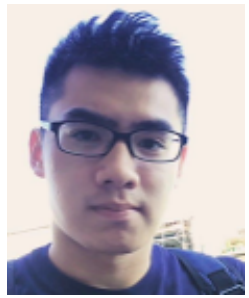
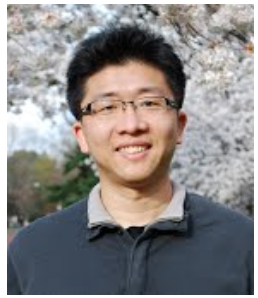
ICCV17

VSLAB

Anticipating Daily Intention using On-Wrist Motion Triggered Sensing

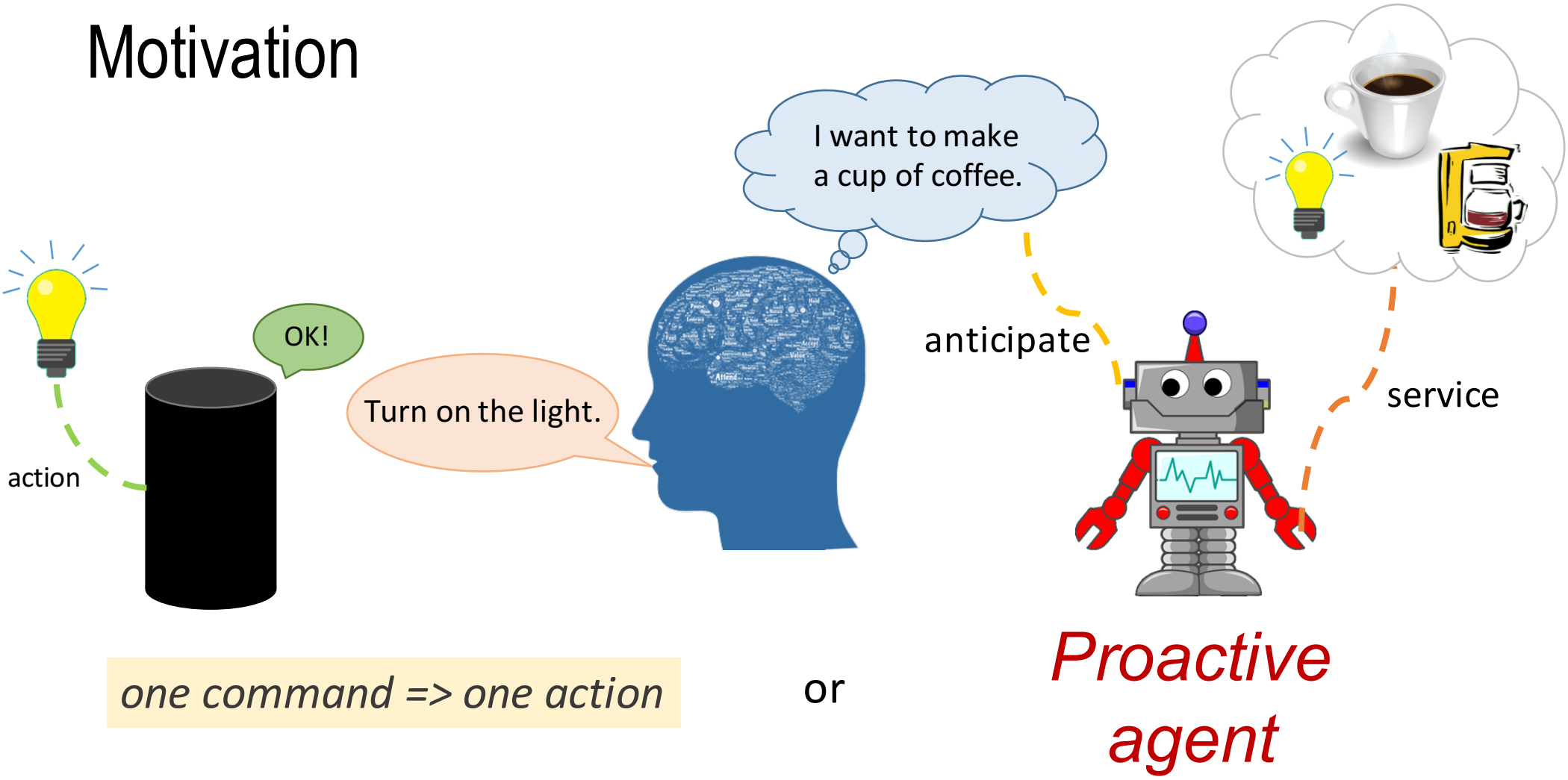


Tz-Ying Wu*, Ting-An Chien*, Cheng-Sheng Chan, Chan-Wei Hu, Min Sun
(*indicate equal contribution)

ICCV 2017 Spotlight

Motivation



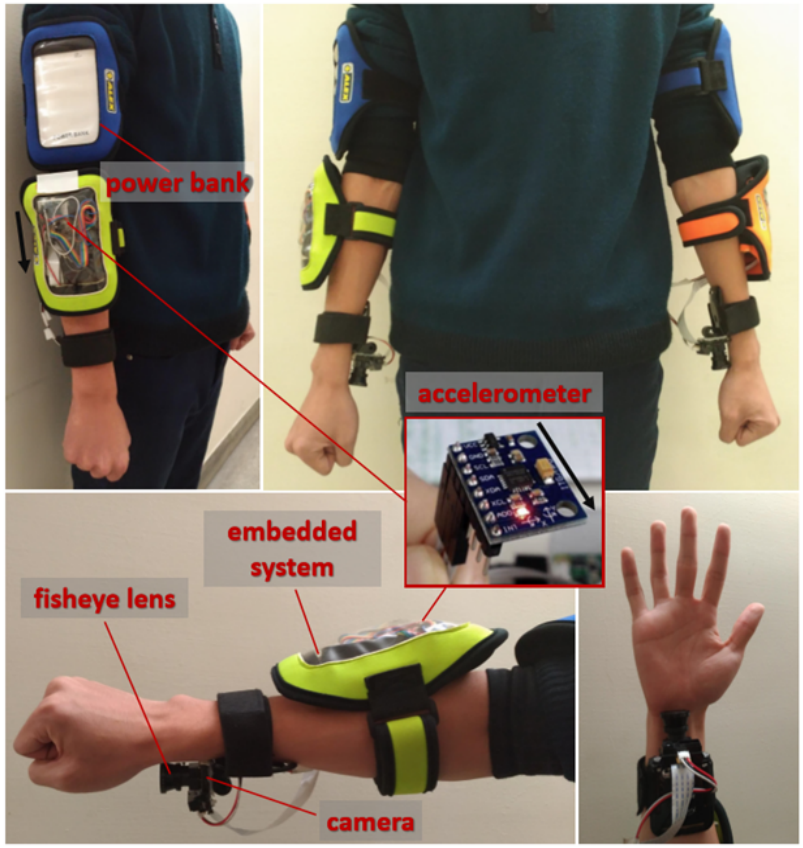
one command => one action

or

Proactive agent

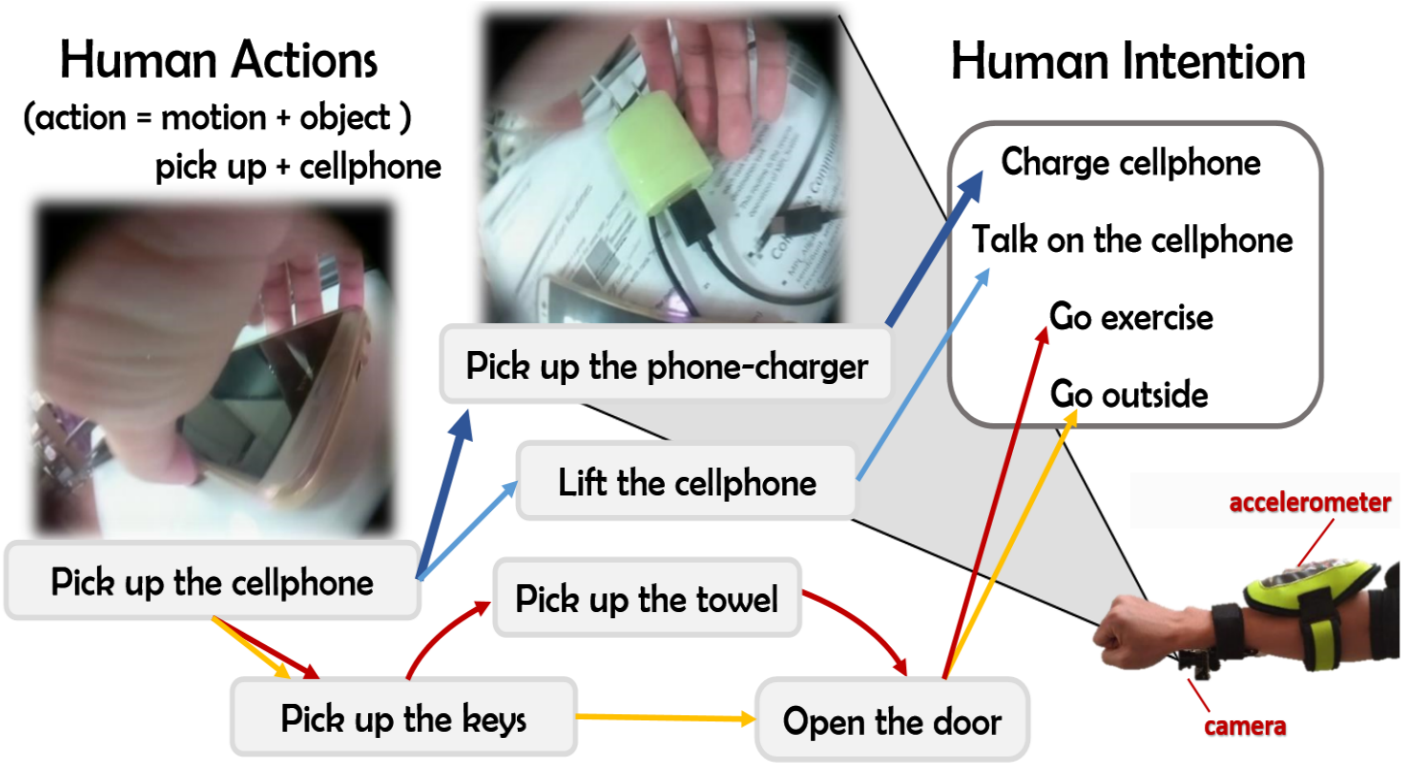
Overview

We adapt on-wrist sensors[1] to reliably capture daily human actions

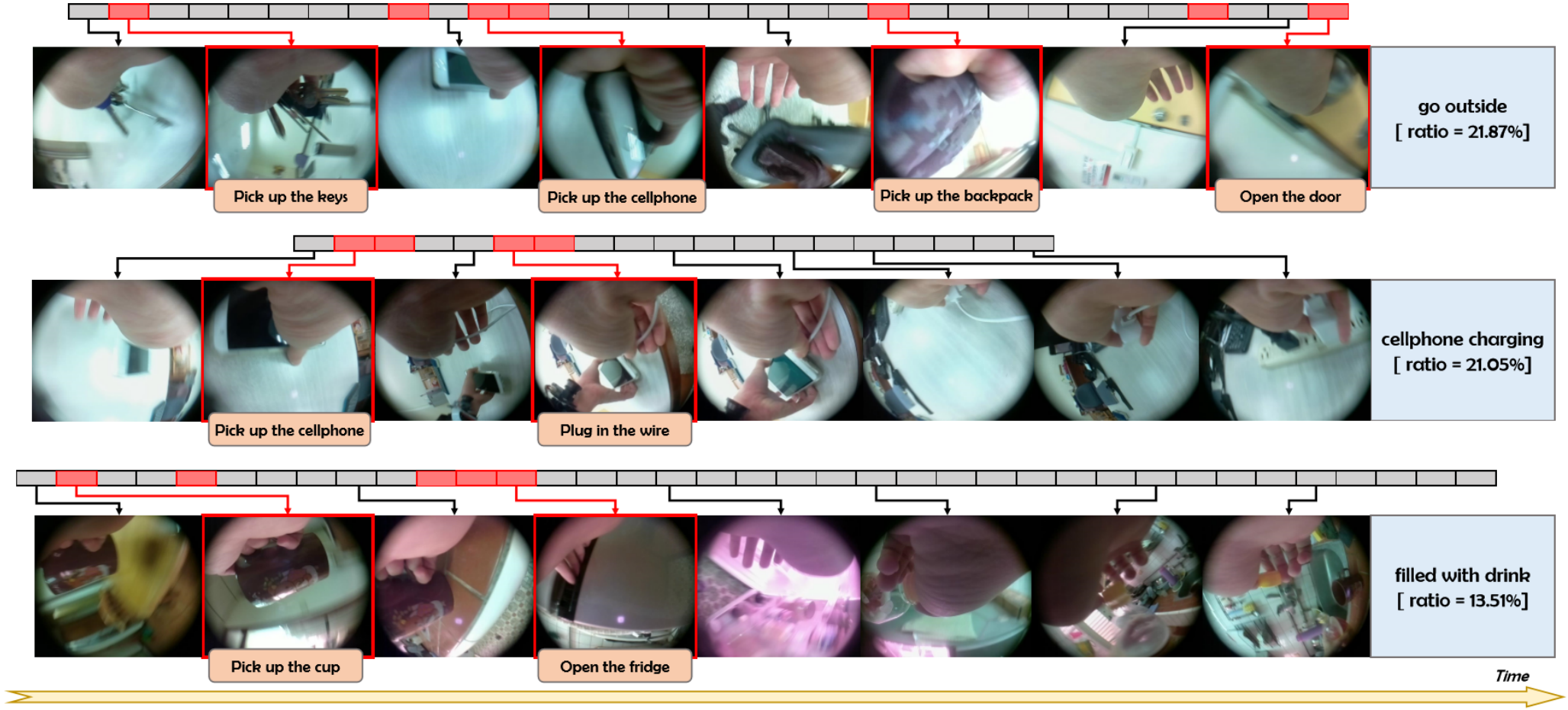


[1] Chan, C.-S., Chen, S.-Z., Xie, P.-X., Chang, C.-C., and Sun, M. (2016a). Recognition from hand cameras: A revisit with deep learning. In ECCV.

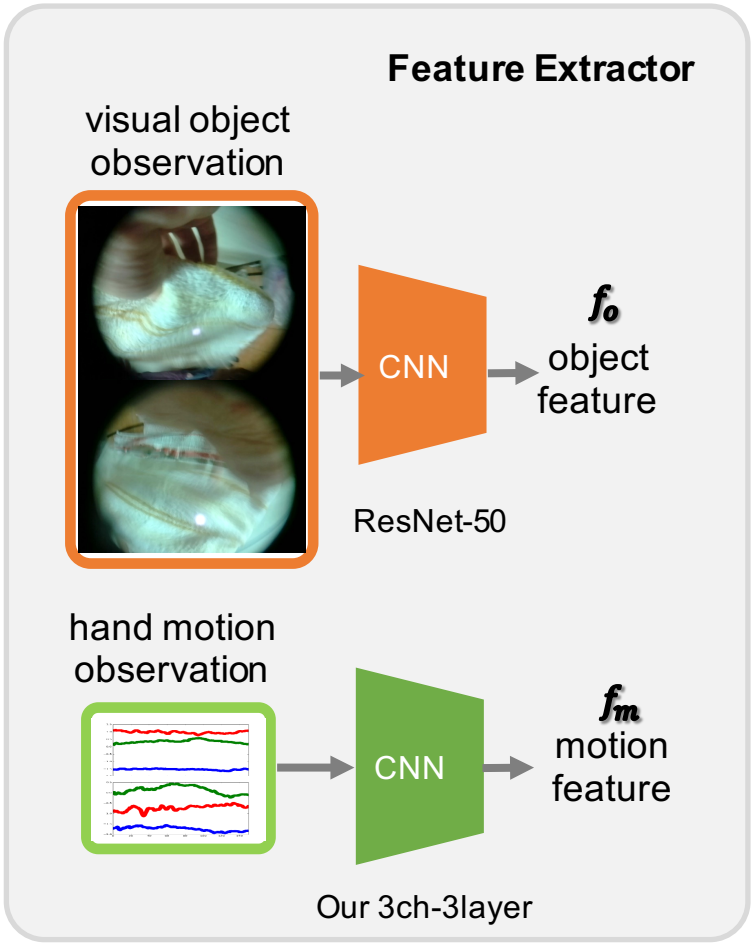
We collected one of the first daily intention dataset

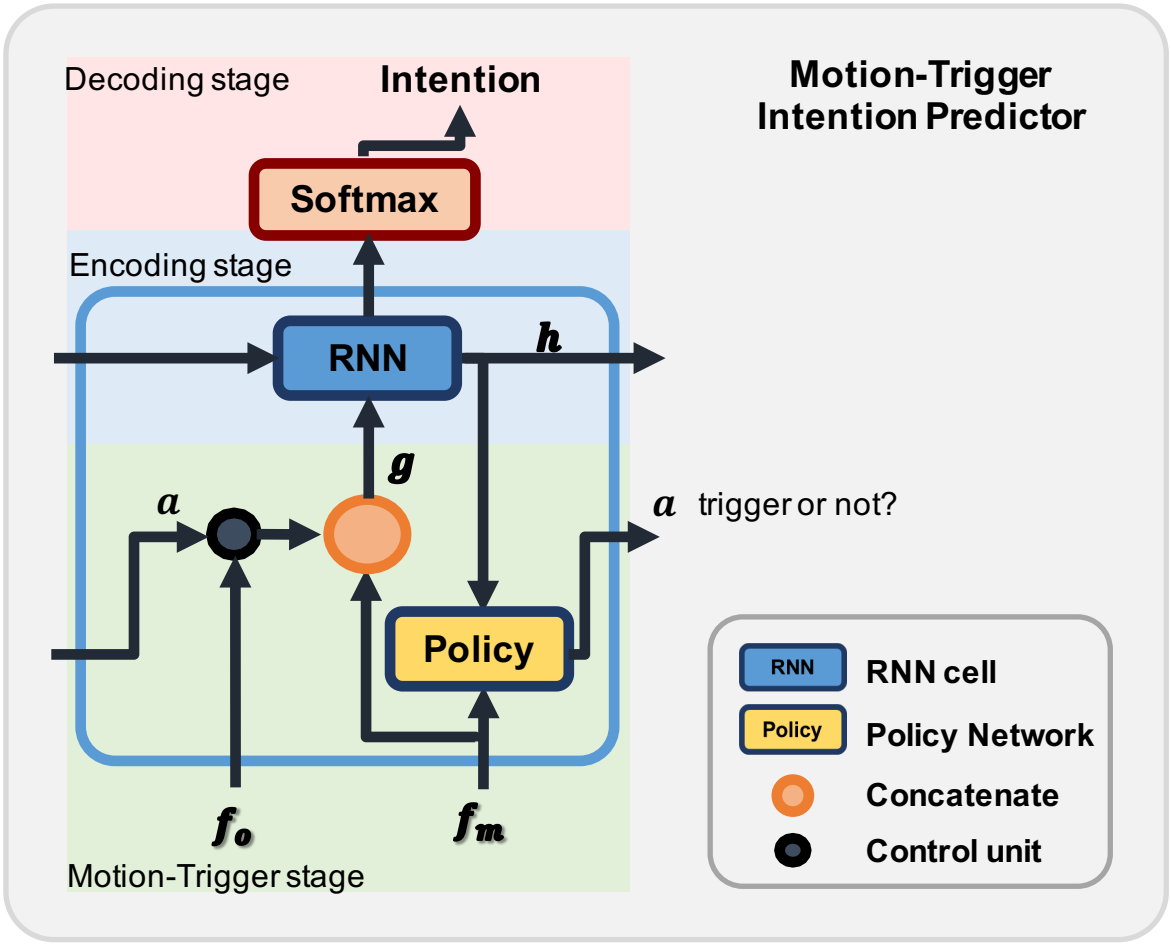


Our policy network effectively select the important images



Model Overview

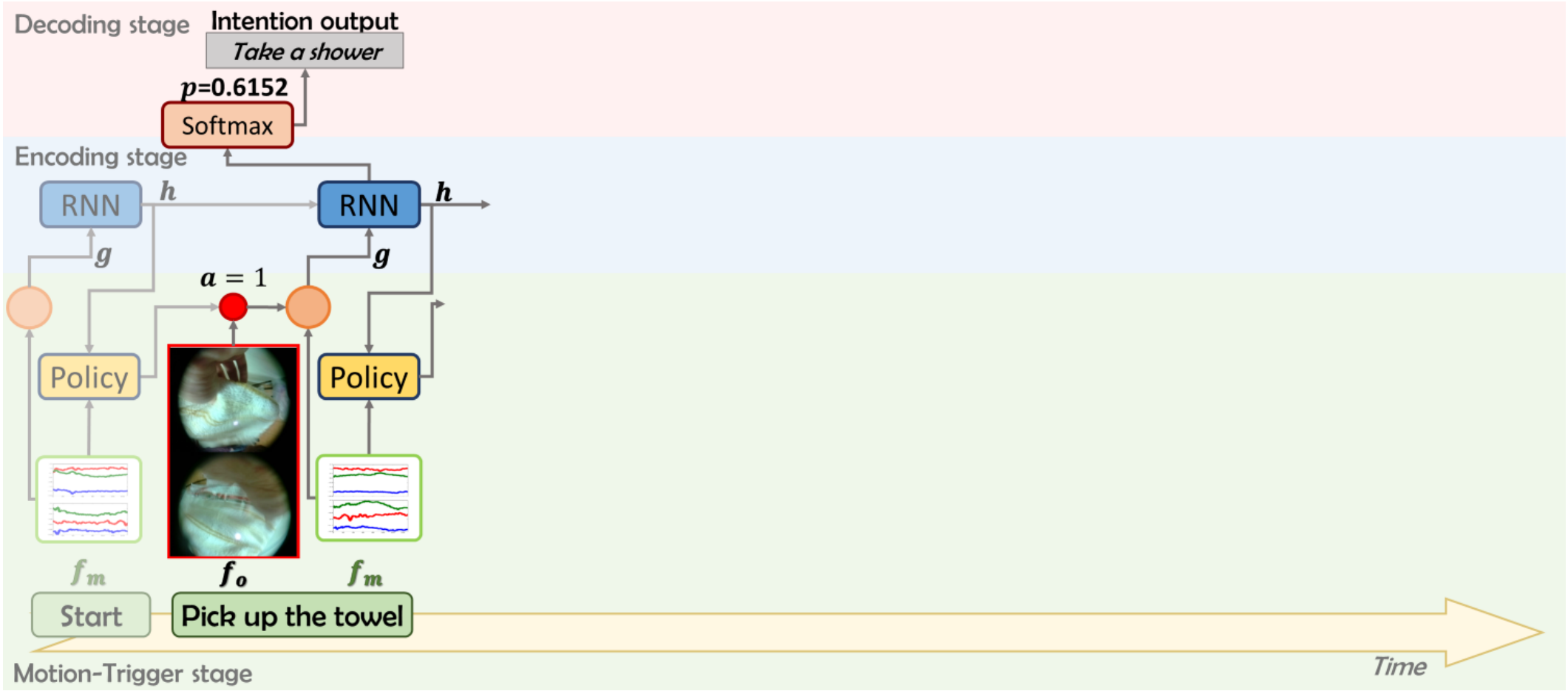




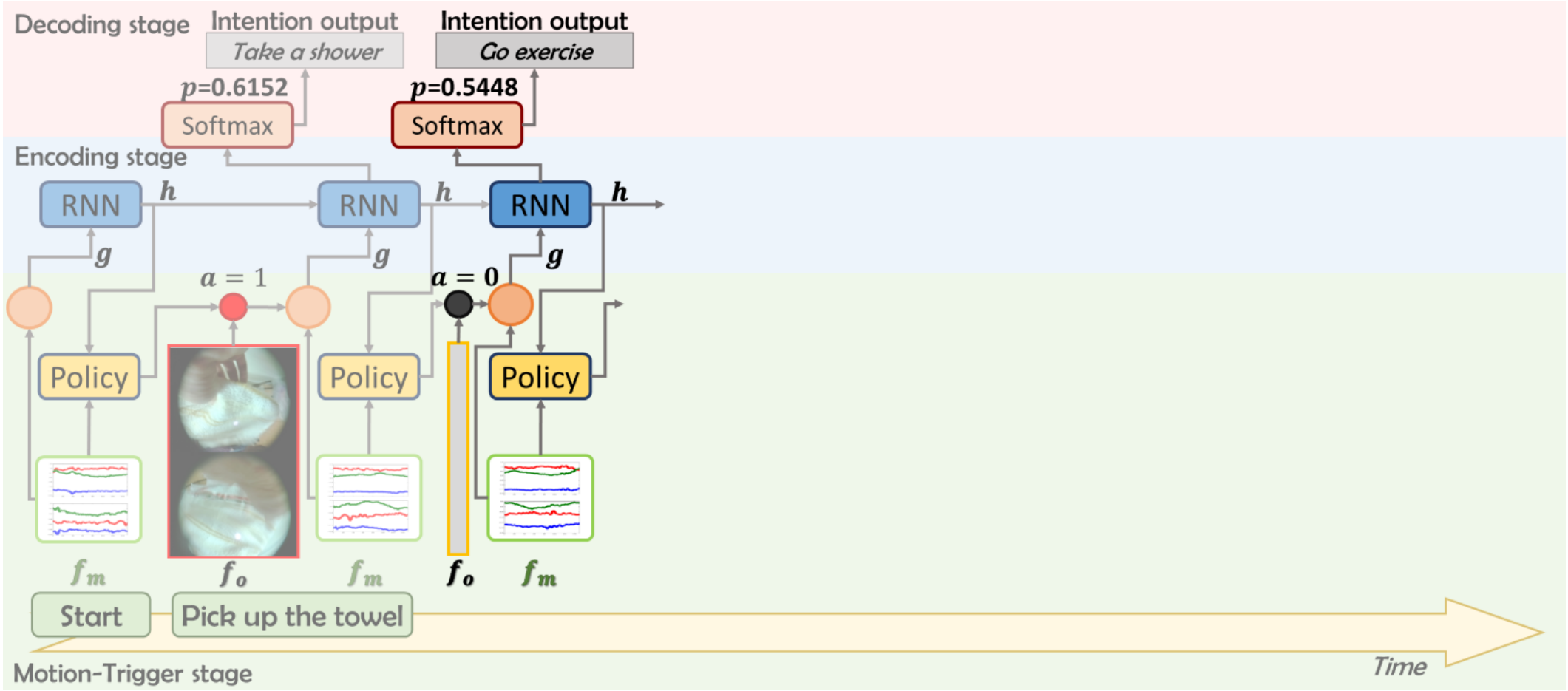
Approach


Decoding stage
Encoding stage
RNN
h
g
Policy
f_m
Start
Motion-Trigger stage
Time

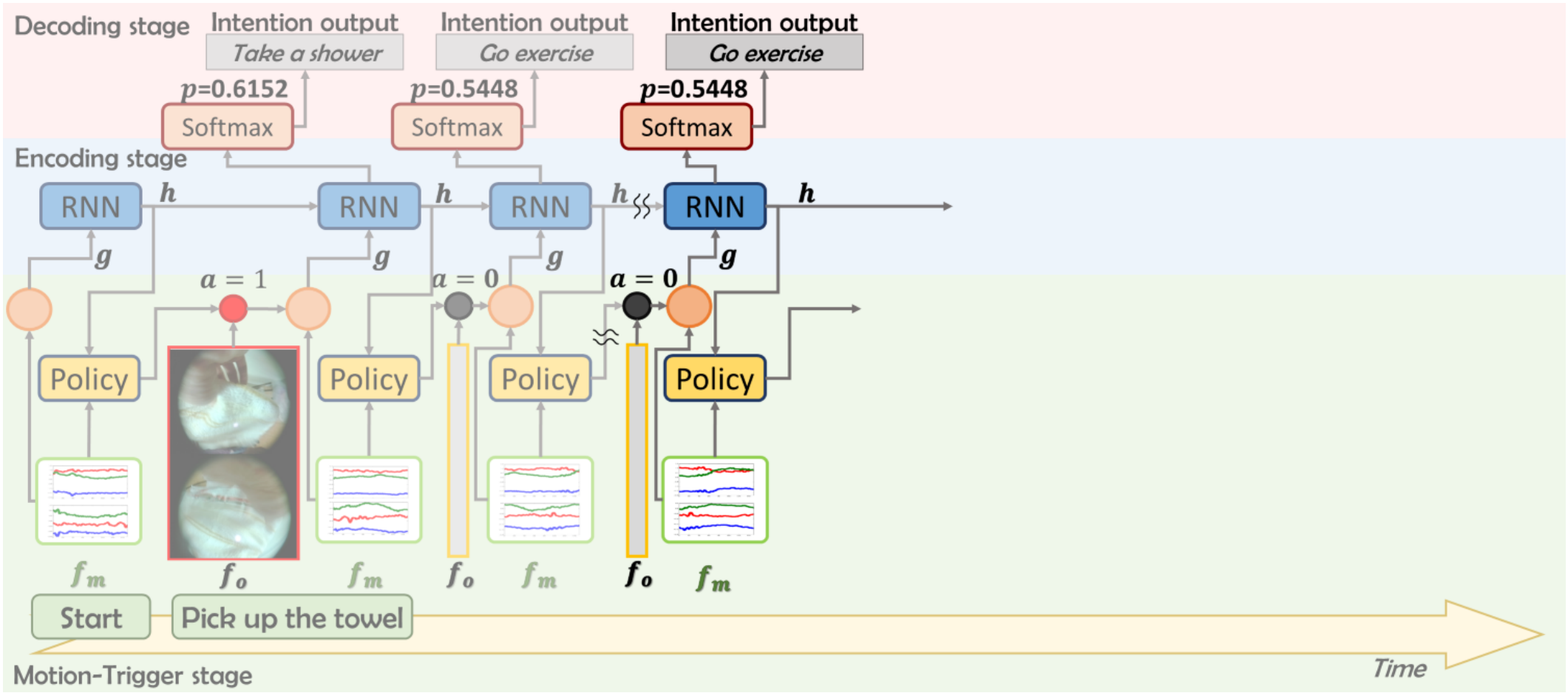
Approach


Decoding stage
Intention output
Take a shower
p=0.6152
Softmax
Encoding stage
RNN
h
RNN
h
g
g
a = 1
Policy
Policy
f_m
f_o
f_m
Start
Pick up the towel
Motion-Trigger stage
Time

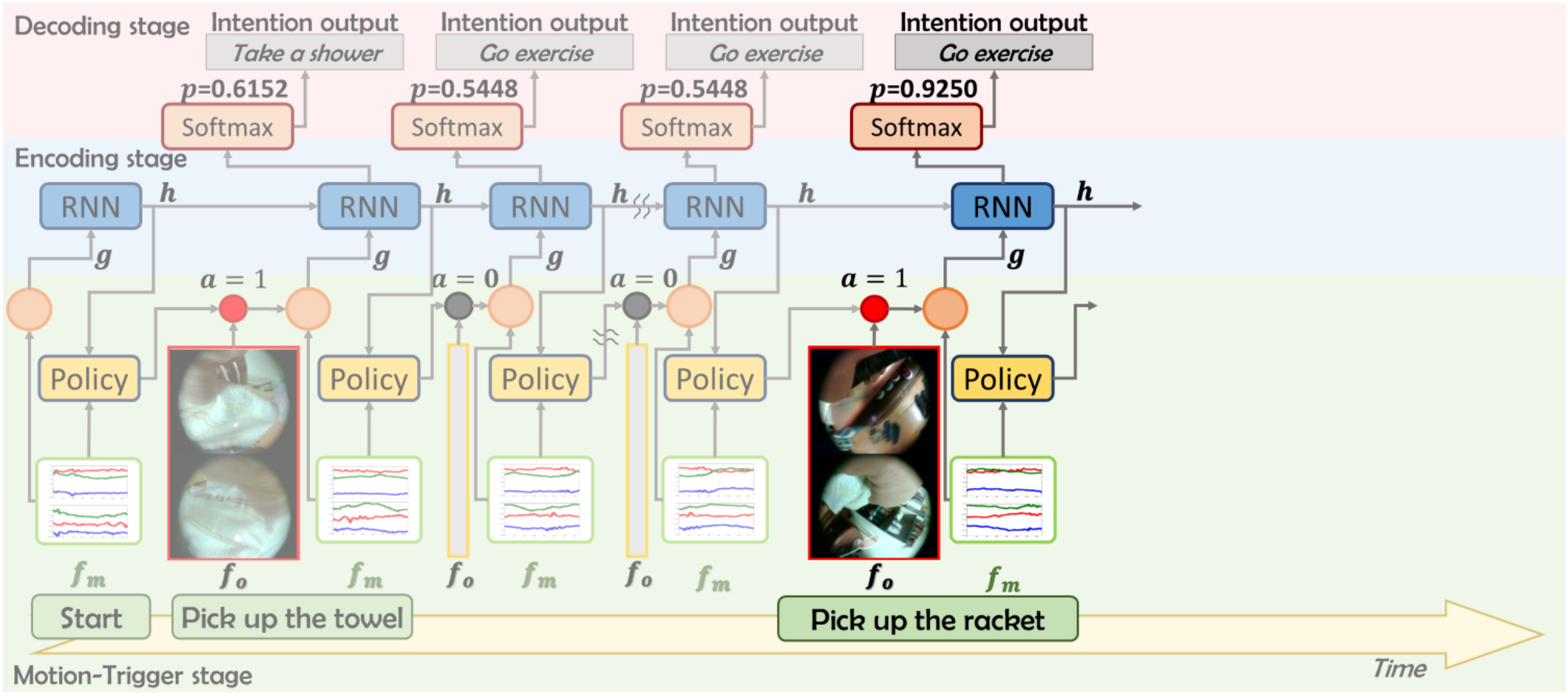
Approach


Decoding stage
Intention output
Take a shower
Intention output
Go exercise
p=0.6152
p=0.5448
Softmax
Softmax
Encoding stage
RNN
RNN
RNN
h
h
h
g
g
g
a = 1
a = 0
Policy
Policy
Policy
f_m
f_o
f_m
f_o
f_m
Start
Pick up the towel
Motion-Trigger stage
Time

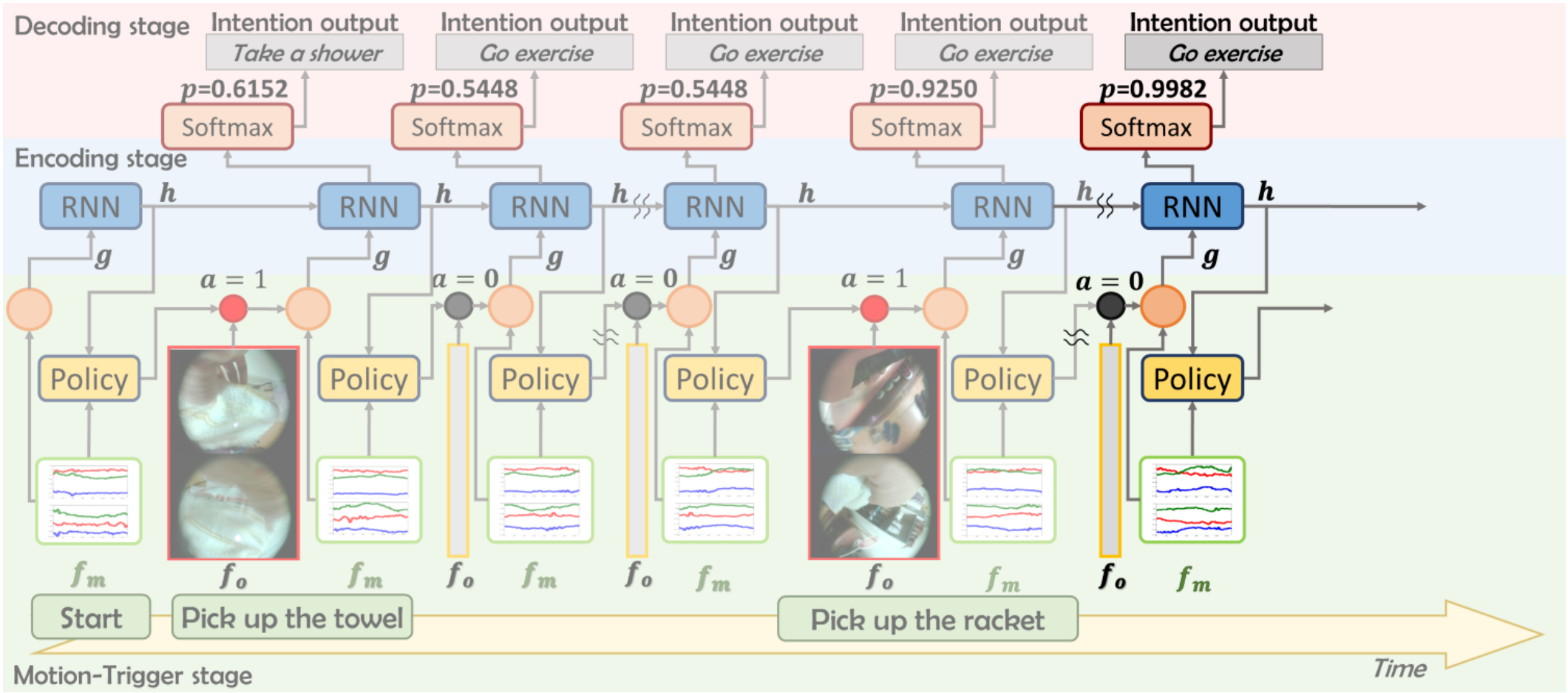
Approach


Decoding stage
Intention output
Take a shower
Intention output
Go exercise
Intention output
Go exercise
p=0.6152
p=0.5448
p=0.5448
Softmax
Softmax
Softmax
Encoding stage
RNN
RNN
RNN
RNN
h
g
a = 1
a = 0
a = 0
Policy
Policy
Policy
Policy
f_m
f_o
f_m
f_o
f_m
f_o
f_m
Start
Pick up the towel
Motion-Trigger stage
Time

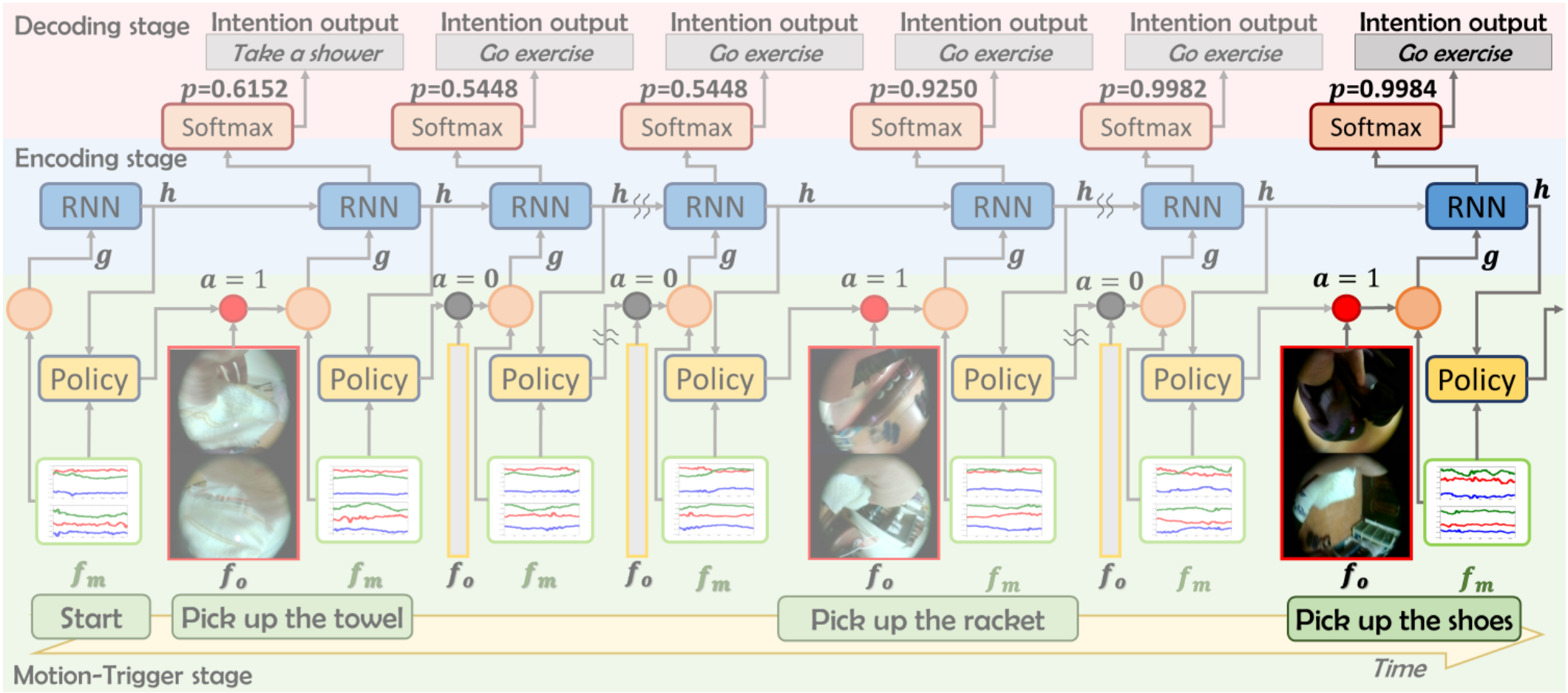
Approach


Decoding stage
Intention output
Take a shower
Intention output
Go exercise
Intention output
Go exercise
Intention output
Go exercise
p=0.6152
p=0.5448
p=0.5448
p=0.9250
Softmax
Softmax
Softmax
Softmax
Encoding stage
RNN
h
g
a = 1
a = 0
a = 0
a = 1
Policy
f_m
f_o
Start
Pick up the towel
Pick up the racket
Motion-Trigger stage
Time

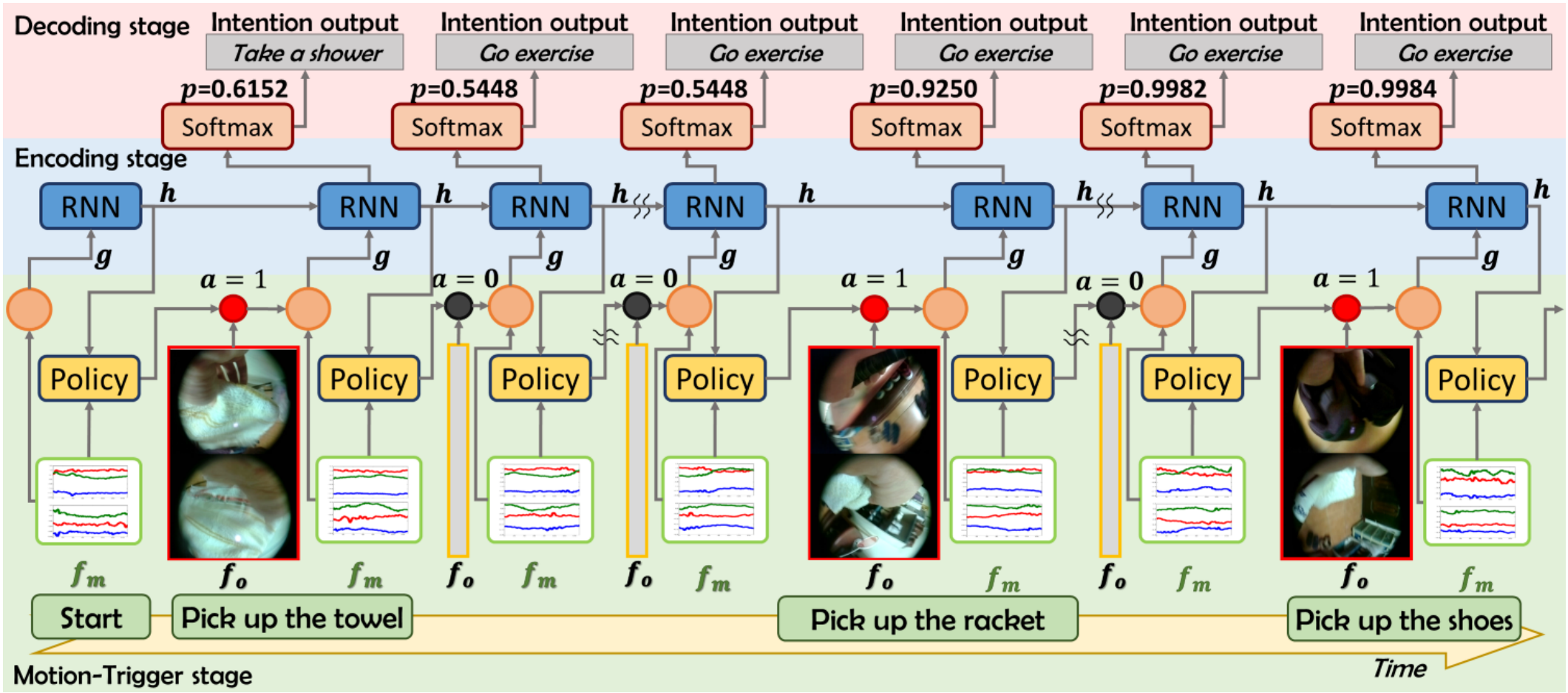
Approach


Decoding stage
Intention output
Take a shower
Go exercise
Go exercise
Go exercise
Go exercise
p=0.6152
p=0.5448
p=0.5448
p=0.9250
p=0.9982
Softmax
Softmax
Softmax
Softmax
Softmax
Encoding stage
RNN
h
g
a = 1
a = 0
a = 0
a = 1
a = 0
Policy
f_m
f_o
Start
Pick up the towel
Pick up the racket
Motion-Trigger stage
Time

Approach


Decoding stage
Intention output
Take a shower
Go exercise
Go exercise
Go exercise
Go exercise
Go exercise
p=0.6152
p=0.5448
p=0.5448
p=0.9250
p=0.9982
p=0.9984
Softmax
Encoding stage
RNN
h
g
a = 1
a = 0
Policy
f_m
f_o
Start
Pick up the towel
Pick up the racket
Pick up the shoes
Motion-Trigger stage
Time

Approach


Decoding stage
Intention output
Take a shower
Go exercise
Go exercise
Go exercise
Go exercise
Go exercise
p=0.6152
p=0.5448
p=0.5448
p=0.9250
p=0.9982
p=0.9984
Softmax
Encoding stage
RNN
h
g
a = 1
a = 0
Policy
f_m
f_o
Start
Pick up the towel
Pick up the racket
Pick up the shoes
Motion-Trigger stage
Time

RNN for Anticipation

- Fusing **visual object** $f_{o,t}$ and **hand motion** $f_{m,t}$ representations

$$g_t = \text{Emb}(W_{emb}, \text{con}(f_{m,t}, f_{o,t})) \,, \quad (1)$$

$$h_t = \text{RNN}(g_t, h_{t-1}) \,, \quad (2)$$

$$p_t = \text{Softmax}(W_y, h_t) \,, \quad (3)$$

$$y_t = \arg\max_{y \in \mathcal{Y}} p_t(y) \,, \quad (4)$$

- Fusing left and right hands $f_i = \text{con}(f_i^R, f_i^L)$, where $i \in \{o, m\}$
- Using anticipation loss (exponential loss) [1]

$$\sum_{t=1}^{T} L_t^A = \sum_{t=1}^{T} -\log p_t(y^{\text{gt}}) \cdot e^{\log(0.1)\frac{T-t}{T}} \,, \quad (5)$$

[1] A.Jain,H.S.Koppula,B.Raghavan,S.Soh,andA.Saxena. Car that knows before you do: Anticipating maneuvers via learning temporal driving models. In ICCV, 2015

RL-based Policy Network (PN)

- Policy network π (parameterized by W_p) to sample trigger or not
  - Observation $o_t = (h_t, f_{m,t})$ (RNN hidden representation + motion feature)
  - Action $a_t = \begin{cases} 1 & , trigger \\ 0 & , not\ trigger \end{cases}$

$$a_t = \arg\max_a \pi(a \mid (h_t, f_{m,t}); W_p) \in \{0, 1\} , \quad (6)$$

$$\hat{f}_{o,t+1} = (1 - a_t) \cdot \hat{f}_{o,t} + a_t \cdot f_{o,t+1}(I_{t+1}) , \quad (7)$$

Keep previous frame or process new frame

$$g_{t+1} = \mathrm{Emb}(W_{emb}, \mathrm{con}(f_{m,t+1}, \hat{f}_{o,t+1})) , \quad (8)$$

RNN input depends on PN

Design of Reward Function

Encourage **less triggered operations** and **correct intention anticipation**

$$R = \begin{cases} p_t(y^{\text{gt}}) \cdot R^{+} \cdot (1 - \frac{n}{T}), & \text{if } y = y^{\text{gt}} \\ p_t(y^{\text{gt}}) \cdot R^{-} \cdot \frac{n}{T}, & \text{if } y \neq y^{\text{gt}} \end{cases} \tag{9}$$

y^{gt}: ground truth intention; y: predicted intention

n: number of triggered operations in the video

p_t: probability of anticipated intention

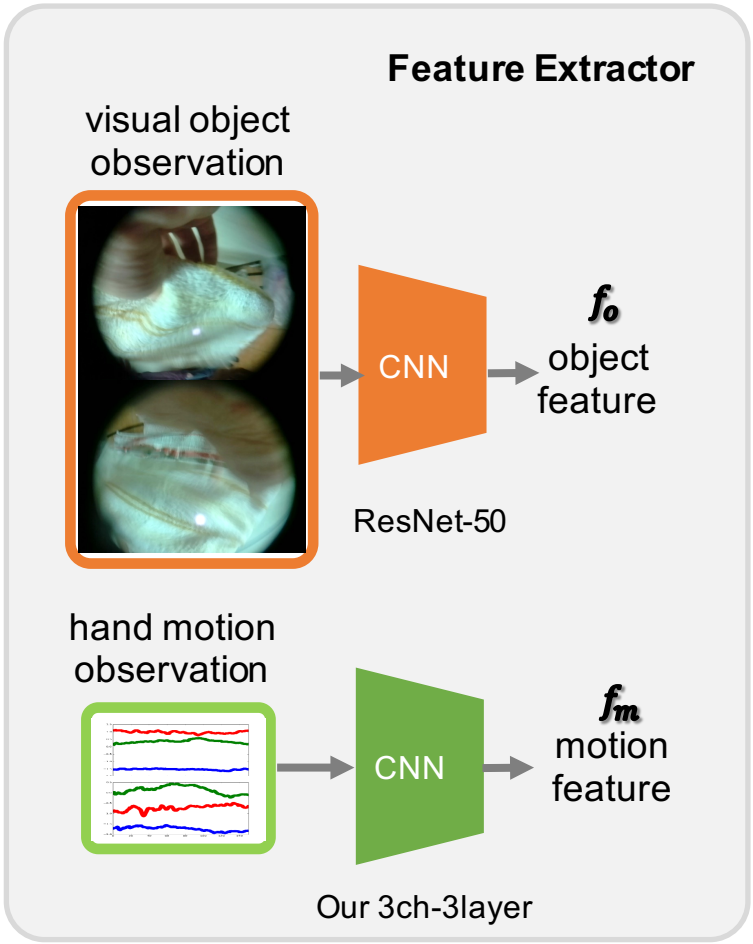
R^{+}/R^{-}: positive/negative reward for correct/incorrect intention anticipation (choose 100/-100)

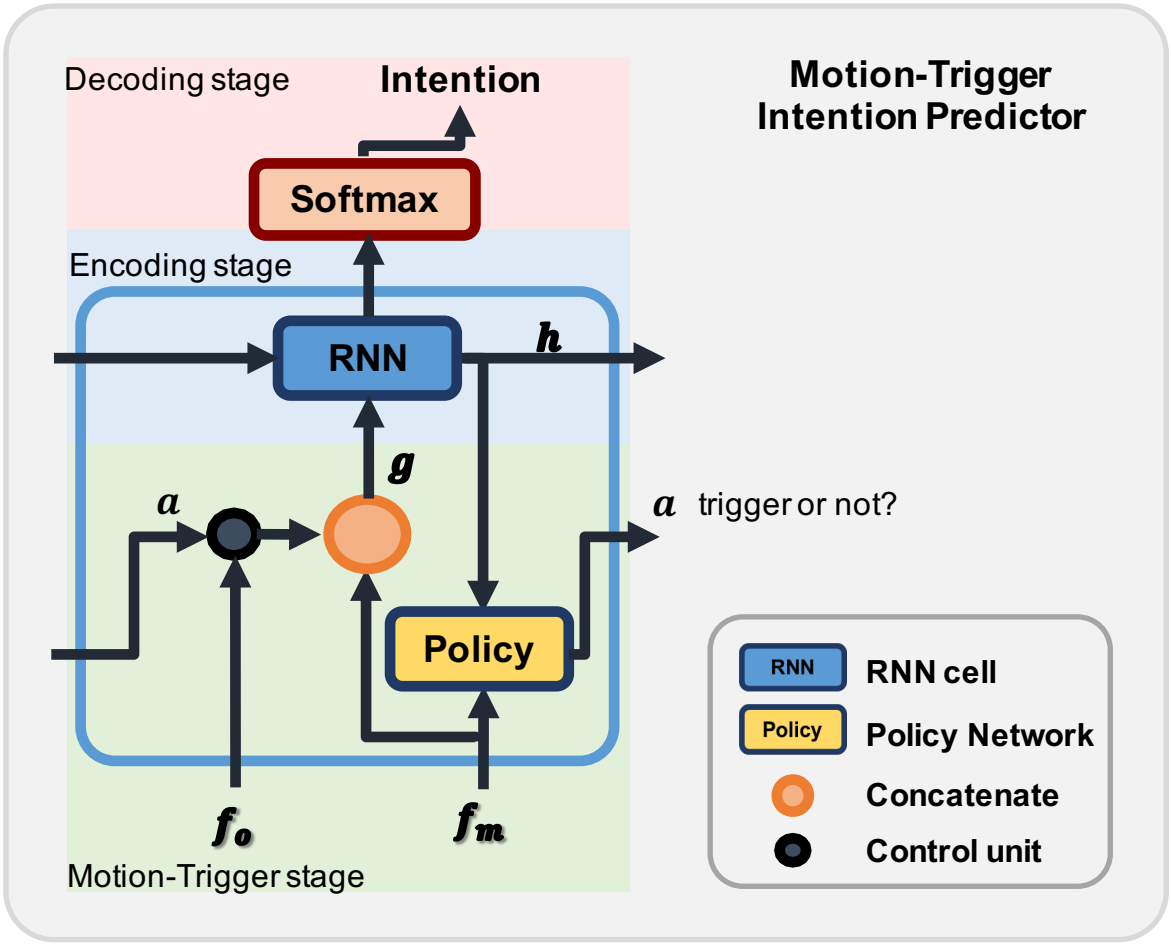
Policy Loss:

$$L^{P} = -\frac{1}{KT} \sum_{k=1}^{K} \sum_{t=1}^{T} \log(\pi(a_t^k \mid (h_t^k, f_{m,t}^k); W_p)) \cdot R_t^k \ ,$$

K: number of sequence; T: number of timesteps in the video

Model Overview



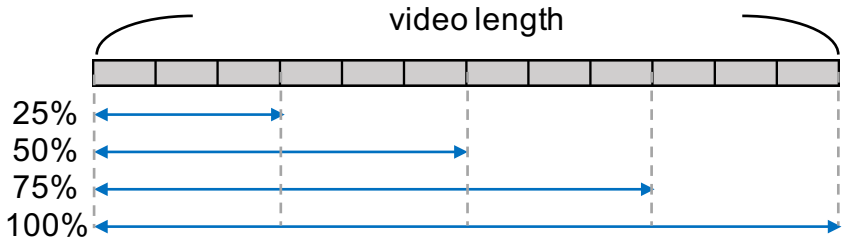


Motion Triggered Intention Anticipation

- PN parsimoniously triggers the process of visual observation (nearly 29%) while maintaining a high anticipation accuracy

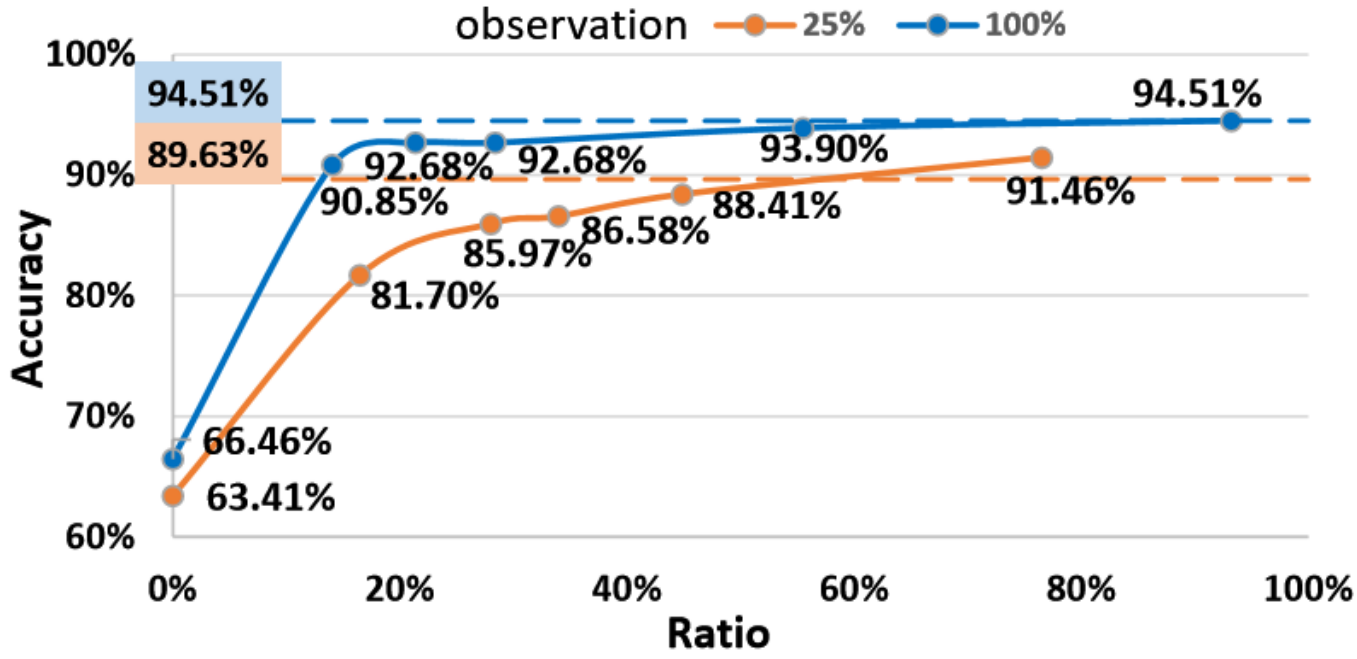
	User A				User B				User C			
	25%	50%	75%	100%	25%	50%	75%	100%	25%	50%	75%	100%
Con.	88.41%	90.24%	92.07%	93.29%	90.85%	92.68%	94.51%	95.12%	97.56%	97.56%	98.17%	98.17%
Mtr.	86.58%	90.24%	92.07%	92.68%	84.75%	88.41%	88.41%	90.85%	94.51%	96.34%	97.56%	97.56%
Ratio	34.00%	32.34%	30.72%	28.42%	31.13%	33.23%	30.88%	29.67%	33.40%	33.88%	30.89%	29.17%

Concatenation (**Con.**), Motion-triggered (**Mtr.**), triggered ratio (**Ratio**)



Adjust the threshold of motion triggers

- More triggers => higher accuracy



Video 1

Ground truth intention: brush the teeth

Predicted intention: brush the teeth

Trigger ratio: 14.29%

Monochrome: not triggered

Colored frame: triggered