Language Model Guided Interpretable Video Action Reasoning

Ning Wang[1], Guangming Zhu[1], HS Li[1], Liang Zhang[1], Syed Afaq Ali Shah[2], Mohammed Bennamoun[3]
[1]Xidian University, [2]Edith Cowan University, [3]University of Western Australia
{ningwang, hsli}@stu.xidian.edu.cn, {gmzhu, liangzhang}@xidian.edu.cn,
afaq.shah@ecu.edu.au, mohammed.bennamoun@uwa.edu.au

In this supplementary material, we first describe the inference procedure of the proposed model. Then, we present some additional experimental results. Finally, to better demonstrate the interpretability of our model, we provide an interpretable analysis of prediction failure cases and additional visualization examples.

1. Inference Procedure

In the inference stage, the proposed method only considers the video model, which takes RGB as input to directly predict action and provide explanations. Figure 1 shows the detailed reasoning process of the video model. First, given a video, a visual encoder encodes the visual relation representations of all person-object pairs. Second, the visual relation representations are fed to the DT-Former module, which selects out key relations and models relations transitions to get the action category. Finally, key visual relationship representations are mapped into the joint embedding space, and the semantic representation closest to the visual representation is considered as its semantic label. Semantic-level relation transitions explicitly explain the action reasoning process.

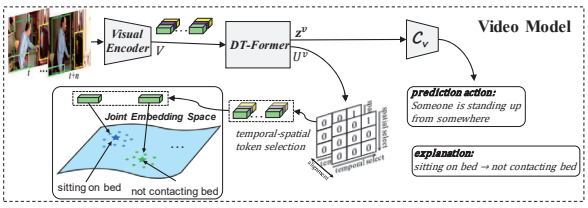

Figure 1. The inference procedure of the video model. Given an input video, the video model selects key relations, and models key relation transitions to identify actions. The key relations are mapped into a joint embedding space to enable explicit explanation of the action reasoning process.

2. Additional Experimental Results

We used the ResNet-101 network [5] as an object detector and visual human-object relation feature extractor in video frames. The ResNet-101 network is pre-trained on the ImageNet dataset [2]. As reported in the sixth row of Table 1, our model achieved state-of-the-art action recognition performance on Charades (e.g., mAP score outperforms OR2G [7] by 1.1 %). As OR2G has been used for comparison, we also adopt the R01-I3D-NL network [9] as the video feature extractor. The R01-I3D-NL network are pre-trained on Kinetics-400 first, and then fine-tuned on the target dataset. The Kinetics-400 is a large video benchmark and its action categories are partially overlapped with Charades. Such overlap may lead to overestimation of the mAP score of models due to the strong prior information in Kinetics-400. Thus, we fused our predictions with the R01-I3D-NL network pre-trained on Kinetics-400, and the model achieved 45.1% mAP performance on Charades benchmark. To achieve this, we process the output from the I3D-NL network with a sigmoid activation function, combining it with our model's confidence score to determine the final predictions. Such notable performance enhancement indicates that our framework is effective in the new action

Table 1. Experimental results of different methods pre-trained on ImageNet or Kinetics-400. See main text in the manuscript for detailed explanation.

Pre-train	Method	Backbone	mAP
ImageNet	Two-stream [8]	R101	18.6
	ActionVLAD [4]	R101	21.0
	TRN [10]	R101	25.2
	OR2G [7]	R101	34.2
	Ours	R101	**35.3**
Kinetics-400	I3D [1]	R101	32.9
	Timeception [6]	R101	37.2
	I3D-NL [9]	R101-I3D-NL	37.5
	SlowFast [3]	R101	42.1
	OR2G [7]	R101-I3D-NL	44.9
	Ours	R101-I3D-NL	**45.1**
	Ours Oracle	R101-I3D-NL	67.4
	OR2G Oracle [7]	R101-I3D-NL	**67.5**

Table 2. This table outlines the division of the Charades dataset into five subsets, ensuring that there is no overlap between the scenes used for training and testing.

Subdataset	Training Scene	Test Scene
Scenario1	Stairs,Laundry room,Home Office, Hallway,Bedroom,Pantry,Dining room,Entryway	Living room,Closet,Kitchen,Bathroom, Garage,Recreation room,Basement,Other
Scenario2	Laundry room,Bathroom,Pantry,Closet, Entryway,Recreation room,Garage,Other	Bedroom,Living room,Kitchen,Home Office, Hallway,Stairs,Basement,Dining room
Scenario3	Stairs,Laundry room,Bedroom,Basement, Bathroom,Entryway,Recreation room,Other	Living room,Closet,Kitchen,Home Office, Garage,Hallway,Pantry,Dining room
Scenario4	Kitchen,Stairs,Laundry room,Home Office, Bedroom,Bathroom,Pantry,Dining room	Living room,Closet,Garage,Hallway, Recreation room,Entryway,Basement,Other
Scenario5	Kitchen,Laundry room,Hallway,Basement, Dining room,Living room,Closet,Other	Bedroom,Home Office,Bathroom,Garage, Stairs,Recreation room,Entryway,Pantry

category of Charades and significantly augments deep models pretrained on Kinetics-400. In particular, our method achieves competitive results with the Oracle version, even without the use of ground-truth scene graphs.

3. Subdataset Division Details

Our framework's resilience to domain shifts is showcased by dividing the Charades dataset into five subsets. As Table 2 indicates, the training scenes for each subset are distinct from those used for testing. Despite variations in scenes between training and testing datasets, our method maintains strong performance.

4. Qualitative Visualization Results

4.1. Special Failure Case

Our experiments on the Charades and CAD-120 datasets validate our method's capability to identify important relationships by evaluating the contribution of each token to the classification task. However, due to the data-driven nature of our approach, which relies on learning action-specific relation transition patterns automatically, it occasionally discards significant tokens. An instance of this can be seen in Figure 2, where our system erroneously omits a crucial frame, leading to incorrect relationship sequencing and misclassification of actions as "null": the relation transition to be **'apart'** → **'apart'** instead of **'apart'** → **'contacting'** → **'apart'**, thus actions are misidentified as "null". Manual oversight, coupled with enhancing the dataset with additional annotations for relation transition patterns per action category, could mitigate such errors. We are further enhancing the dataset by adding new annotations. These annotations will include one or more relation transition patterns for each action category, aiming to improve the accuracy and depth of analysis in our model.

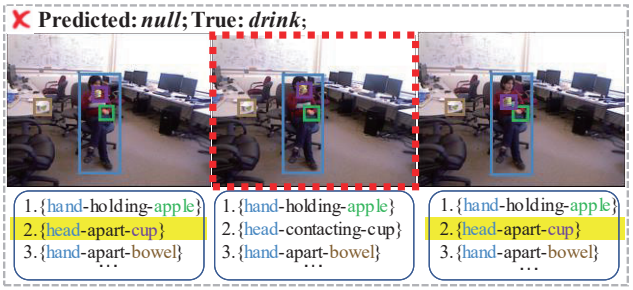

Figure 2. An illustration of a specific failure instance in our approach. The key frame, marked by the red dotted box, is incorrectly discarded by our method. This frame is essential for correctly identifying the action 'drink'. The relationships shown in the figure are those that our method has selected.

4.2. Additional Visualization Results

Figure 3 illustrates the process of selecting key relationships. First, the temporal token selection module estimates the contribution of each frame to the recognition result, and frames with low contributions (*i.e.*, scores below 0.5) are excluded. In this example, frames 1, 4, 5, 7, 9, 12, 13, and 14 are omitted, and their tokens are not used in later calculations. Then, a spatial token selection module evaluates each token within visual relationship pairs for their contribution to recognition. Tokens scoring under 0.5 are also excluded. Finally, the remaining tokens are then assigned semantic labels from a joint embedding space to serve as explanations. In the case of human-doorway interactions, only tokens from frames 3 and 4 are kept. The token from frame 3 is labeled **'in doorway'**, while the token from frame 4 is labeled **'behind doorway'**. The consequences **'in doorway'** → **'behind doorway'** leads to the inference of the action "Walking through a doorway". In a similar vein, for human-box interactions, tokens from frames 3 and 6 labeled **'holding box'**, the token from frame 8 is labeled **'touching box'**, and those from frames 10 and 11 labeled **'not con-**

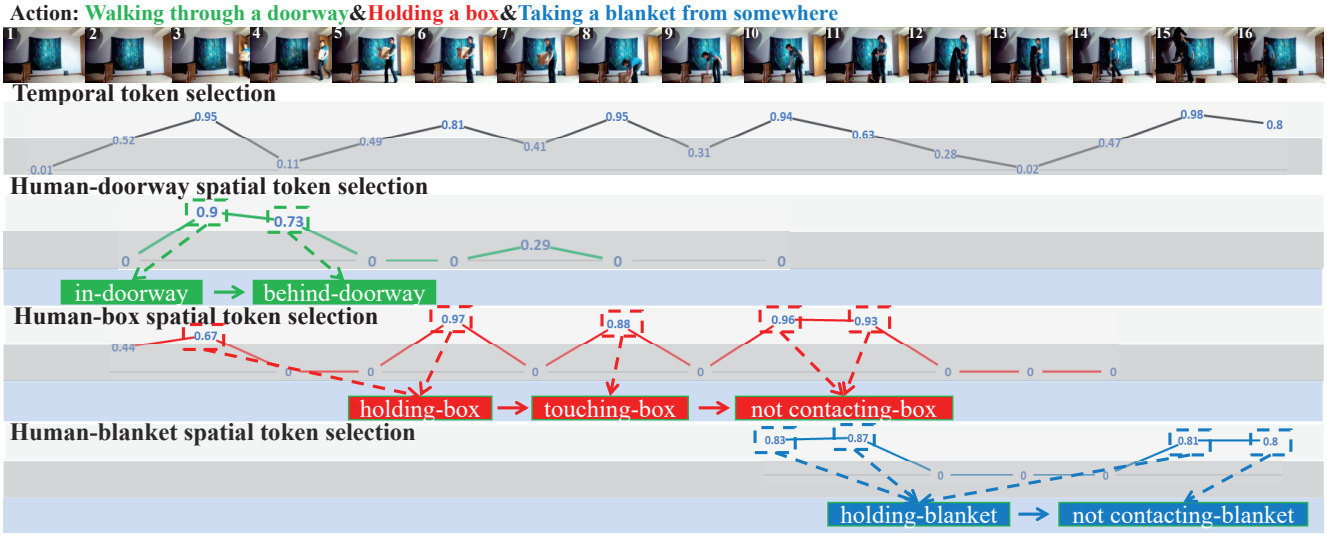

Figure 3. An example of action recognition performed by the proposed method and its corresponding process of providing explanations. Actions and their corresponding relation transitions are marked with the same color.

tacting box', suggest the action 'Holding a box'. Lastly, for human-blanket interactions, tokens from frames 10 and 11 labeled **'holding blanket'**, and the tokens from frames 15 and 16 labeled **'not contacting blanket'**, indicate the action 'Taking a blanket from somewhere'.

Figure 4, 5, and 6 provides additional examples. It details the patterns of relationship transitions for some action class, as identified by our method. This visual representation shows how the reasoning behind each action can be explained through these relationship transitions, thereby demonstrating the effectiveness and interpretative diversity of our approach.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[3] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 1

[4] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 971–980, 2017. 1

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[6] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. Timeception for complex action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 254–263, 2019. 1

[7] Yangjun Ou, Li Mi, and Zhenzhong Chen. Object-relation reasoning graph for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20133–20142, 2022. 1

[8] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 1

[9] 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

[10] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 803–818, 2018. 1

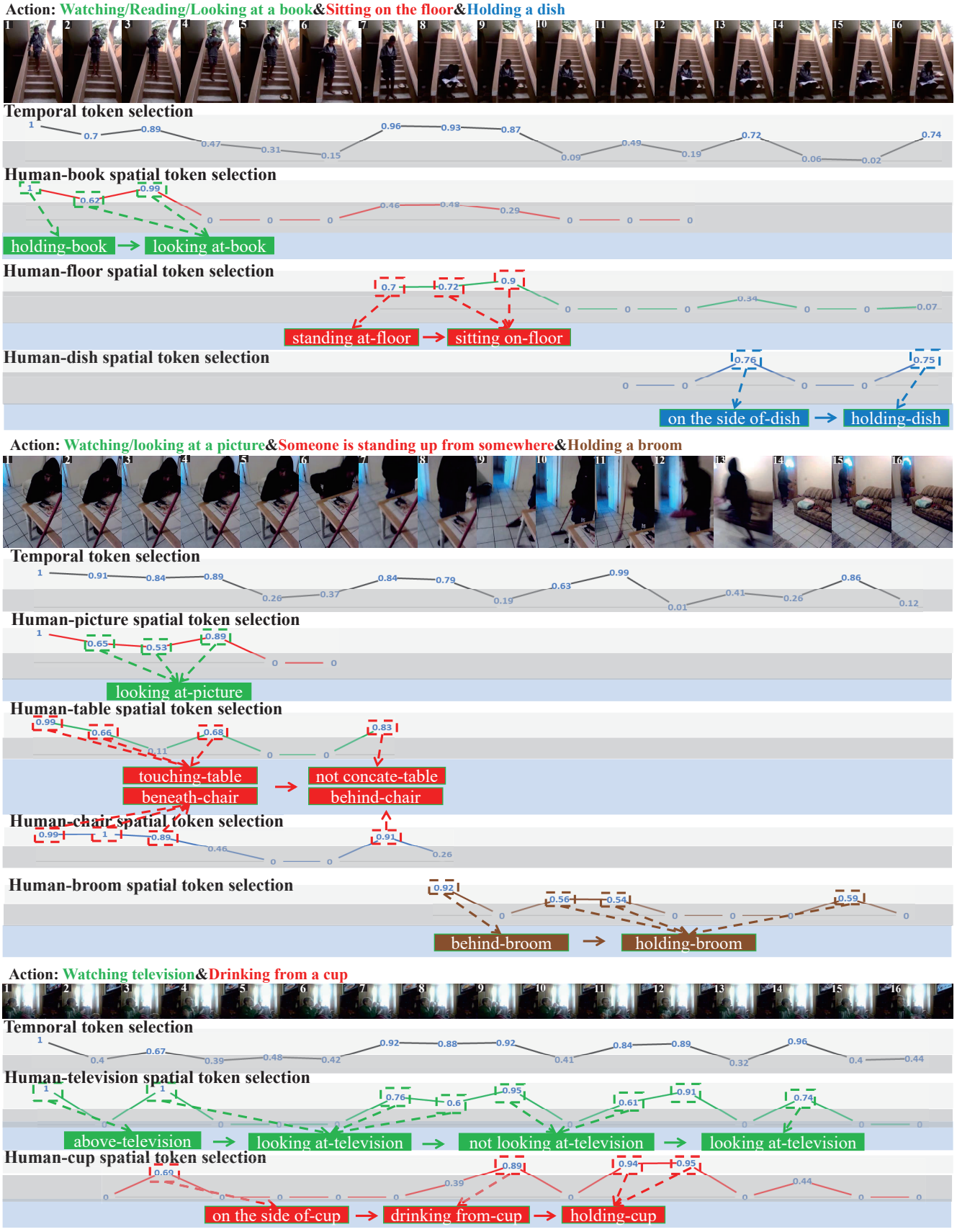

Figure 4. Some examples of the action-specific relation transitions provided by our method. Actions and their corresponding relation transitions are marked with the same color. Areas with no background color indicate that our method did not select the corresponding relation.

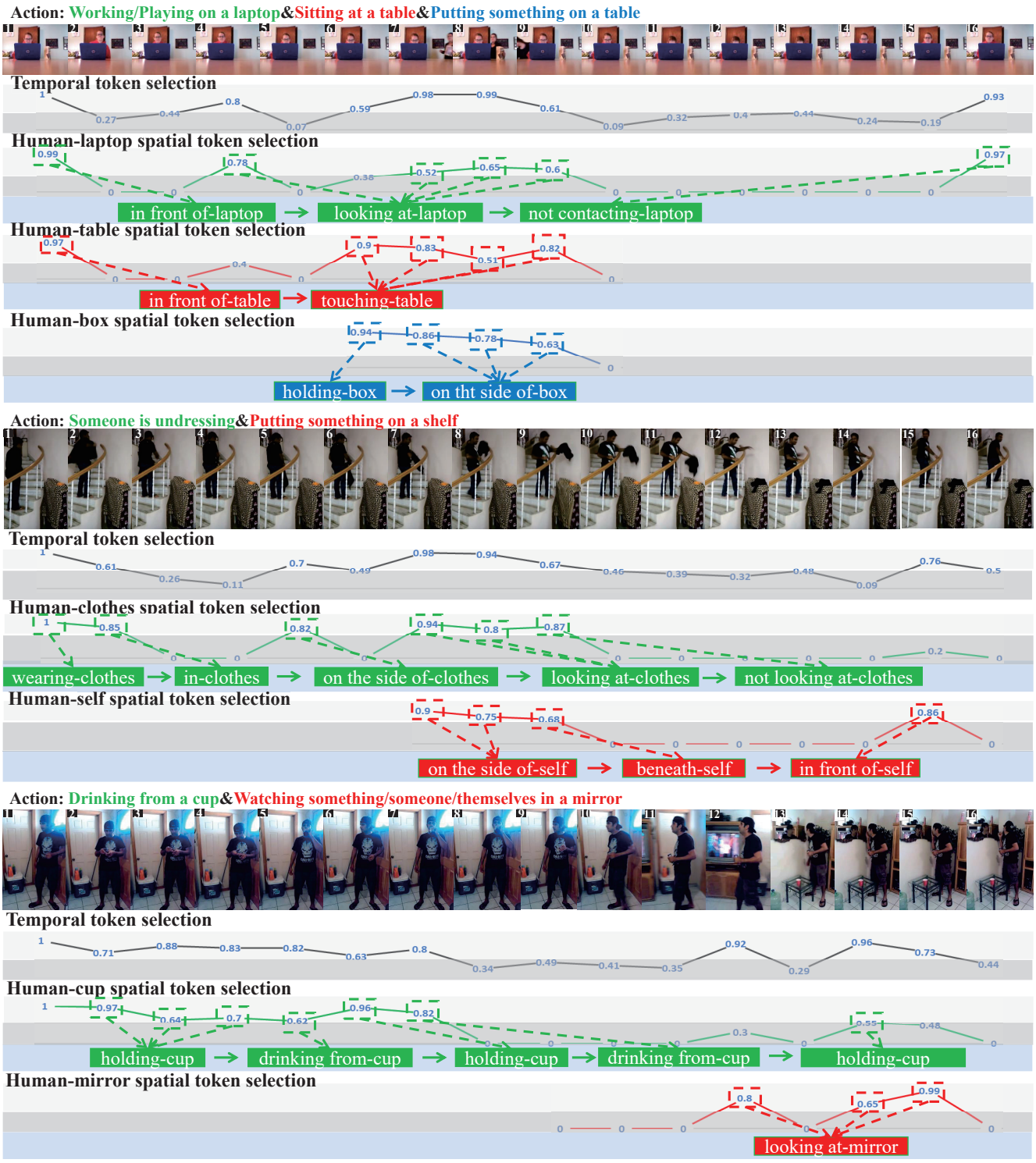

Figure 5. Some examples of the action-specific relation transitions provided by our method. Actions and their corresponding relation transitions are marked with the same color. Areas with no background color indicate that our method did not select the corresponding relation.

Figure 6. Some examples of the action-specific relation transitions provided by our method. Actions and their corresponding relation transitions are marked with the same color. Areas with no background color indicate that our method did not select the corresponding relation.