Supplementary Material for "Complex Video Action Reasoning via Learnable Markov Logic Network"

Yang Jin[1,2], Linchao Zhu[3], Yadong Mu[1*]
[1]Peking University, [2]Baidu Research,[3]ReLER Lab, AAII, University of Technology Sydney
jiny@stu.pku.edu.cn, linchao.zhu@uts.edu.au, myd@pku.edu.cn

In this supplementary, we firstly present some experimental results about the late-fusion between our model and 3D convolution networks pre-trained on Kinetics-400 [2] in Section 1. More ablation studies with respect to hyper parameters in our action reasoning module are given in Section 2. Finally, for better demonstrating the explainability of our model, we provide the details of our user study and more visualization examples about the learned formula and weight in Section 3. Our code will be available at here[1].

1. Additional Experimental Results

The predicates in our logic formula are visual relationships in single video frame. To train the scene graph predictor, we adopt ResNet-101 as the backbone image feature extractor, which is pre-trained on ImageNet for a better weight initialization. The most advanced 3D deep models are usually pre-trained on Kinetics-400 [2] first, and then finetuned on the target dataset. The Kinetics-400 is a large video benchmark and its action categories are partially overlapped with Charades. We argue that such overlap may lead to overestimation of the mAP score of models due to the strong prior information in Kinetics-400. To validate this conjecture, we fused our predictions with an advanced 3D model [5] pre-trained on Kinetics-400, which achieved 42.5% mAP performance on Charades benchmark. In detail, we pass the output of [5] through a sigmoid activation function, and then add it to the confidence score given by our model as the final predictions. The experimental results are shown in Table 1.

Table 1. The experimental results of late fusions between our model and different advanced 3D deep models pre-trained on Kinetics-400. See main text for more explanation.

Network	Pre-train	Method	mAP (%)
R50-I3D-NL	K-400	3D CNN [4]	38.3
		LFB [5]	40.3
		Ours + LFB	**44.5**
R101-I3D-NL	K-400	3D CNN [4]	40.4
		LFB [5]	42.5
		X3D-XL [1]	43.4
		Ours + LFB	**45.2**

It can be seen that, after fused with the deep models pre-trained on Kinetics-400, our model achieved state-of-the-art action recognition performance on Charades (*e.g*., mAP score outperforms X3D-XL [1] by 1.8 %). Such a huge performance improvement demonstrates that our framework performs well on the novel action categories on Charades and is significantly complementary with deep models that are pre-trained on Kinetics-400.

2. More Ablation Studies

In this section, more ablation studies about the hyper parameters in our action reasoning module are conducted, mainly the number of relationship predicates T in each formula and the total formula number k of per action category.

In the experiments, we adopt a multi-size sliding window to generate short video snippets from a whole video v, where the kernal size $L = \{25, 50, 75, ..., v_L\}$ and v_L is the length of video v. For each snippet, only $M = 5$ frames are uniformly sampled to predict the corresponding scene graphs. Then, we perform the probability inference on the sampled frames. Here, we explore the effect of formula length and the the number of formula sampled from our rule policy network for action recognition. The detailed experimental results on Charades [3] and CAD-120 [6] are presented in Table 2.

From the results shown in Table 2, we can find that the best formula length setting is $T = 3$. In addition, when the formula number k increases, the overall mAP performance becomes better on both benchmarks. This phenomenon can be intuitively understood since more formulae can better generalize the underlying temporally-evolving patterns of specific action, and the weight learning on MLN will assign a lower weight for noisy formulae. In spite of better mAP performance, we finally adopt $k = 20$ in all our experiments

[*] Yadong Mu is the corresponding author. Part of the work was performed when Yang Jin was an intern at Baidu Research.

[1] https://github.com/rain1011/VideoMLN

Table 2. The ablation studies of different settings about formula length T and formula number k for action recognition on Charades [3] and CAD-120 [6]. See main text for more explanation.

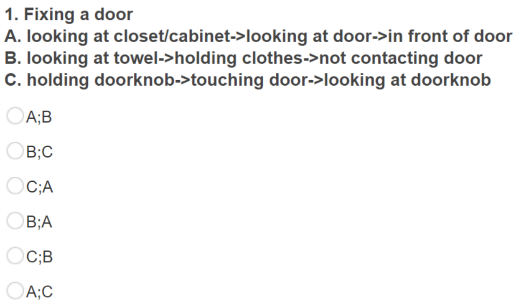
T	k	mAP on Charades(%)	mAR on CAD-120
3	10	36.7	0.80
	20	38.4	0.83
	50	**38.7**	**0.85**
5	10	36.1	0.78
	20	37.3	0.81
	50	**37.5**	**0.82**

to balance accuracy and computational efficiency.

3. User Study and Visualization results

3.1. Details of user study

To further demonstrate the explainability of our method, we conduct a user study to evaluate the generated formulae. In our framework, the formulae with a higher weight should provide more convincing evidence to recognize an action. To validate this, the range of formula weight given by our model is uniformly trisected, where the formulae are accordingly denoted as *good*, *neutral* and *bad* ones. For a subset of 20 actions on Charades, we randomly sample 1 formula from each type. Then, we design a questionnaire (See figure 1) to invite 21 subjects to rank the shuffled formulae based on the relevance to the action.



1. Fixing a door
A. looking at closet/cabinet->looking at door->in front of door
B. looking at towel->holding clothes->not contacting door
C. holding doorknob->touching door->looking at doorknob

○A;B

○B;C

○C;A

○B;A

○C;B

○A;C

Figure 1. An example of our questionnaire in user study. Participants need to pick the formulae that are most relevant and least relevant to the action. For example, if you consider the most relevant one to action *Fixing a door* is A and the leaset one is B, then you should pick the first option.

The detailed results are given in Table 3. Each row represents the formula type given by our model (based on the learned weight), and each row represents the types given by participants. From the results shown in Table 3, we can observe that the results show high consistency between the learned weight and human commonsense. For example, 78.75% *good* formulae are still marked as *good*, and 55.5% *bad* formulae are still marked as *bad*.

Table 3. The detailed experimental results of user study regarding the explainability (*i.e.*, being human-friendly) of the learned formulae. The formulae are categoried as *good*, *neutral* and *bad*. See main text for more explanation.

	good	*neutral*	*bad*
good	**78.75%**	15.5%	5.75%
neutral	15.5%	**45.75%**	38.75%
bad	5.75%	38.75%	**55.5%**

3.2. More visualization examples

We present more specific examples in this section. In detail, for each action category, the corresponding formulae and their weights given by our model are illustrated in Figures 2 to 4. Formulae shown in green represent those higher weights and the orange ones are with lowest weights. It can be observed that the formulae with higher weights often provide better reasoning for the interested actions, which demonstrates the explainability and diversity of our generated formulae.

References

[1] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–213, 2020. 1

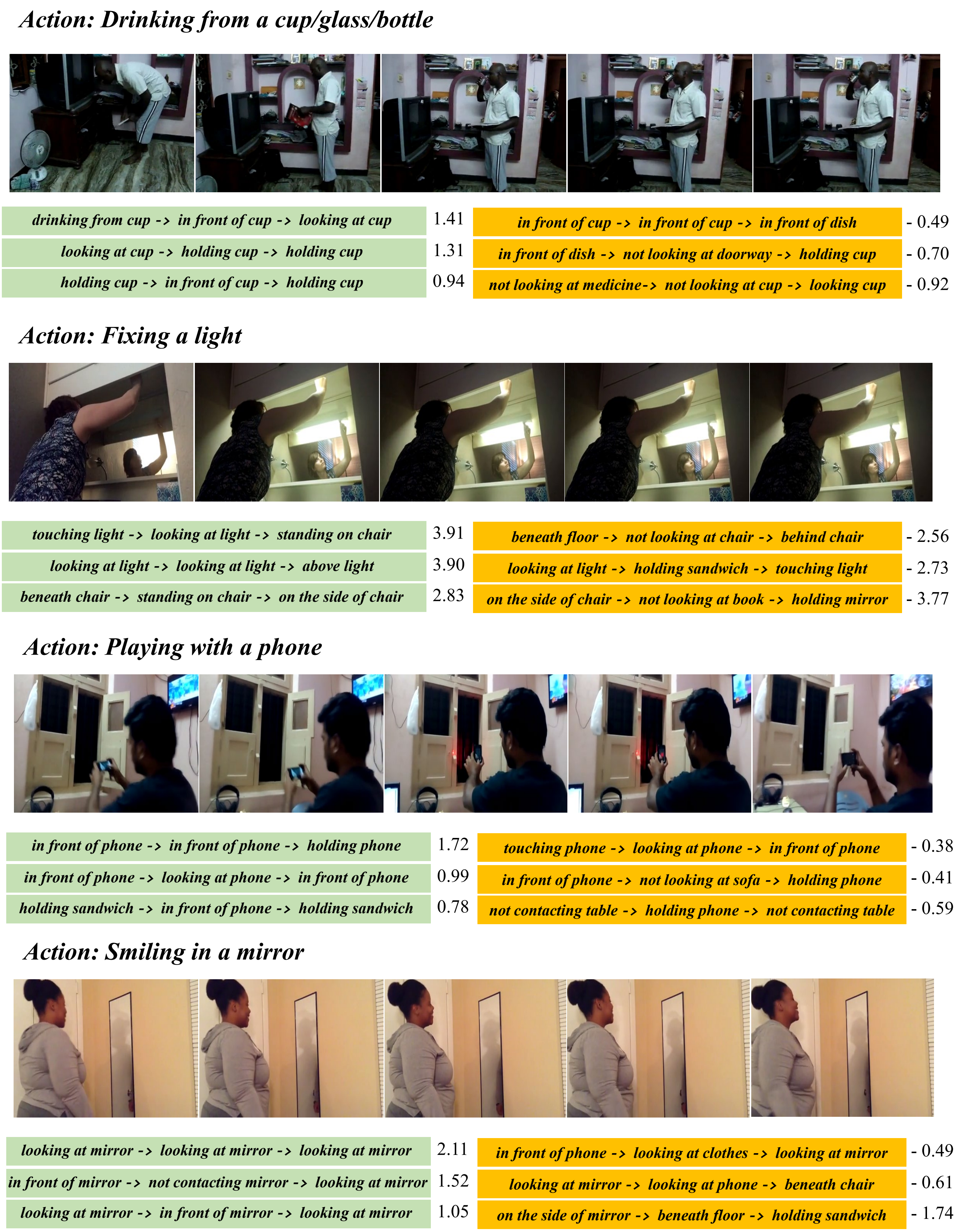
[2] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[3] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016. 1, 2

[4] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 1

[5] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019. 1

[6] Tao Zhuo, Zhiyong Cheng, Peng Zhang, Yongkang Wong, and Mohan Kankanhalli. Explainable video action reasoning via prior knowledge and state transitions. In *Proceedings of the 27th acm international conference on multimedia*, pages 521–529, 2019. 1, 2



Action: Drinking from a cup/glass/bottle

drinking from cup -> in front of cup -> looking at cup	1.41	*in front of cup -> in front of cup -> in front of dish*	- 0.49
looking at cup -> holding cup -> holding cup	1.31	*in front of dish -> not looking at doorway -> holding cup*	- 0.70
holding cup -> in front of cup -> holding cup	0.94	*not looking at medicine-> not looking at cup -> looking cup*	- 0.92

Action: Fixing a light

touching light -> looking at light -> standing on chair	3.91	*beneath floor -> not looking at chair -> behind chair*	- 2.56
looking at light -> looking at light -> above light	3.90	*looking at light -> holding sandwich -> touching light*	- 2.73
beneath chair -> standing on chair -> on the side of chair	2.83	*on the side of chair -> not looking at book -> holding mirror*	- 3.77

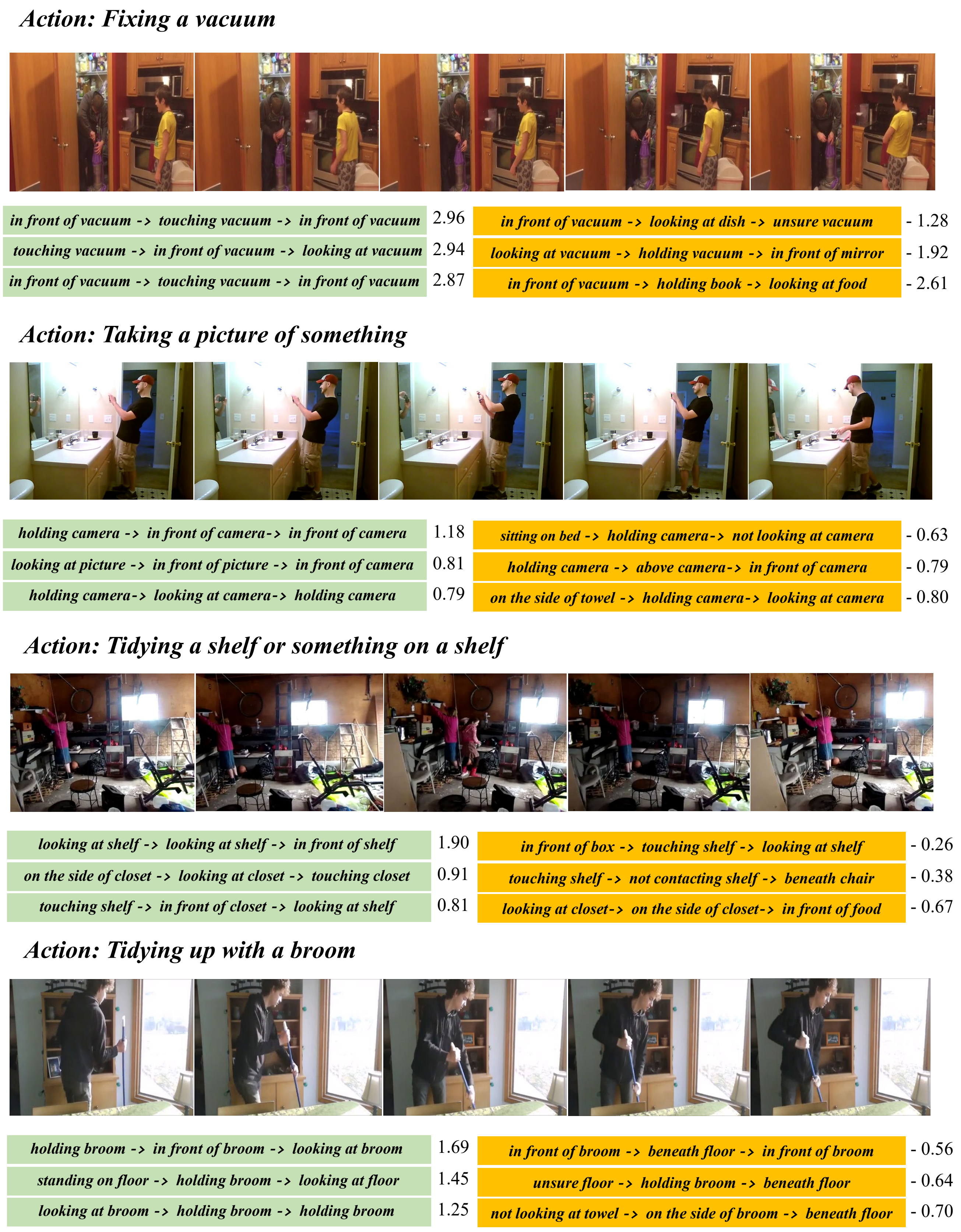
Action: Playing with a phone

in front of phone -> in front of phone -> holding phone	1.72	*touching phone -> looking at phone -> in front of phone*	- 0.38
in front of phone -> looking at phone -> in front of phone	0.99	*in front of phone -> not looking at sofa -> holding phone*	- 0.41
holding sandwich -> in front of phone -> holding sandwich	0.78	*not contacting table -> holding phone -> not contacting table*	- 0.59

Action: Smiling in a mirror

looking at mirror -> looking at mirror -> looking at mirror	2.11	*in front of phone -> looking at clothes -> looking at mirror*	- 0.49
in front of mirror -> not contacting mirror -> looking at mirror	1.52	*looking at mirror -> looking at phone -> beneath chair*	- 0.61
looking at mirror -> in front of mirror -> looking at mirror	1.05	*on the side of mirror -> beneath floor -> holding sandwich*	- 1.74

Figure 2. Some examples of the learned formula and corresponding weights given by our proposed framework.



Action: Fixing a vacuum

in front of vacuum -> touching vacuum -> in front of vacuum	2.96	*in front of vacuum -> looking at dish -> unsure vacuum*	- 1.28
touching vacuum -> in front of vacuum -> looking at vacuum	2.94	*looking at vacuum -> holding vacuum -> in front of mirror*	- 1.92
in front of vacuum -> touching vacuum -> in front of vacuum	2.87	*in front of vacuum -> holding book -> looking at food*	- 2.61

Action: Taking a picture of something

holding camera -> in front of camera-> in front of camera	1.18	*sitting on bed -> holding camera-> not looking at camera*	- 0.63
looking at picture -> in front of picture -> in front of camera	0.81	*holding camera -> above camera-> in front of camera*	- 0.79
holding camera-> looking at camera-> holding camera	0.79	*on the side of towel -> holding camera-> looking at camera*	- 0.80

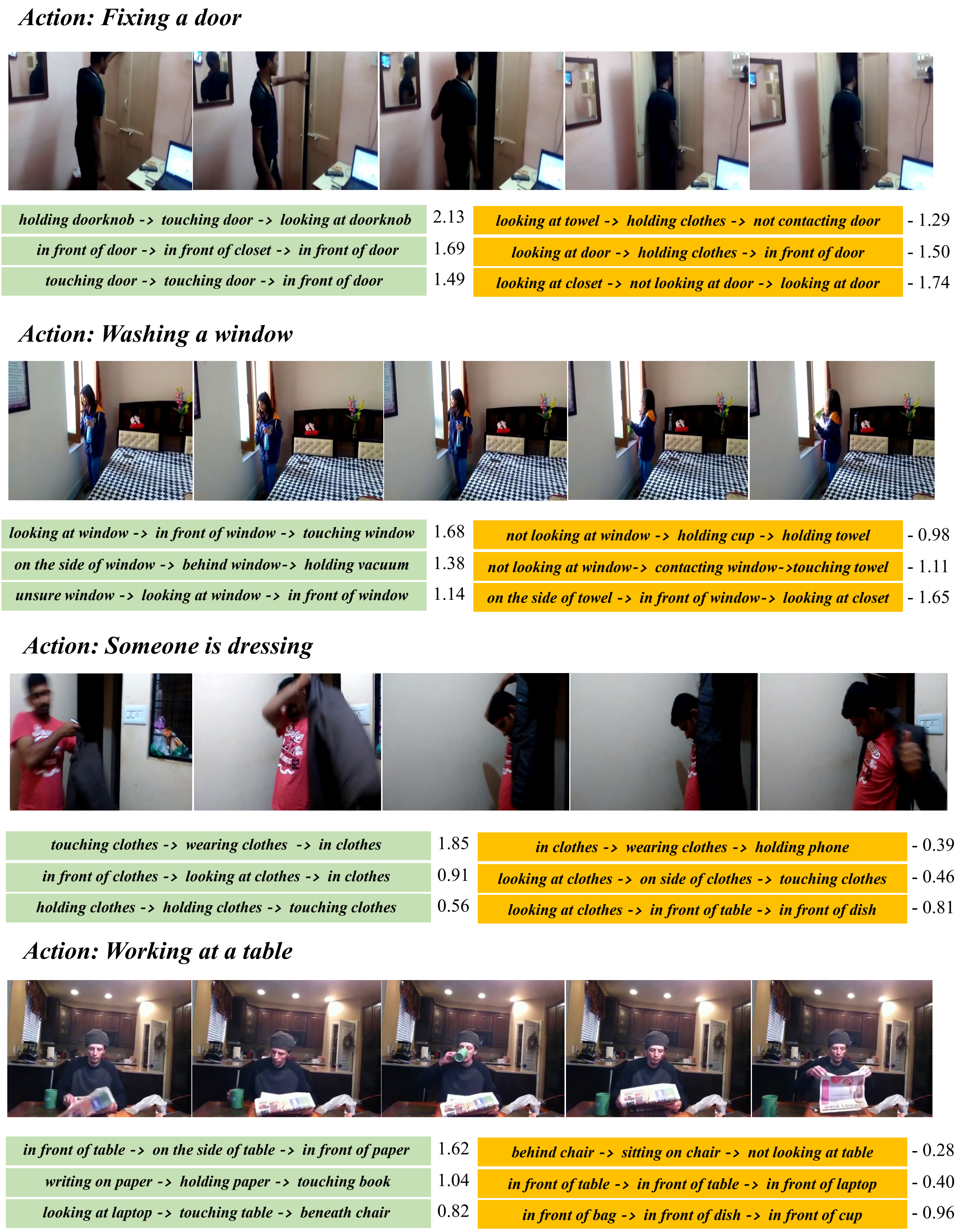
Action: Tidying a shelf or something on a shelf

looking at shelf -> looking at shelf -> in front of shelf	1.90	*in front of box -> touching shelf -> looking at shelf*	- 0.26
on the side of closet -> looking at closet -> touching closet	0.91	*touching shelf -> not contacting shelf -> beneath chair*	- 0.38
touching shelf -> in front of closet -> looking at shelf	0.81	*looking at closet-> on the side of closet-> in front of food*	- 0.67

Action: Tidying up with a broom

holding broom -> in front of broom -> looking at broom	1.69	*in front of broom -> beneath floor -> in front of broom*	- 0.56
standing on floor -> holding broom -> looking at floor	1.45	*unsure floor -> holding broom -> beneath floor*	- 0.64
looking at broom -> holding broom -> holding broom	1.25	*not looking at towel -> on the side of broom -> beneath floor*	- 0.70

Figure 3. Some examples of the learned formula and corresponding weights given by our proposed framework.



Action: Fixing a door

holding doorknob -> touching door -> looking at doorknob	2.13	*looking at towel -> holding clothes -> not contacting door*	- 1.29
in front of door -> in front of closet -> in front of door	1.69	*looking at door -> holding clothes -> in front of door*	- 1.50
touching door -> touching door -> in front of door	1.49	*looking at closet -> not looking at door -> looking at door*	- 1.74

Action: Washing a window

looking at window -> in front of window -> touching window	1.68	*not looking at window -> holding cup -> holding towel*	- 0.98
on the side of window -> behind window-> holding vacuum	1.38	*not looking at window-> contacting window->touching towel*	- 1.11
unsure window -> looking at window -> in front of window	1.14	*on the side of towel -> in front of window-> looking at closet*	- 1.65

Action: Someone is dressing

touching clothes -> wearing clothes -> in clothes	1.85	*in clothes -> wearing clothes -> holding phone*	- 0.39
in front of clothes -> looking at clothes -> in clothes	0.91	*looking at clothes -> on side of clothes -> touching clothes*	- 0.46
holding clothes -> holding clothes -> touching clothes	0.56	*looking at clothes -> in front of table -> in front of dish*	- 0.81

Action: Working at a table

in front of table -> on the side of table -> in front of paper	1.62	*behind chair -> sitting on chair -> not looking at table*	- 0.28
writing on paper -> holding paper -> touching book	1.04	*in front of table -> in front of table -> in front of laptop*	- 0.40
looking at laptop -> touching table -> beneath chair	0.82	*in front of bag -> in front of dish -> in front of cup*	- 0.96

Figure 4. Some examples of the learned formula and corresponding weights given by our proposed framework.