Politecnico di Torino

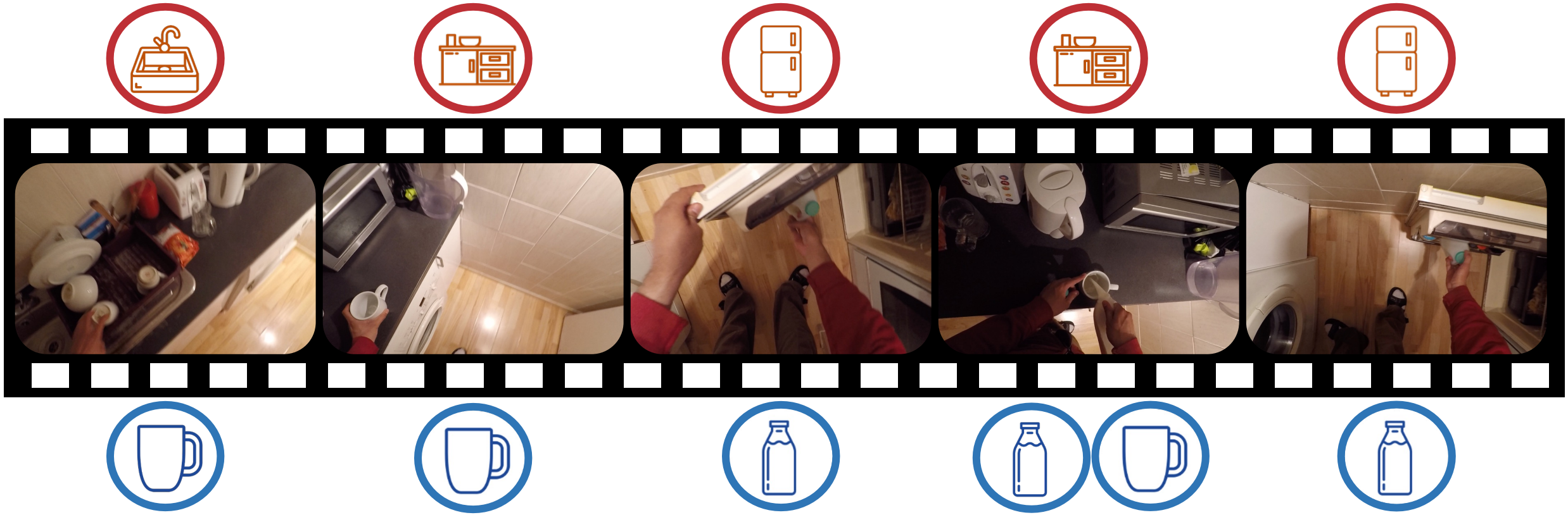
Meta

University of BRISTOL

AMEGO: Active Memory from long EGOcentric videos

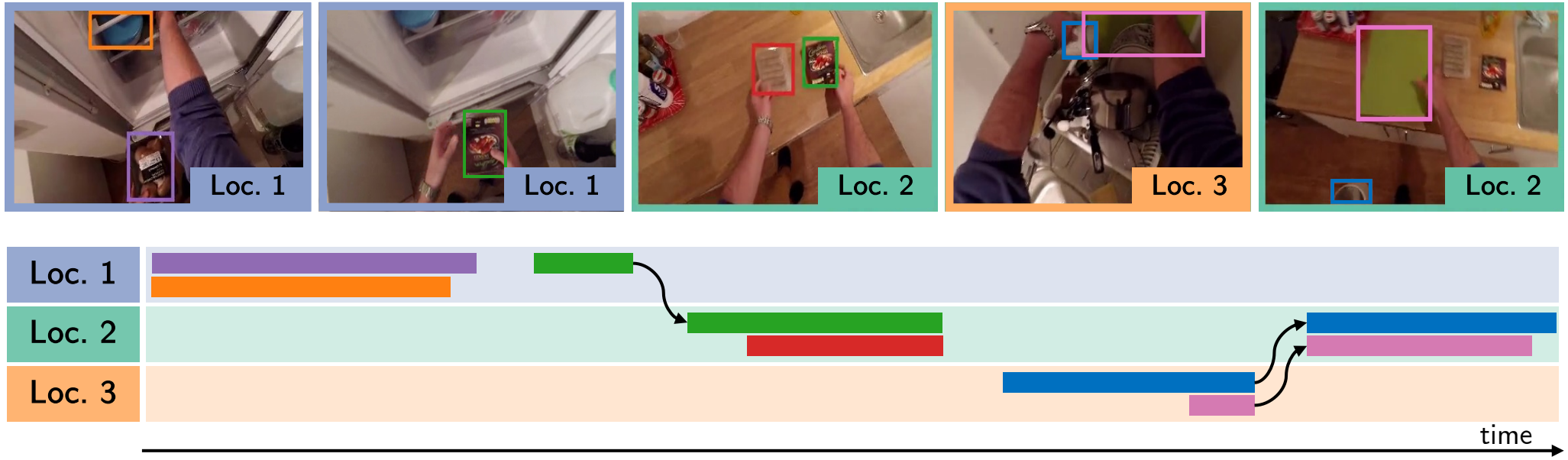
Gabriele Goletto, Tushar Nagarajan,
Giuseppe Averta, Dima Damen

Egocentric videos are NOT just a collection of frames…



…but rather a sequence of interactions of the camera wearer

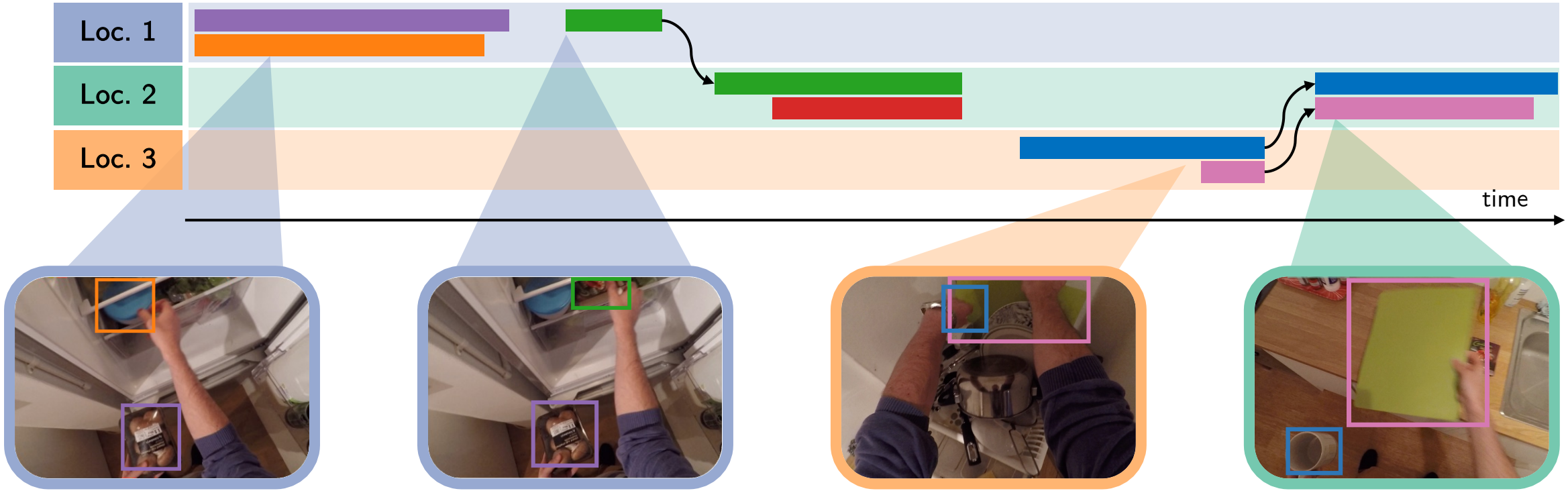
AMEGO is an **online, structured, semantic-free** representation capturing both **objects interacted with, locations visited**, and their **interplay**



Check more AMEGO examples

SCAN ME

AMEGO advantages



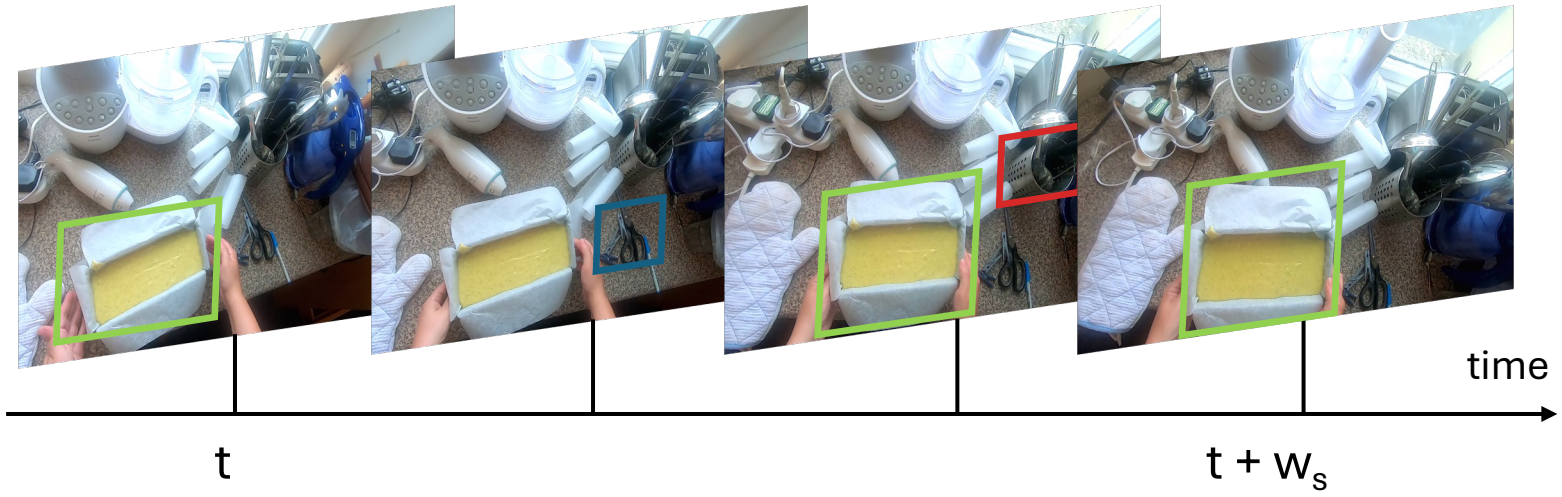
Interpretable representation, e.g. identifying co-occurrences and interaction starts

Human-centric interactions

Semantic-free and class-agnostic

How do we build AMEGO?

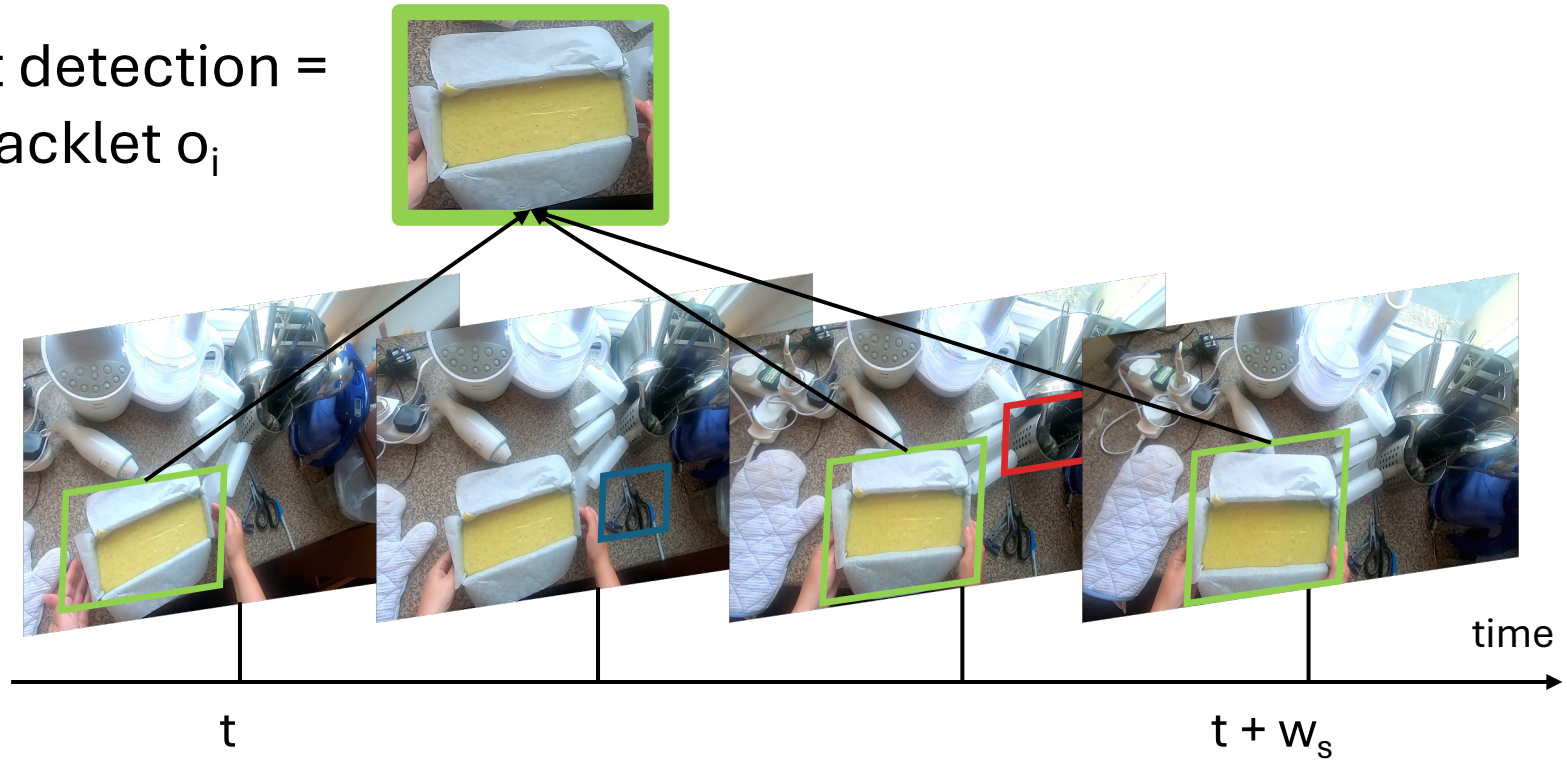
Initialisation



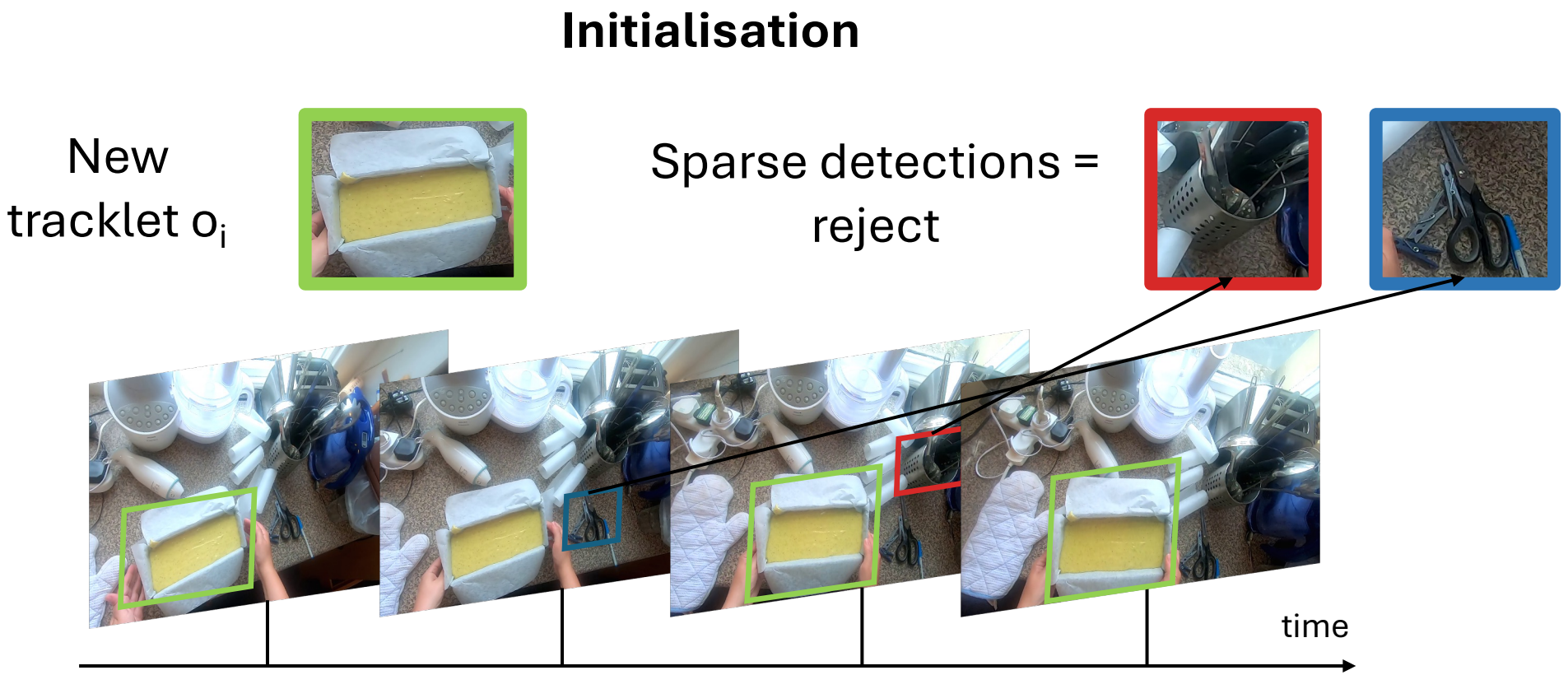
How do we build AMEGO?

Initialisation

Consistent detection = new tracklet o_i



How do we build AMEGO?

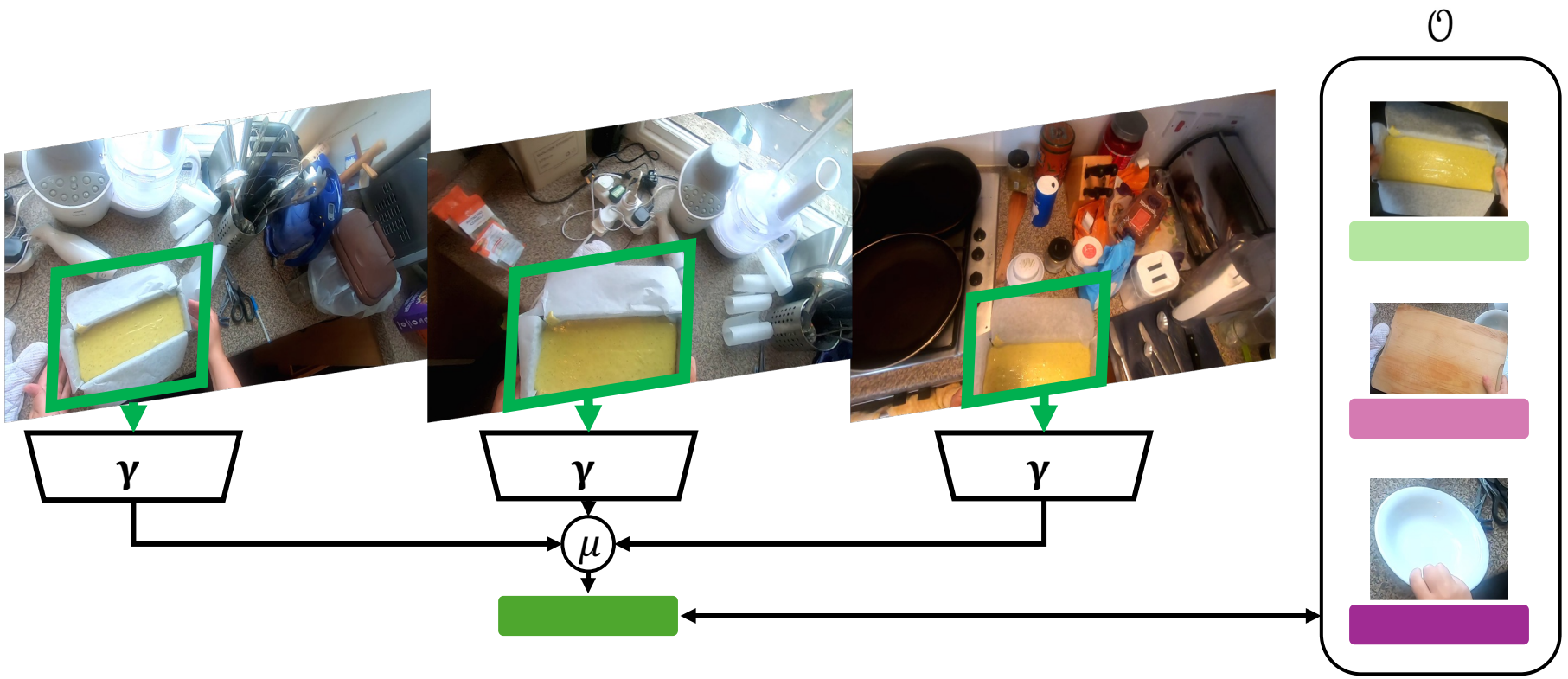


How do we build AMEGO?

Updating

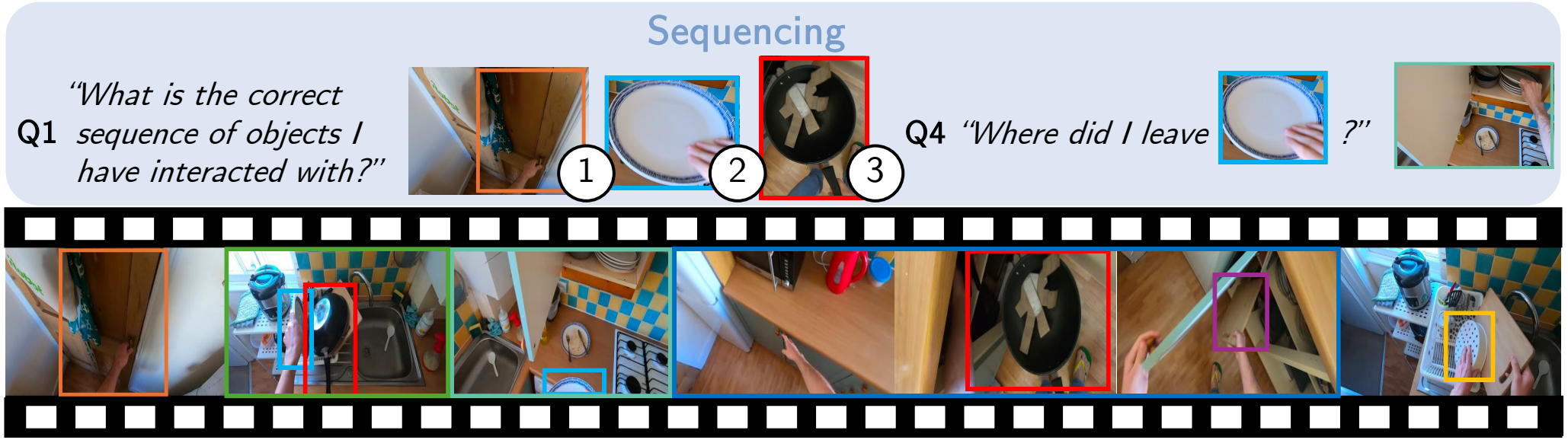
Tracking even when the hands are out of view

How do we build AMEGO?

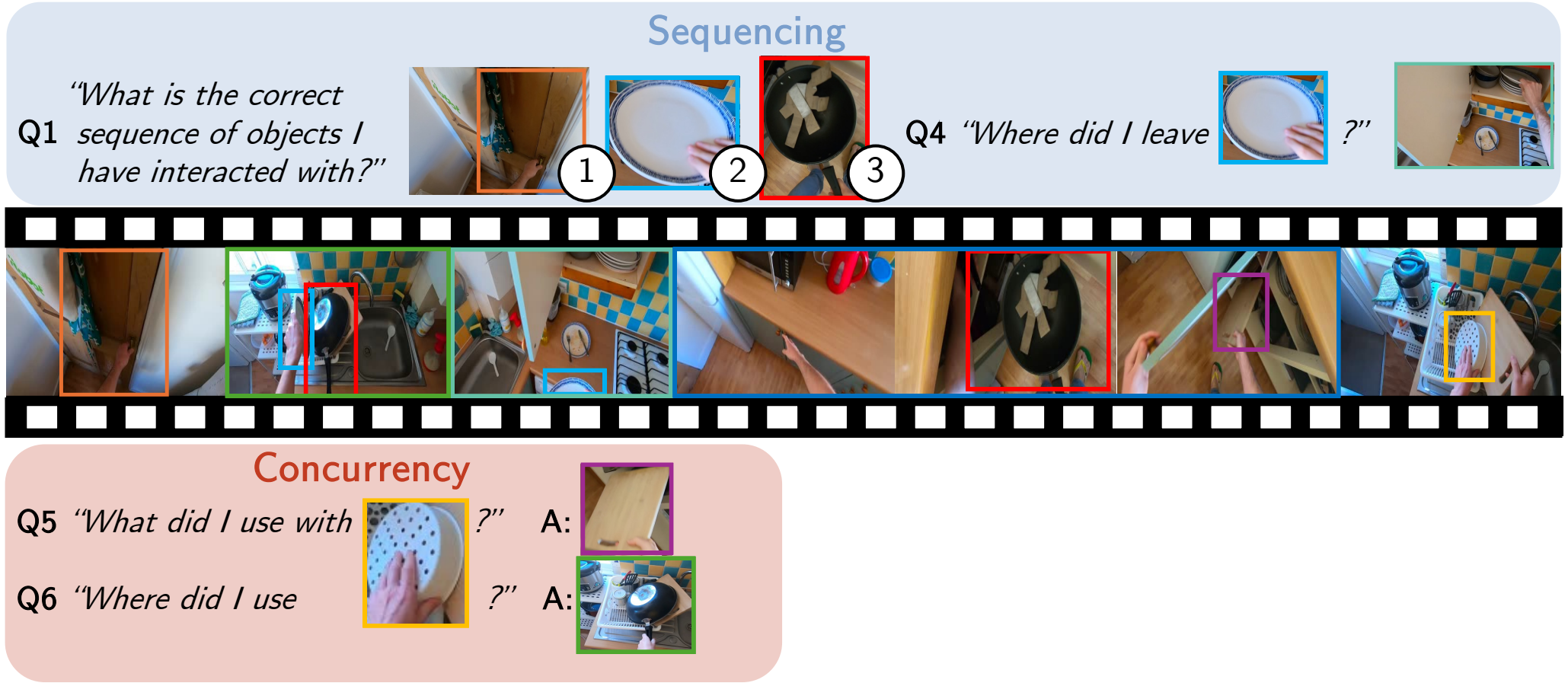


Active Memory Benchmark

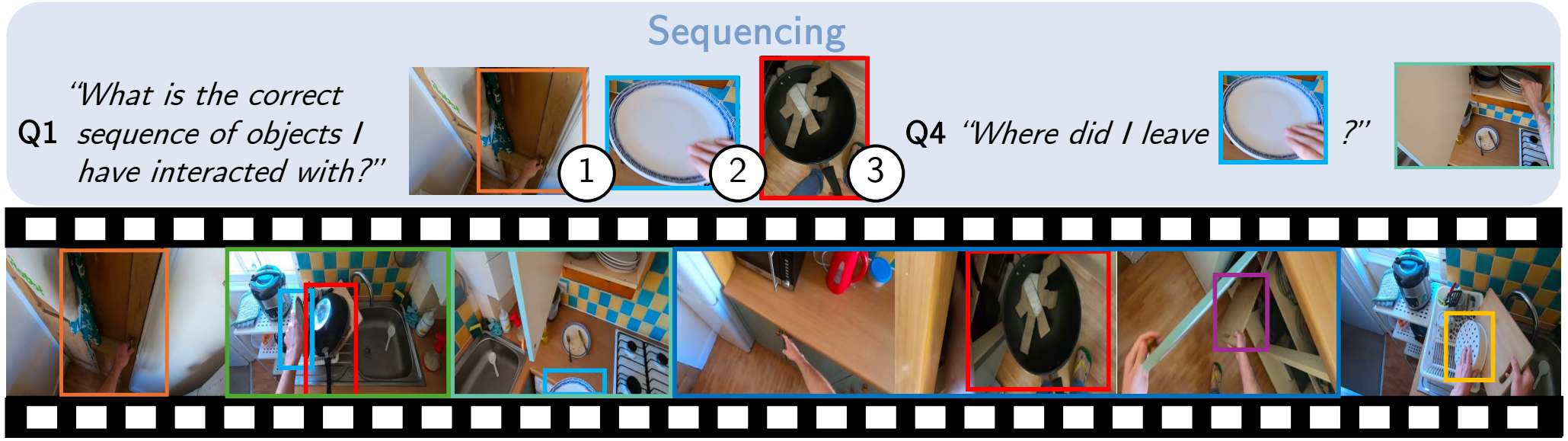
Active Memory Benchmark



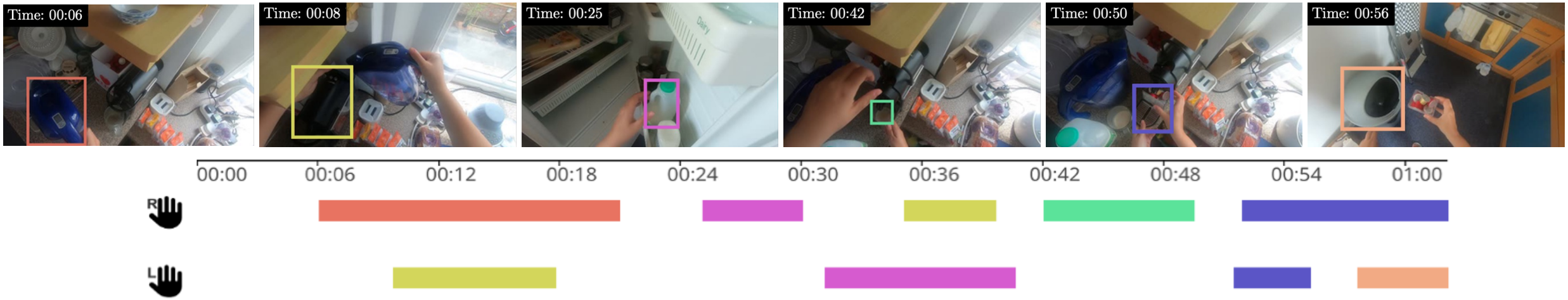
Active Memory Benchmark



Active Memory Benchmark



Querying AMB with AMEGO

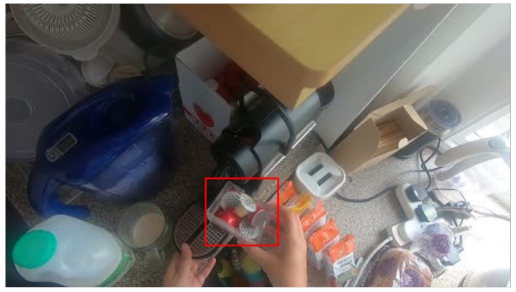


Q2: What did I use with the left hand after [VQ] at time 00:20?

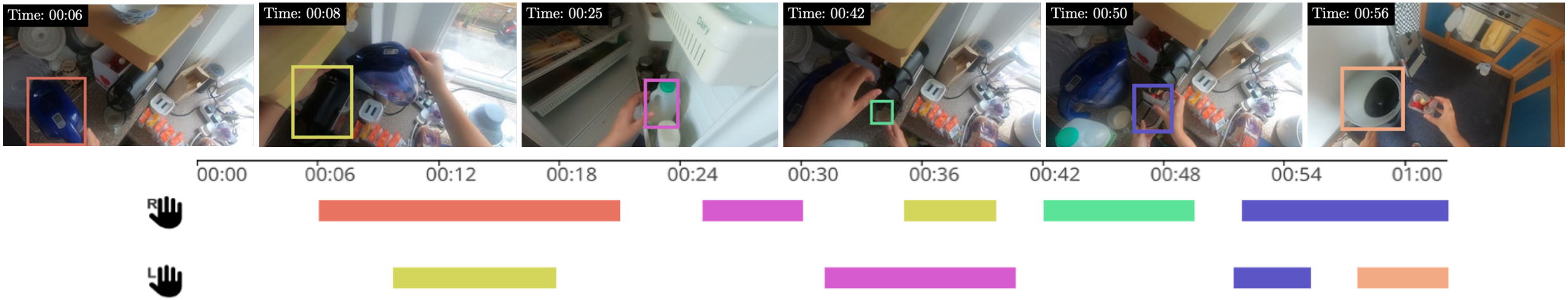
[VQ]

Ans. A

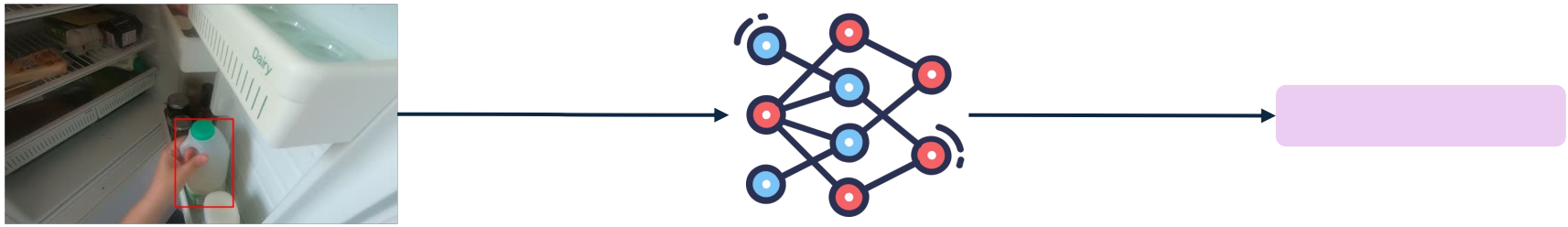
Ans. B



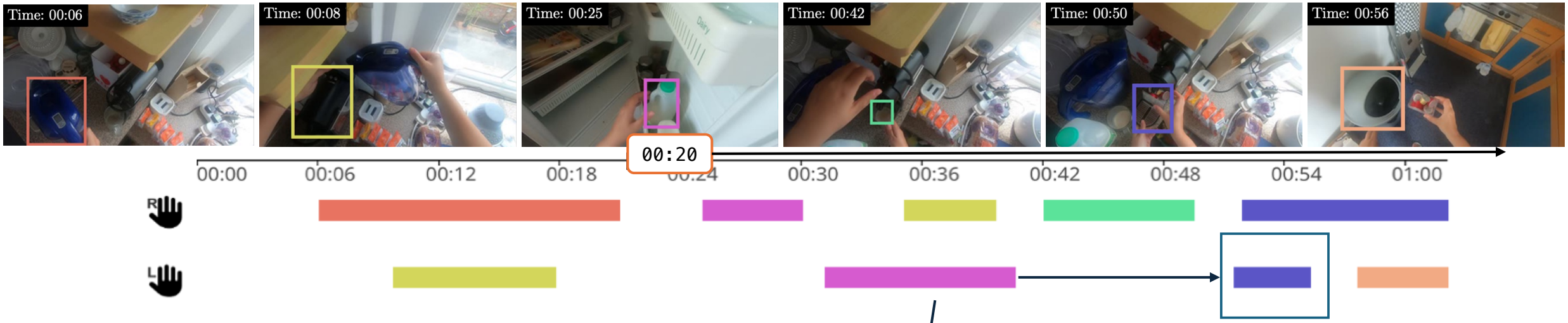
Querying AMB with AMEGO



Step 1: Extract [VQ] features

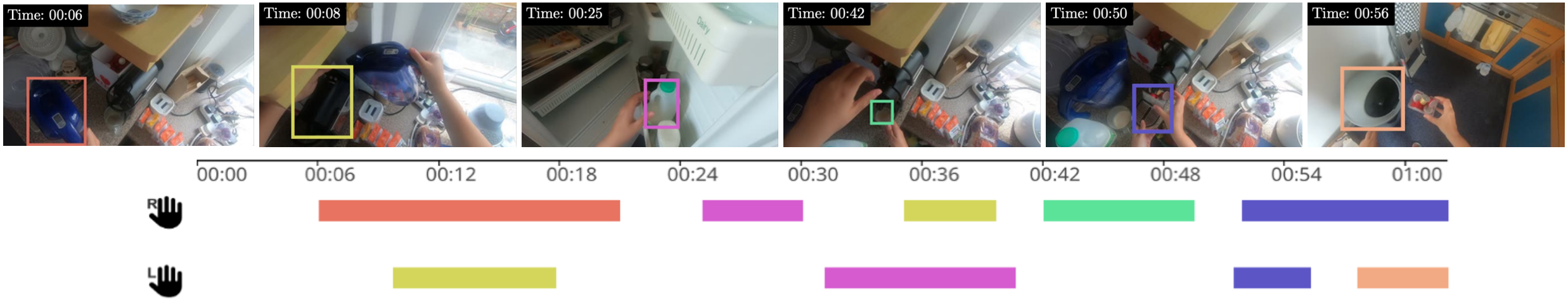


Querying AMB with AMEGO

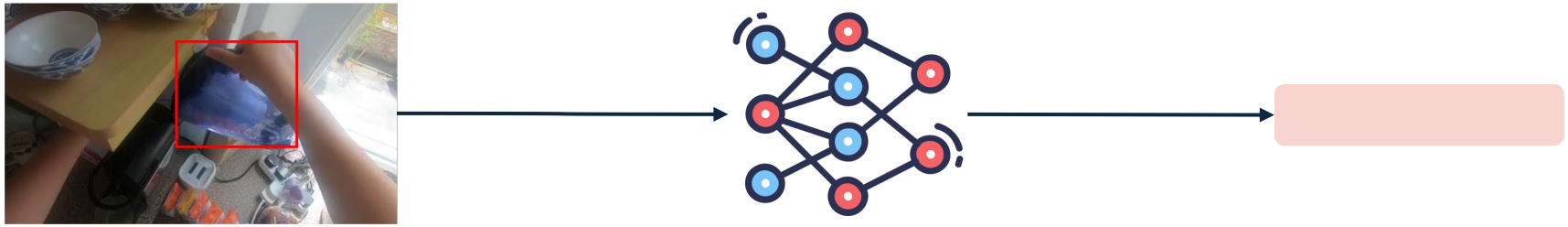


Step 2: Match the extracted features with AMEGO features (only object interactions after 00:20 involving the left hand) and retrieve the next instance

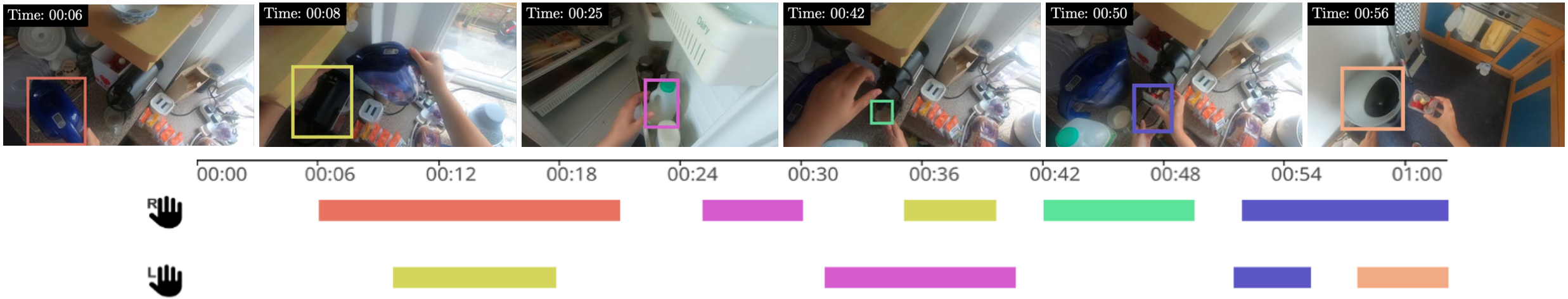
Querying AMB with AMEGO



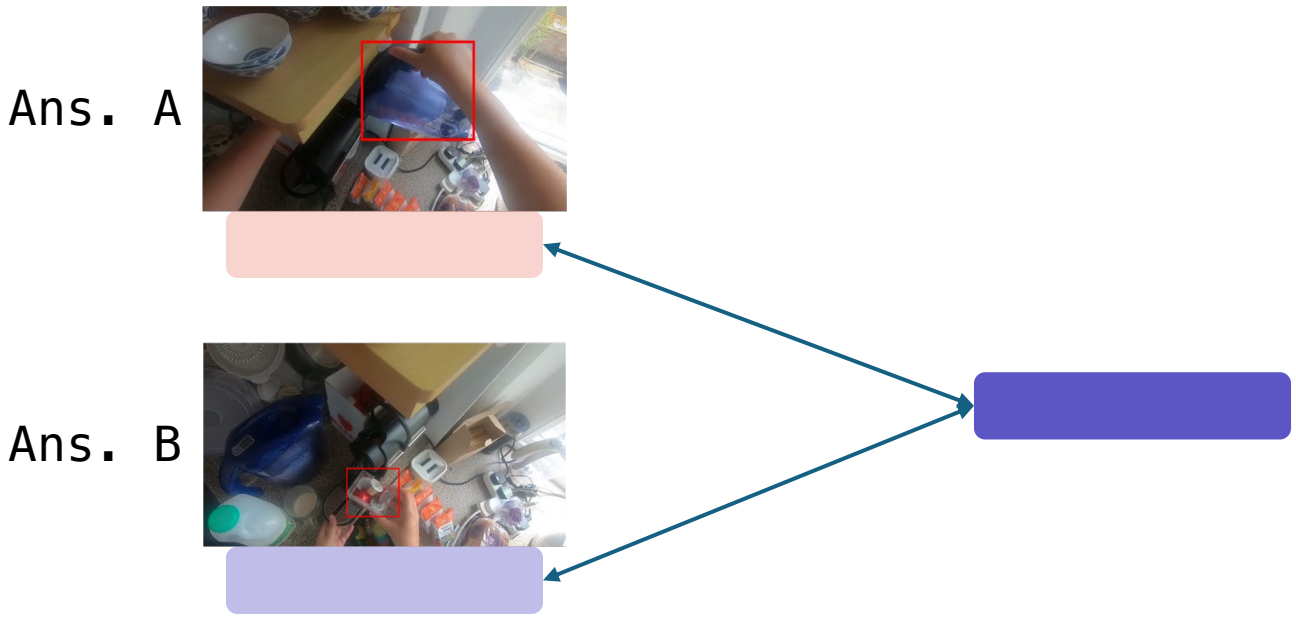
Step 3: Extract features for answer image (Ans. A)



Querying AMB with AMEGO



Step 4: Compare the answer features with the retrieved instance
Pick the answer whose features have the highest similarity



Pick Ans. B

Results

Semantic-Free Vision-Language embedding

Semantic captioning approaches

Multi-round captioning approaches

Method	SQ				CO		TG		Total
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	
Random	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0	20.0
SF-QA	13.7	21.6	22.5	26.8	22.1	31.9	23.7	26.2	22.0
SF-QA (obj)	13.1	23.4	22.6	23.2	21.7	26.1	23.8	25.2	21.2
S-QA (LaViLa)	20.9	20.6	21.2	24.6	24.9	27.1	21.4	22.6	22.4
S-QA (BLIP-2)	23.9	22.0	22.5	23.3	27.5	27.0	20.2	24.1	23.6
S-QA (LaViLa+BLIP-2)	22.8	22.2	21.4	22.6	25.1	26.1	21.4	24.5	22.9
LLoVi (LaViLa)	21.1	20.2	20.8	21.0	21.2	20.3	20.5	21.6	20.8
LLoVi (BLIP-2)	22.3	21.4	21.8	22.2	25.6	26.7	18.1	22.2	22.4
LLoVi (LaViLa+BLIP-2)	22.8	21.9	21.5	24.6	25.3	26.5	18.5	19.8	22.6
AMEGO - S	32.0	35.1	34.8	35.8	24.7	37.8	33.6	44.3	33.8
AMEGO - L	**33.7**	**36.3**	**37.2**	**38.3**	**27.6**	**44.3**	**34.7**	**48.9**	**36.3**