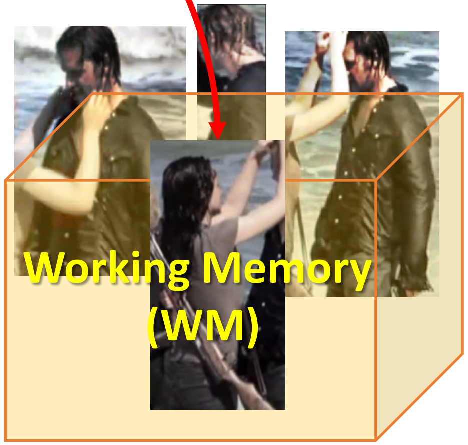
Sustained attention is the human ability to activate an attention network to understand relations amid snapshots and use the working memory to drive the gaze where to look at.



Sustained attention uses the *WM* to prioritize hot regions to build spatio-temporal relations.

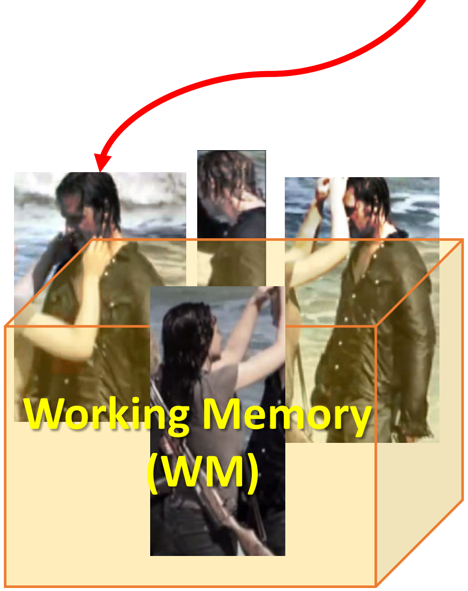
Sustained attention

Sustained attention is the human ability to activate an attention network to understand relations amid snapshots and use the working memory to drive the gaze where to look at.



Sustained attention uses the *WM* to prioritize hot regions to build spatio-temporal relations.

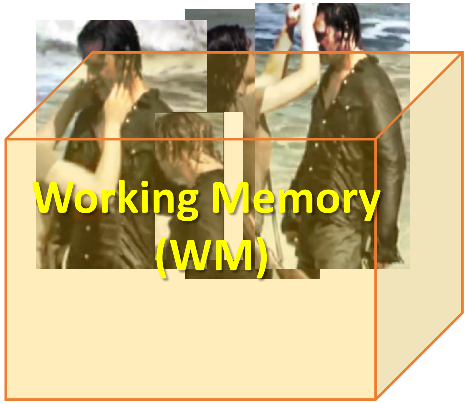
Sustained attention



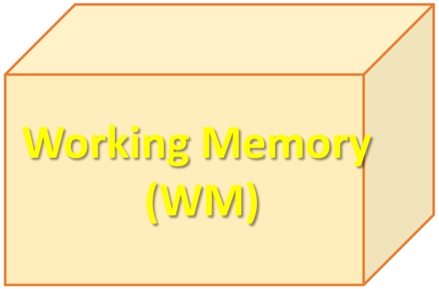
Sustained attention is the human ability to activate an attention network to understand relations amid snapshots and use the working memory to drive the gaze where to look at.

Sustained attention uses the *WM* to prioritize hot regions to build spatio-temporal relations.

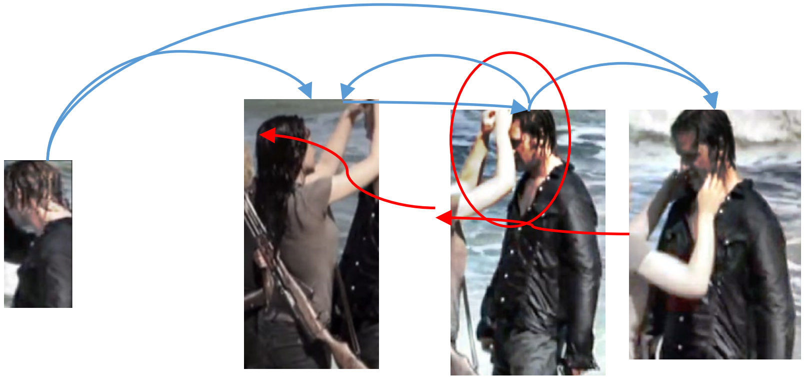
Sustained attention uses the *WM* to prioritize hot regions to build spatio-temporal relations.

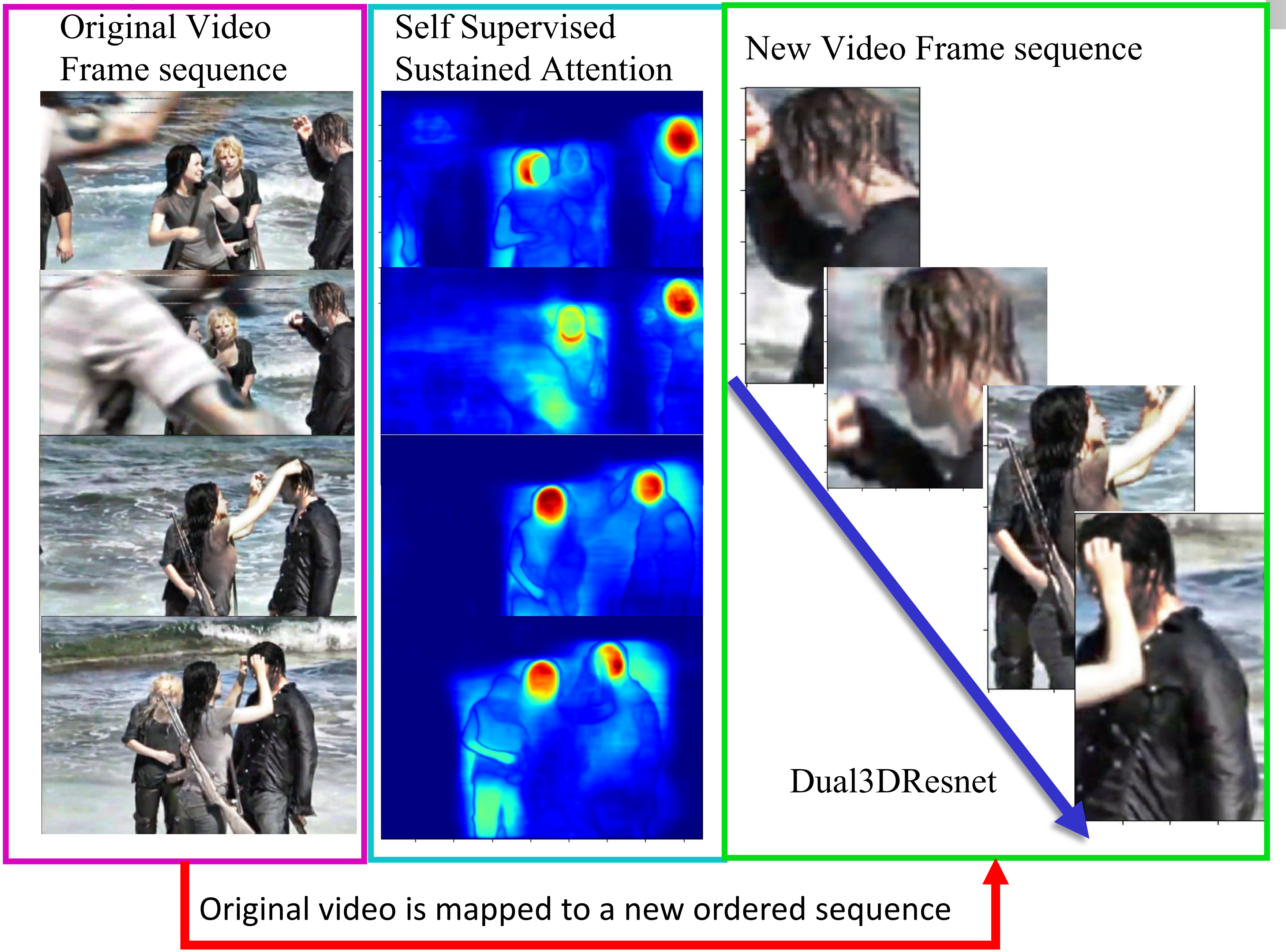


Sustained attention uses the *WM* to prioritize hot regions to build spatio-temporal relations.



Working Memory
(WM)





Original Video Frame sequence

Self Supervised Sustained Attention

New Video Frame sequence

Dual3DResnet

Original video is mapped to a new ordered sequence

Conclusions

A relatively complex task can be reduced to a simple one when the number of events is limited and when “where to look at” is known.

When the task is challenging sustained attention becomes crucial.

We have shown that basing on self-supervised sustained attention we can solve very hard recognition tasks.

- The work on Sustained attention will be on ArXiv soon together with the code.