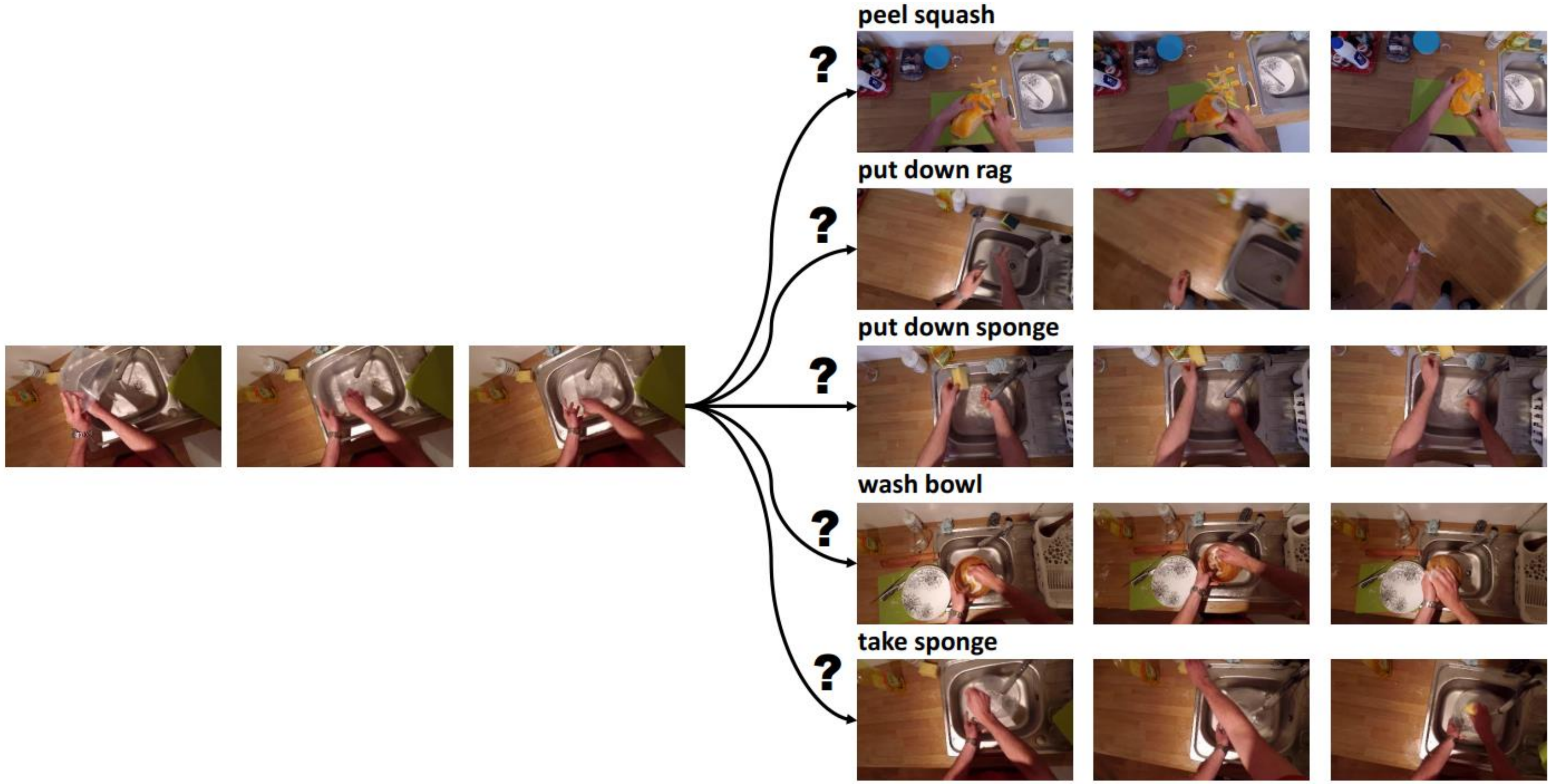
Part5: Cooking action anticipation

Motivation

Predicting what will happen in the future!

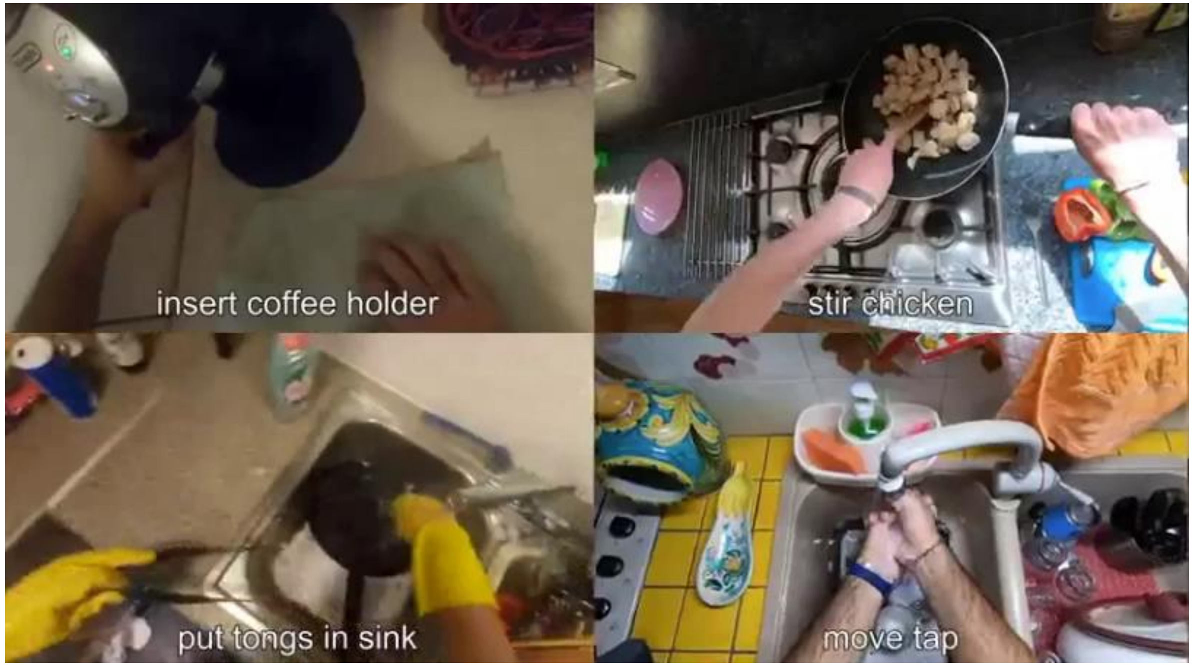


Motivation

- **Egocentric Vision**

□ wearable cameras

□ daily activities



Motivation

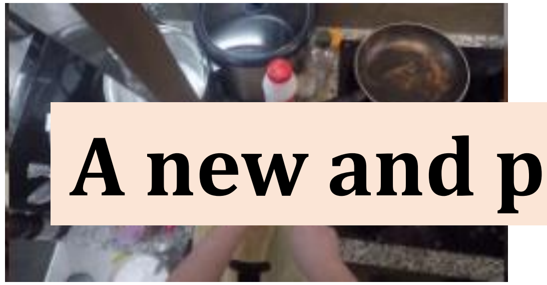
- ➢ in the kitchen environment
- ➢ cooking-related actions

fry egg

pour milk

split salmon

A new and promising area of food computing



slice chilli

flip fish

apply spreads

Motivation

- In the kitchen environment

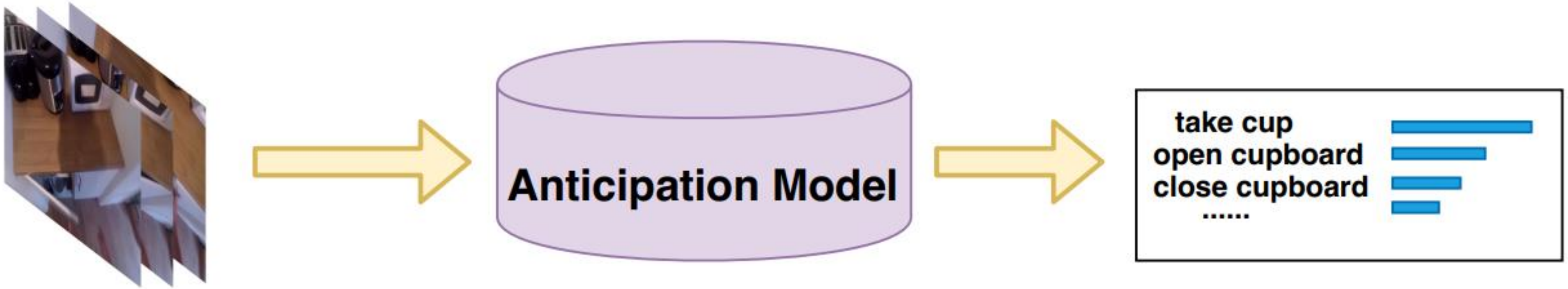
- Application of sevice robots

□ Help those who are disabled to cook recipes, wish dishes, etc

□ Instruct people to learn how to cook

Egocentric Action Anticipation

➢ Definition



Observed video segment

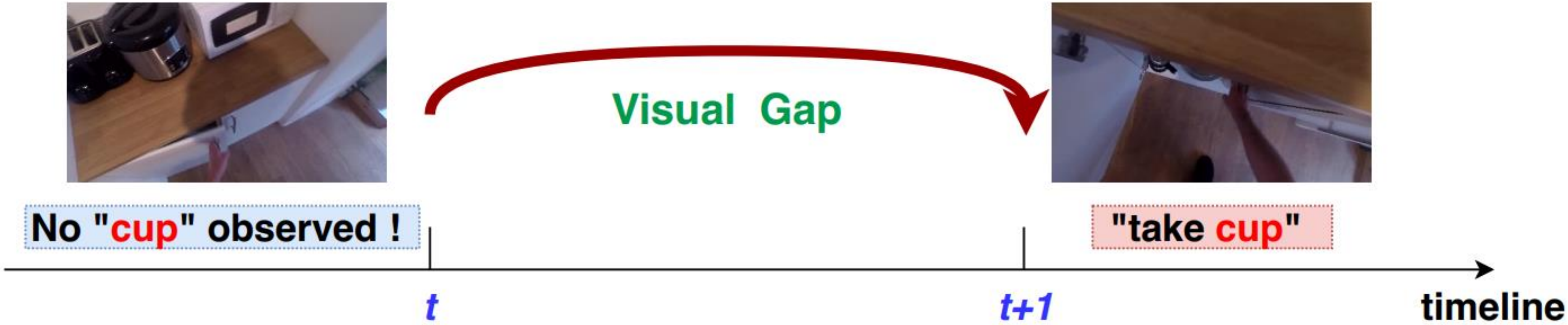
(*ends at time point* ***t***)

Label of action

(*starts at time point* ***t+1***)

Egocentric Action Anticipation

- **Difficulty**



No "cup" observed !

Visual Gap

"take cup"

t

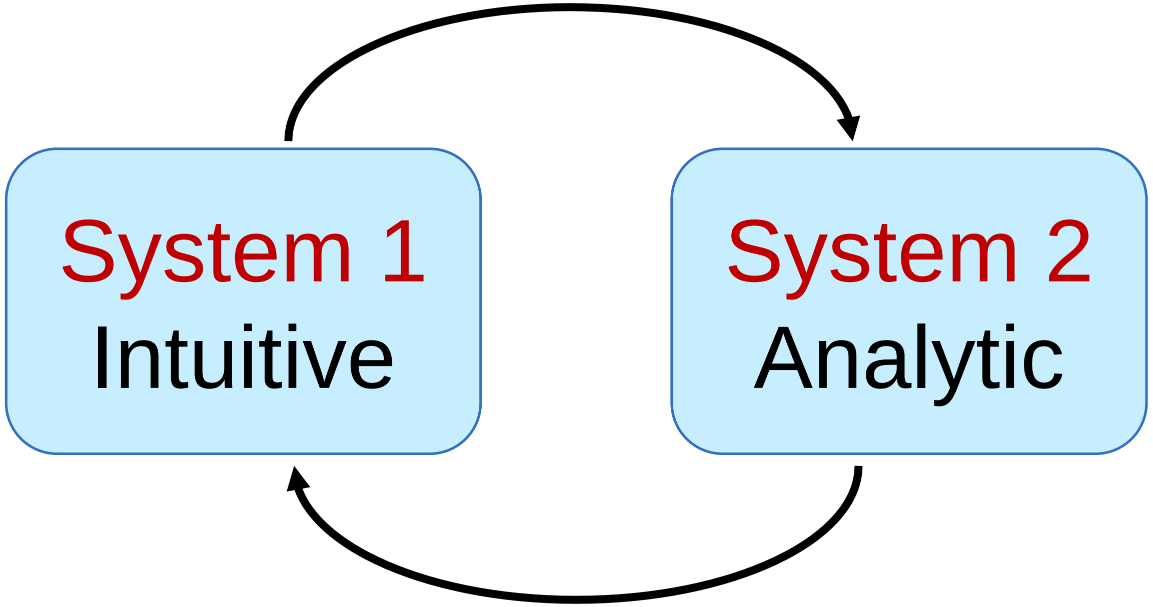
t+1

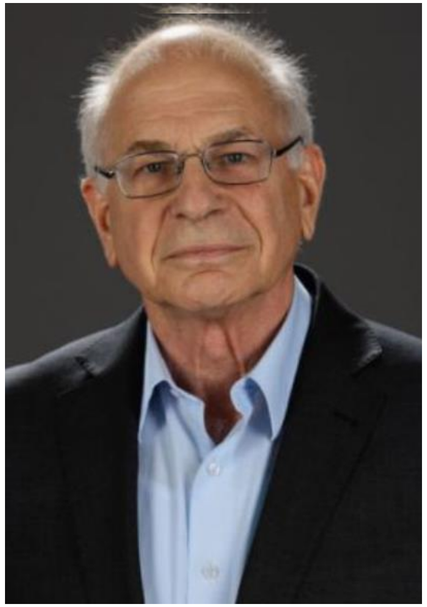
timeline

Exploration from human psychology

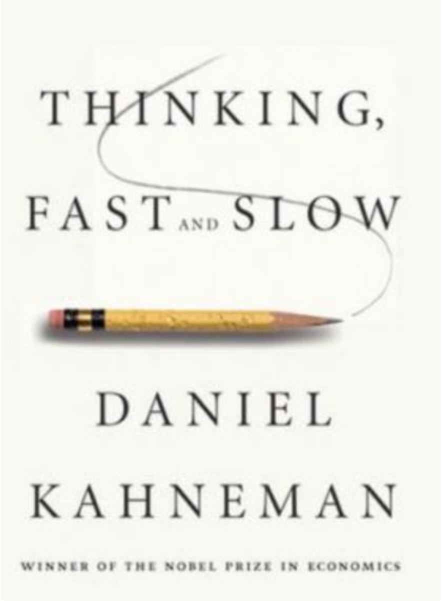
□Two modes in the cognitive system of human brain: **intuition and analysis**

□Intuition and anlysis are both crucial in solving many problems (e.g. making predictions)



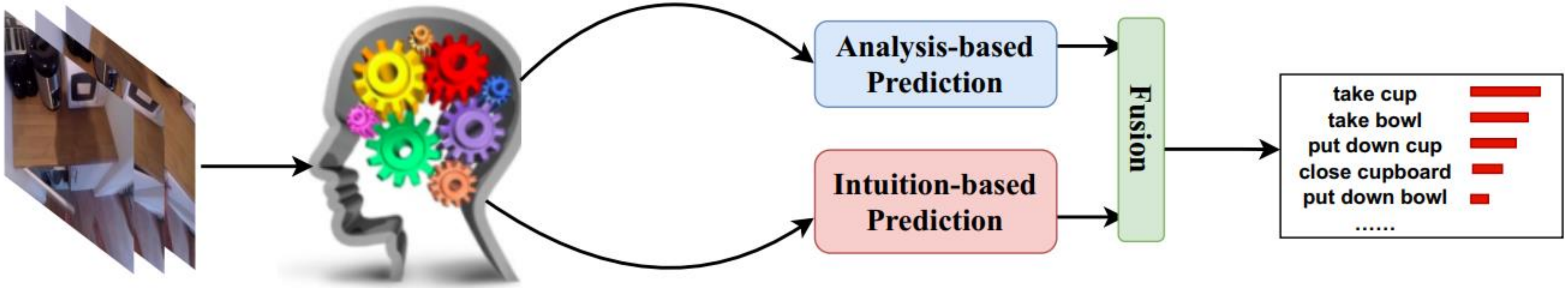


Daniel Kahneman



Exploration from human psychology

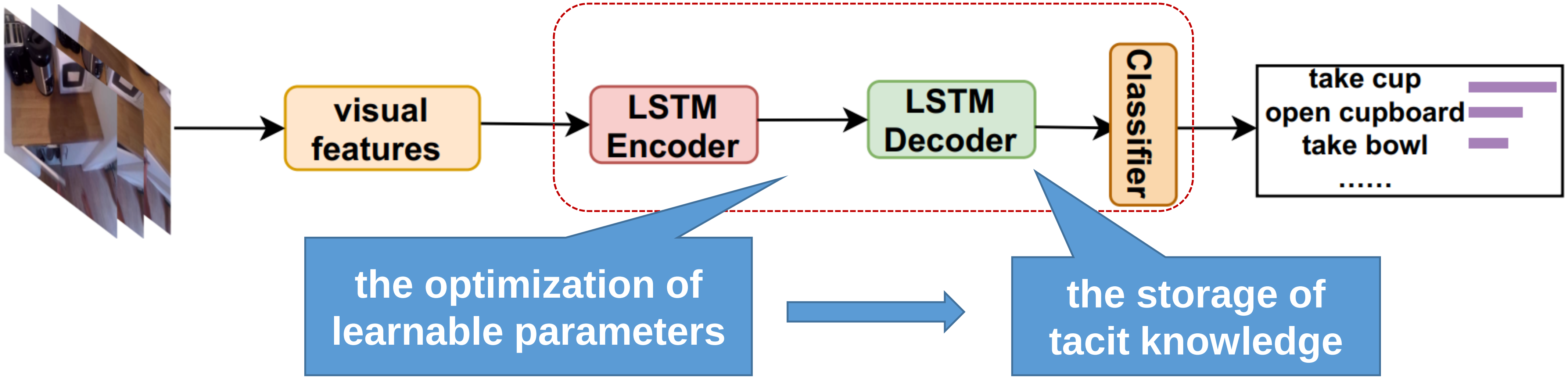
Construct a basic framework that integrates both intuition-based prediction and analysis-based prediction to imitate human beings in making predictions



Exploration from human psychology

□ Intuition-based prediction

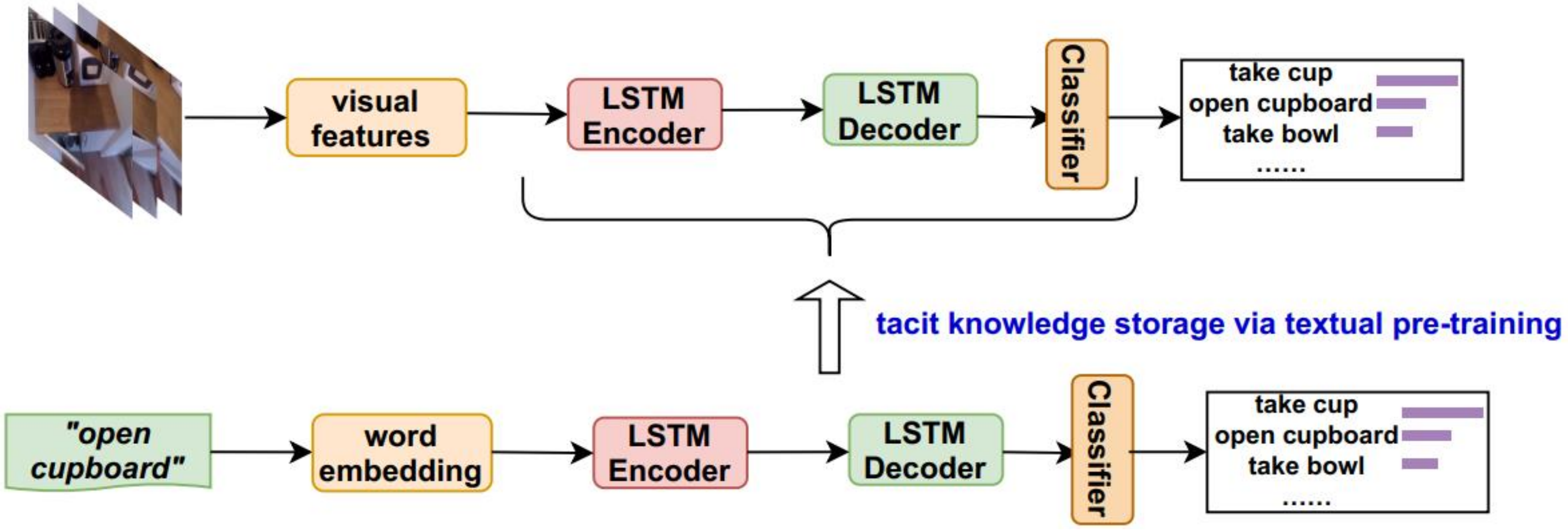
- ➢Subconscious, habitual
- ➢Tacit knowledge (**hard to explain**)
- ➢An **encoder-decoder** structure (a black-box process)



Exploration from human psychology

□ Intuition-based prediction

- ➢Visual information is insufficient to store tacit knowledge
- ➢Introduce textual pre-training to store tacit knowledge in advance



Exploration from human psychology

□ Intuition-based predictior

➢ (b)→(c):

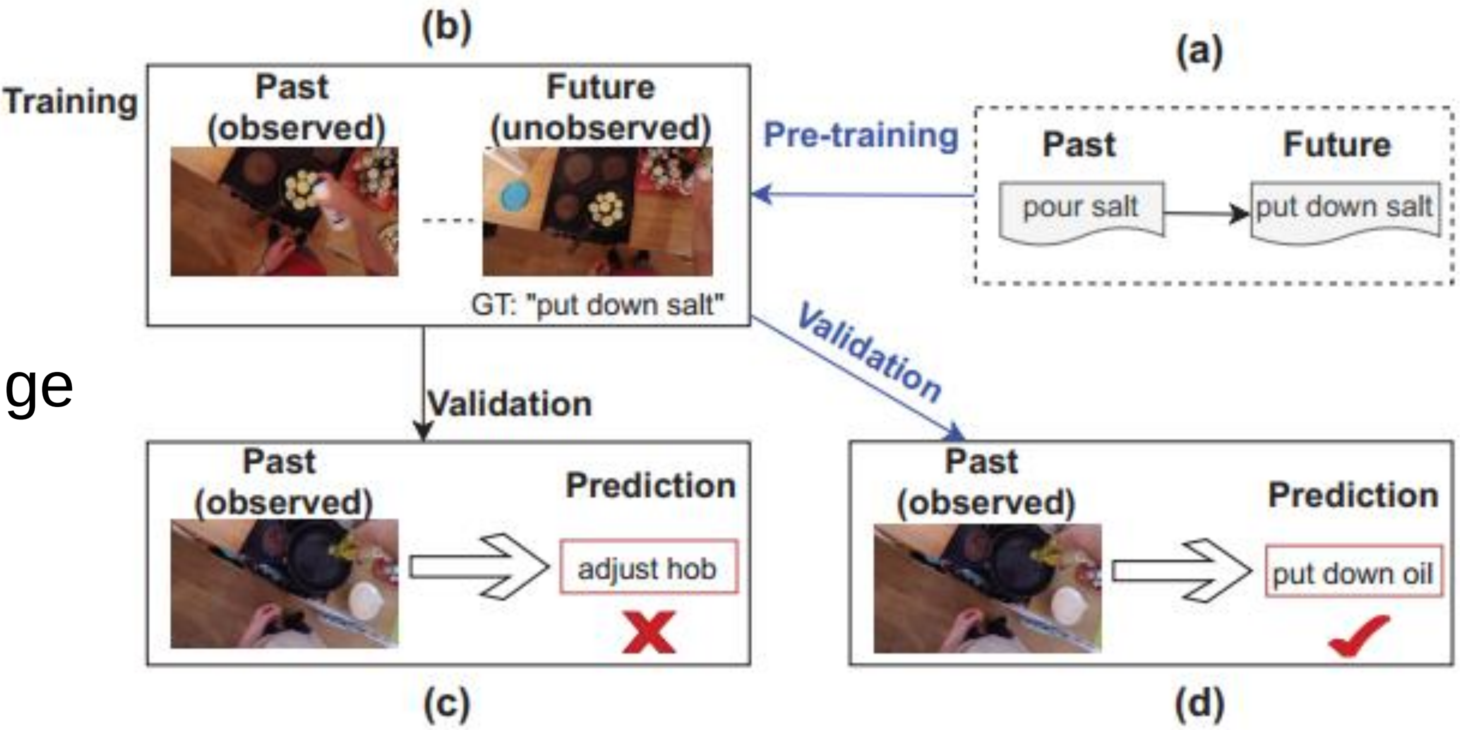
□ Visual information

□ Insufficient to store tacit knowledge

➢ (a)→(b)→(d):

□ Visual + text information

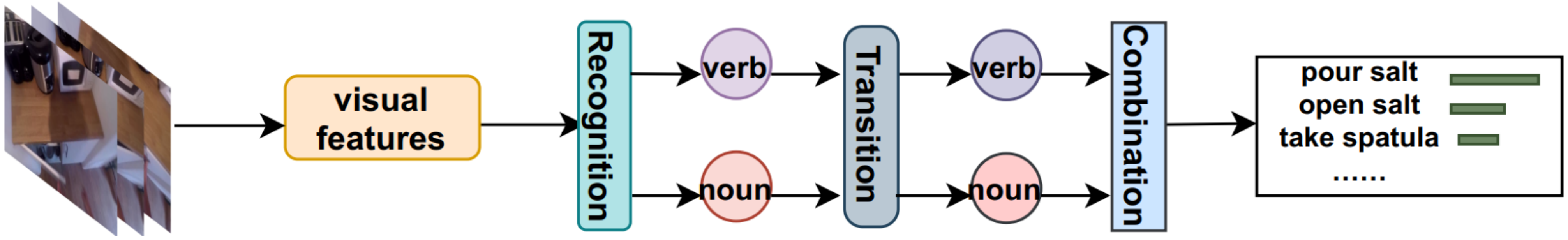
□ Store more reliable tacit knowledge



Exploration from human psychology

□ Analysis-based prediction

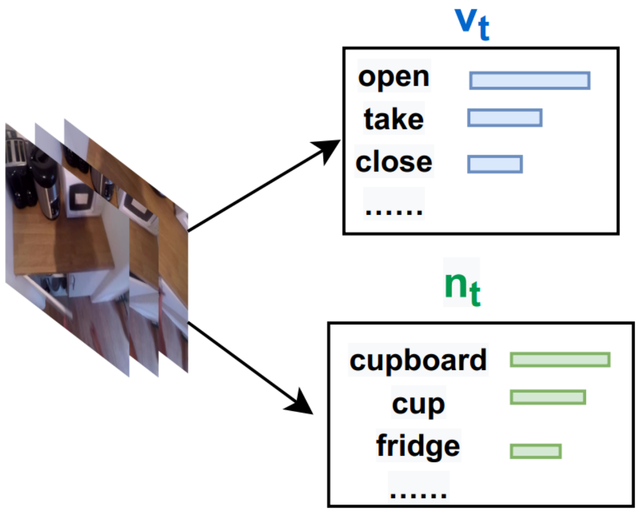
- ➢ Conscious and explicit
- ➢ Tend to process information under given principles
- ➢ An interpretable three-step pipeline



Exploration from human psychology

□ Analysis-based prediction

➢ **Recognition**



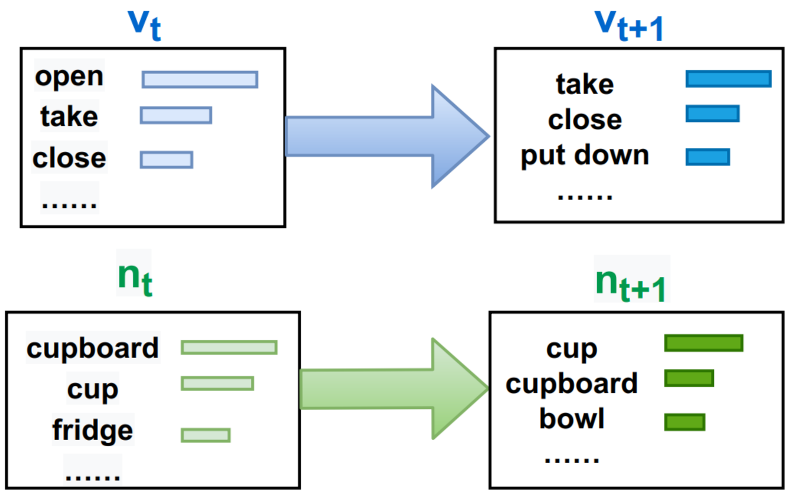
v_t

open
take
close
......

n_t

cupboard
cup
fridge
......

➢ **Transition**



v_t

open
take
close
......

v_{t+1}

take
close
put down
......

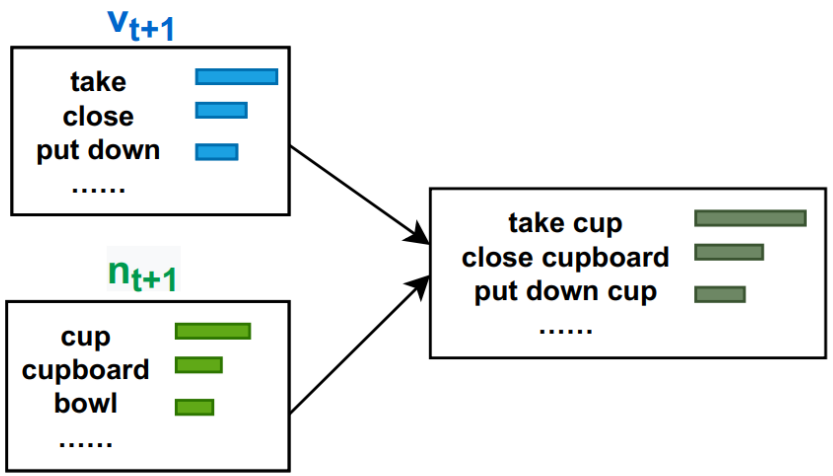
n_t

cupboard
cup
fridge
......

n_{t+1}

cup
cupboard
bowl
......

➢ **Combination**



v_{t+1}

take
close
put down
......

n_{t+1}

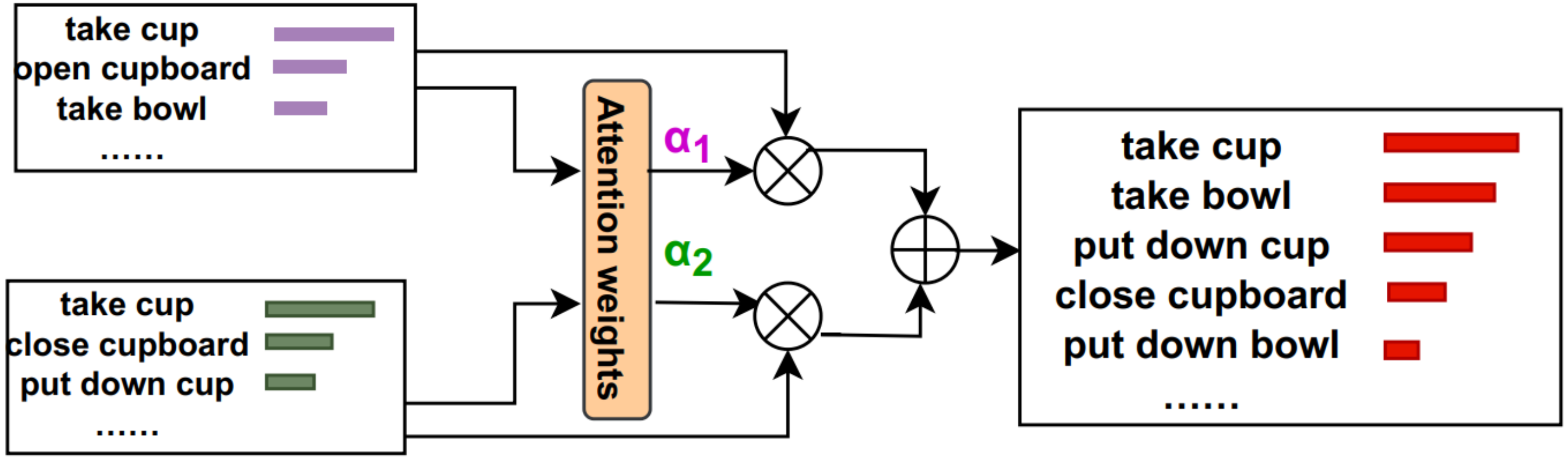
cup
cupboard
bowl
......

take cup
close cupboard
put down cup
......

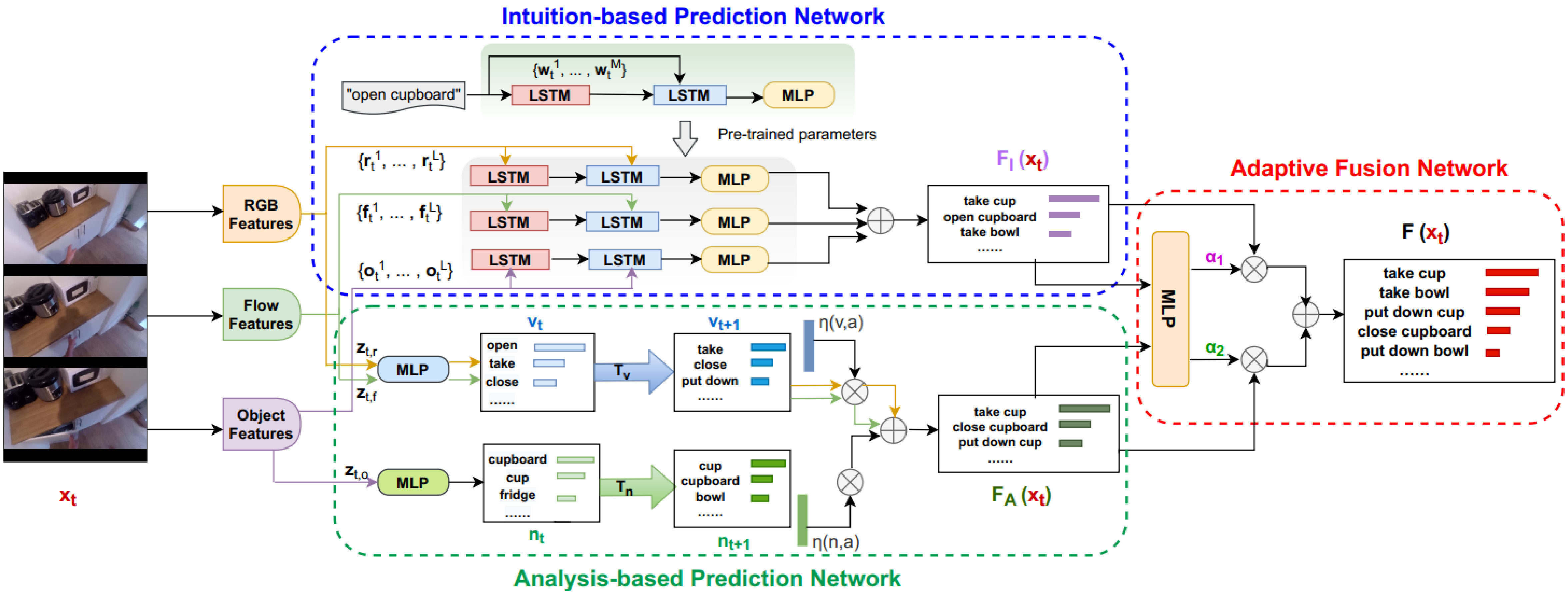
Exploration from human psychology

□ Intuition-analysis fusion

- ➢Both intuition and analysis are crucial and indispensable
- ➢Compute attention weights for intuition-based and analysis-based prediction and integrate them adaptively

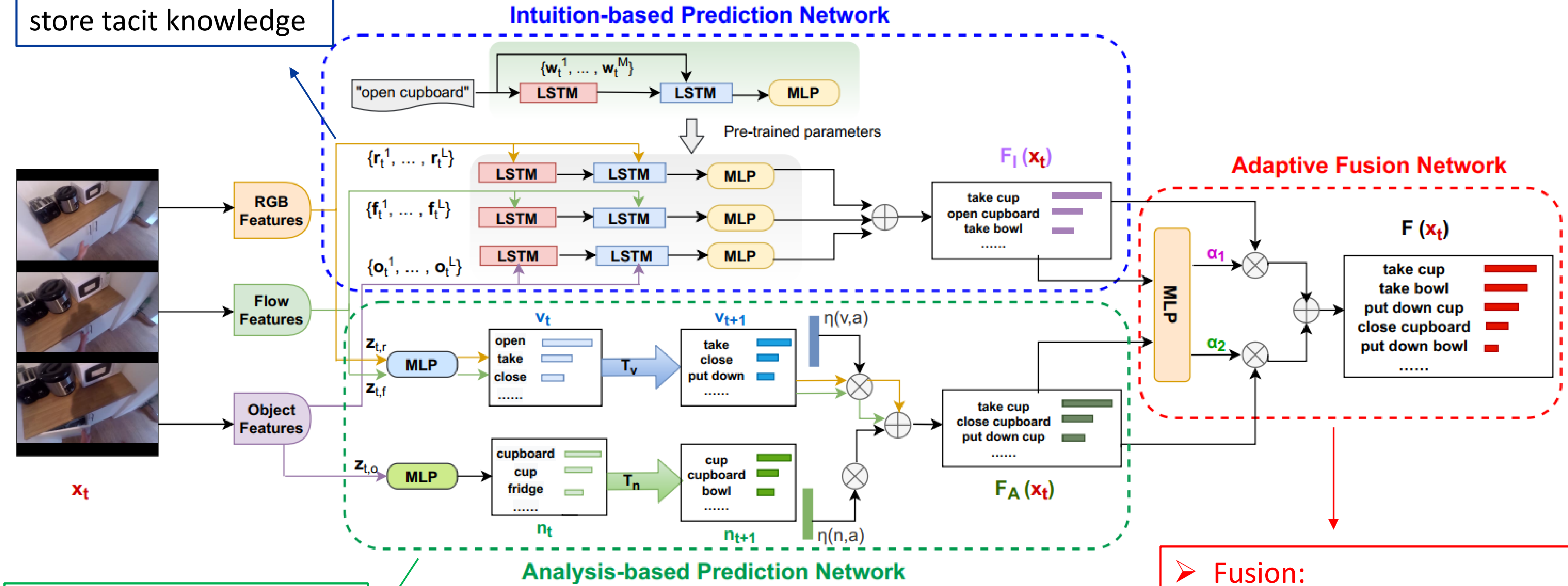


Intuition-Analysis Integrated Framework



Intuition-Analysis Integrated Framework

- LSTM:
store tacit knowledge



- Markov logic:
transit from past to future

- Fusion:
$F(x_t) = a_1{*}F_I(x_t){+}a_2{*}F_A(x_t)$

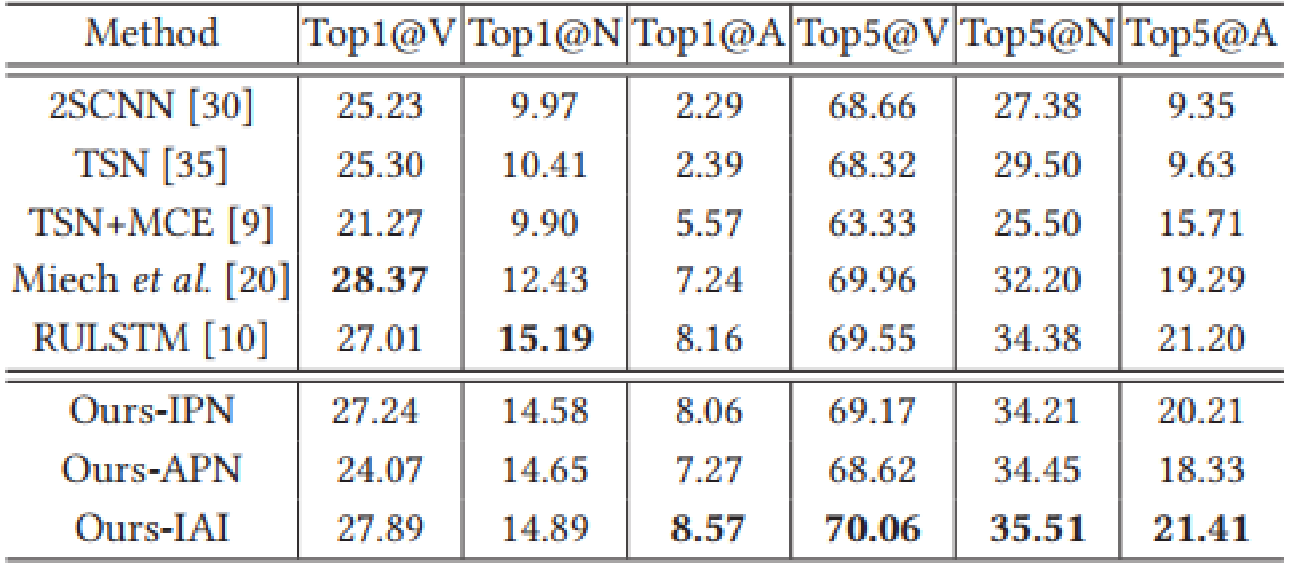
Evaluation

□ EPIC-Kitchens Dataset

- ➢ 32 kitchens - 4 cities
- ➢ Head-mounted camera
- ➢ 55 hours of recording - Full HD, 60fps
- ➢ 11.5M frames
- ➢ 39,594 action segments
- ➢ 125 verb classes
- ➢ 352 noun classes
- ➢ 2,513 action classes

Evaluation

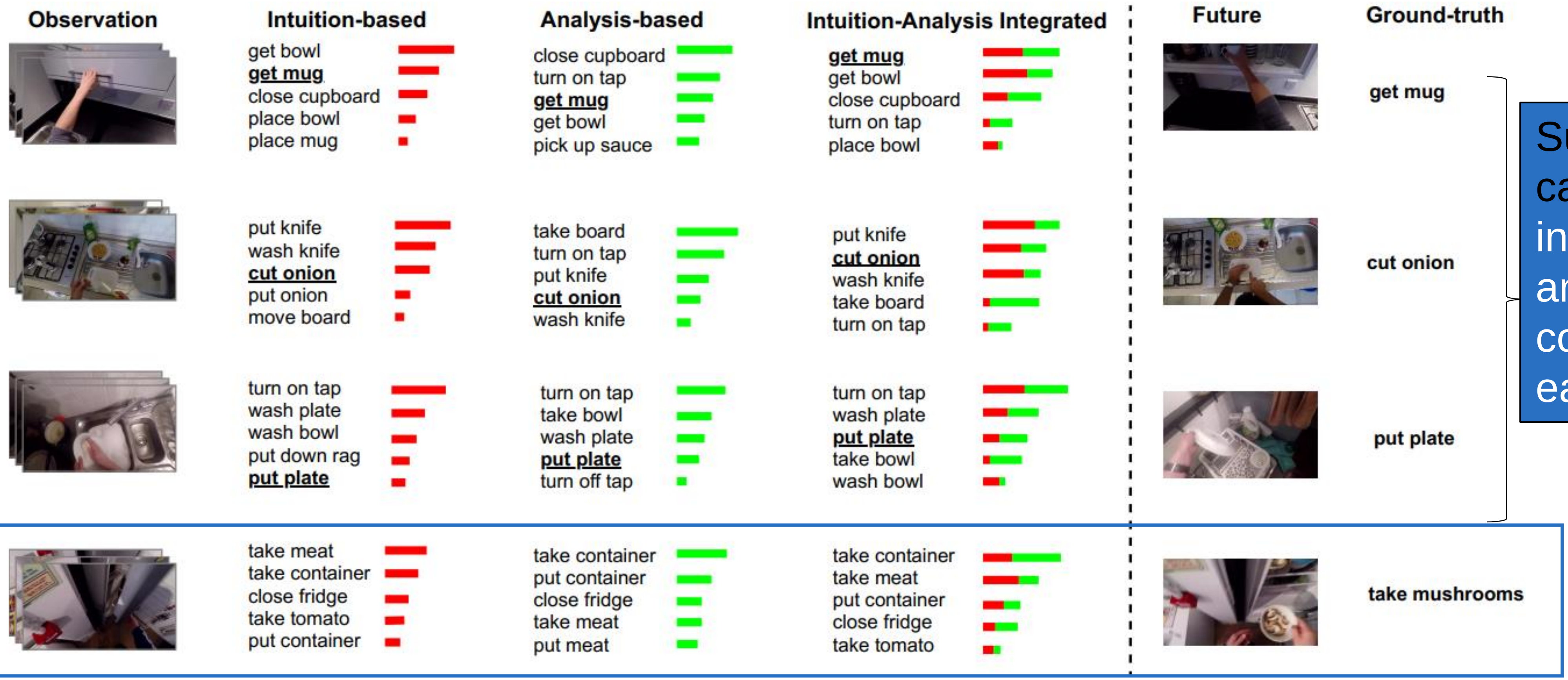
□ Comparison with other methods



Method	Top1@V	Top1@N	Top1@A	Top5@V	Top5@N	Top5@A
2SCNN [30]	25.23	9.97	2.29	68.66	27.38	9.35
TSN [35]	25.30	10.41	2.39	68.32	29.50	9.63
TSN+MCE [9]	21.27	9.90	5.57	63.33	25.50	15.71
Miech *et al.* [20]	**28.37**	12.43	7.24	69.96	32.20	19.29
RULSTM [10]	27.01	**15.19**	8.16	69.55	34.38	21.20
Ours-IPN	27.24	14.58	8.06	69.17	34.21	20.21
Ours-APN	24.07	14.65	7.27	68.62	34.45	18.33
Ours-IAI	27.89	14.89	**8.57**	**70.06**	**35.51**	**21.41**

- ➢ Top1@V/N/A: Top-1 accuracy for verbs/nouns/actions
- ➢ Top5@V/N/A: Top-5 accuracy for verbs/nouns/actions

Evaluation



Sucessful cases: intuition and analysis complement each other

More food-related information from cooking domain is needed !
(e.g., ingredient information of the being-prepared dishes)