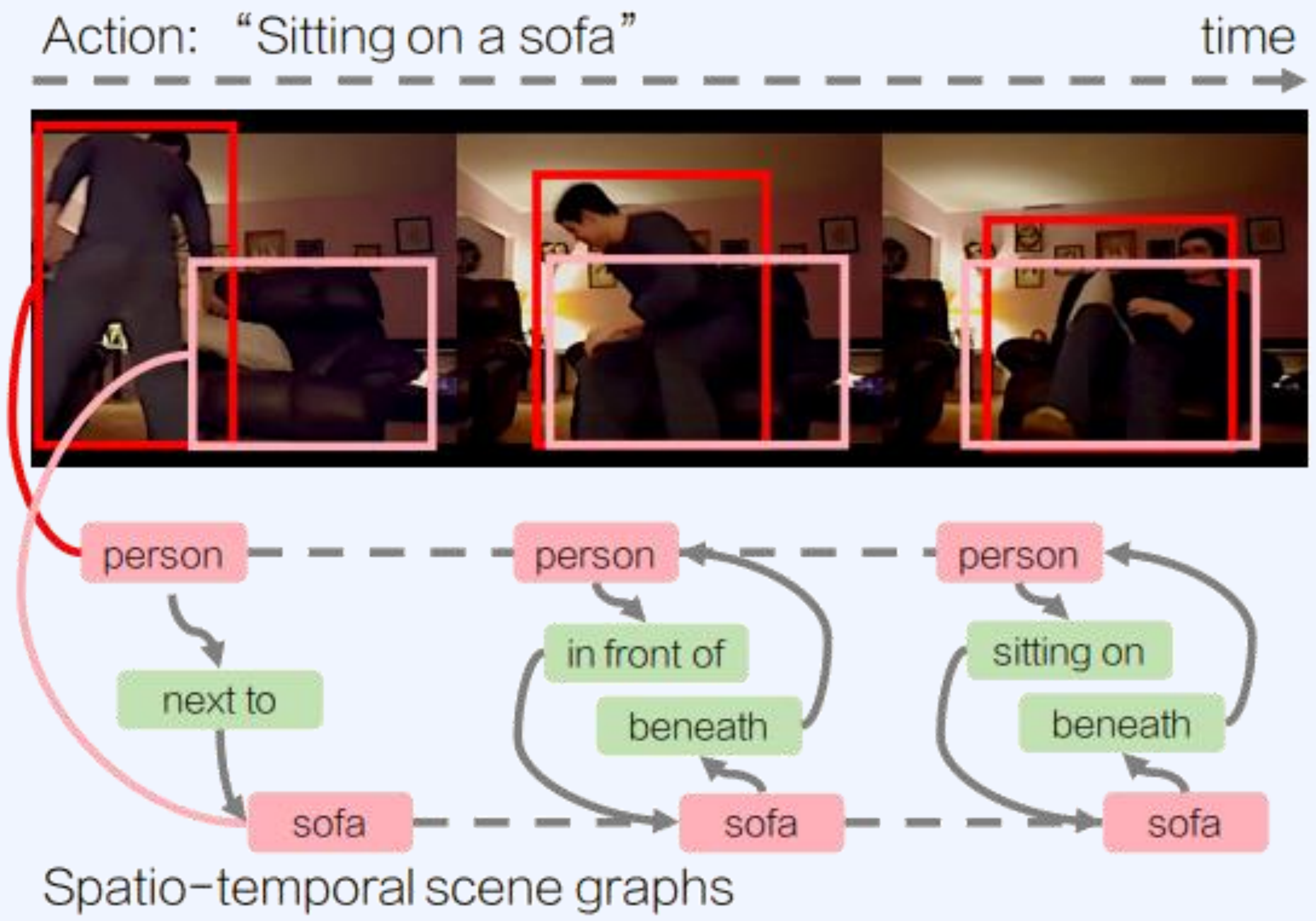
Video Understanding

2 Action Genome

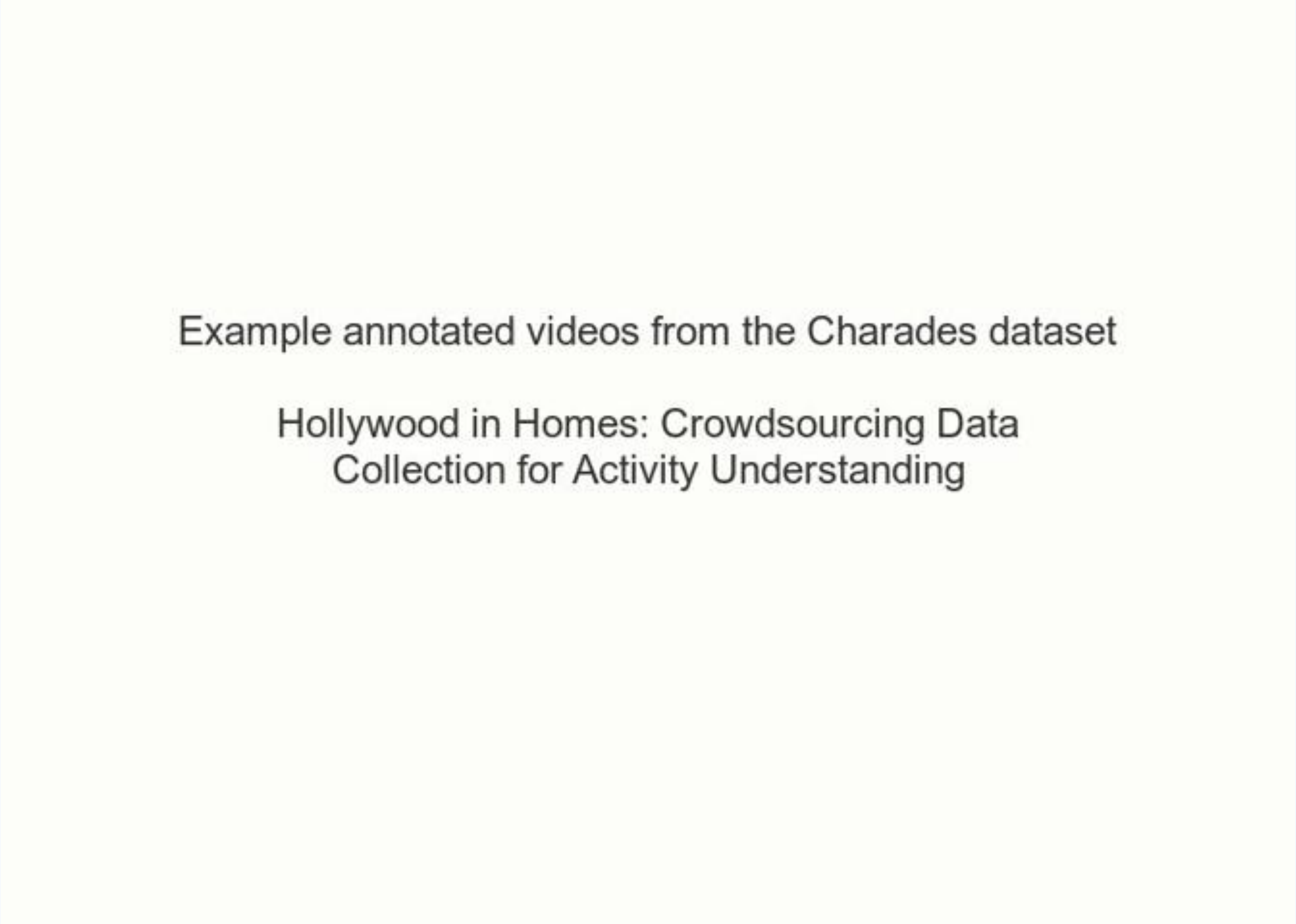
Video Understanding
Action Genome

J. Ji et al. Action Genome: Actions as Composition of Spatio-temporal Scene Graphs. CVPR



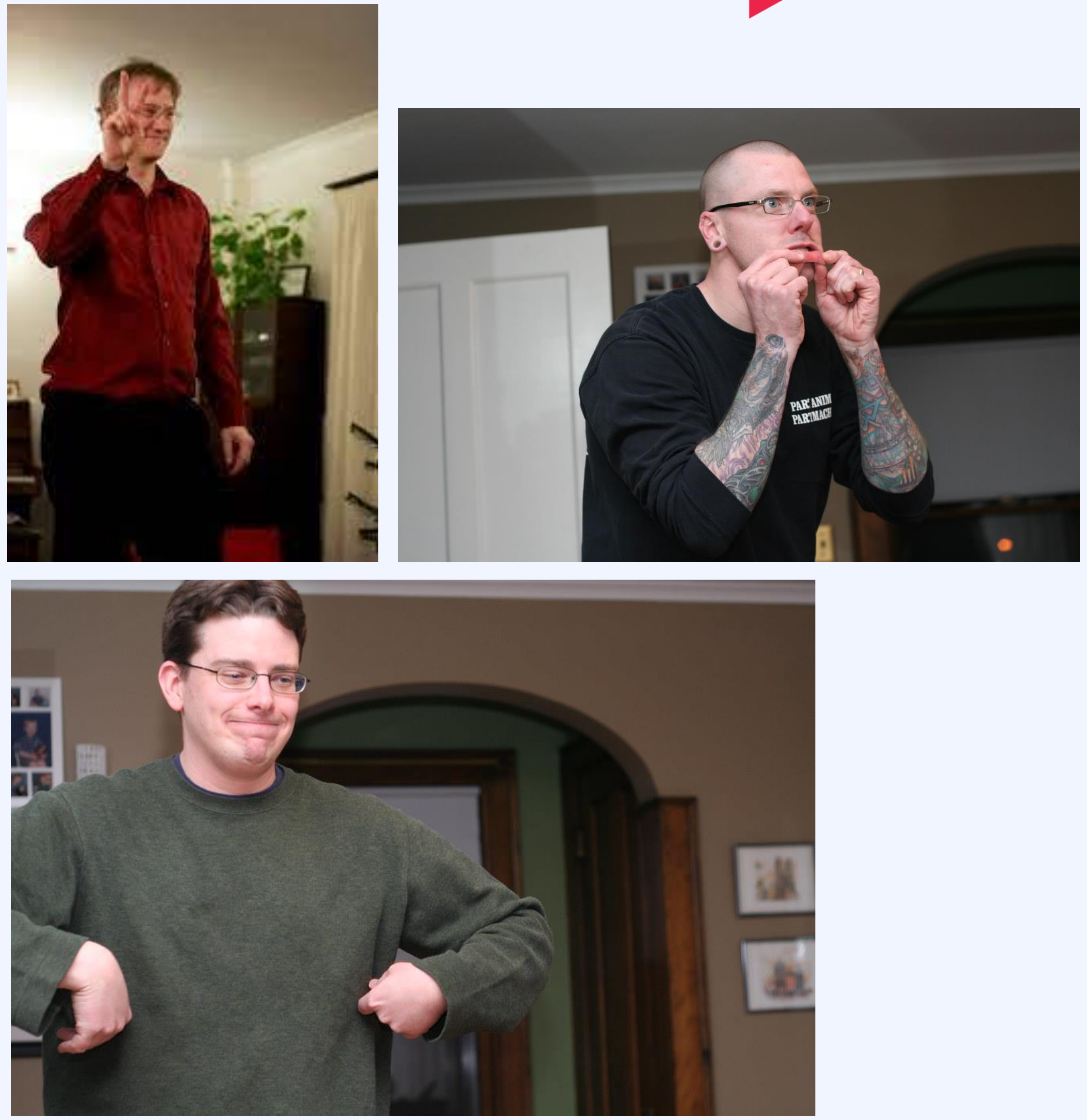
Video Understanding
Action Genome

Charades dataset : 9848 videos / 66,500 temporal annotations for 157 action classes / 41,104 labels for 46 object classes



Example annotated videos from the Charades dataset

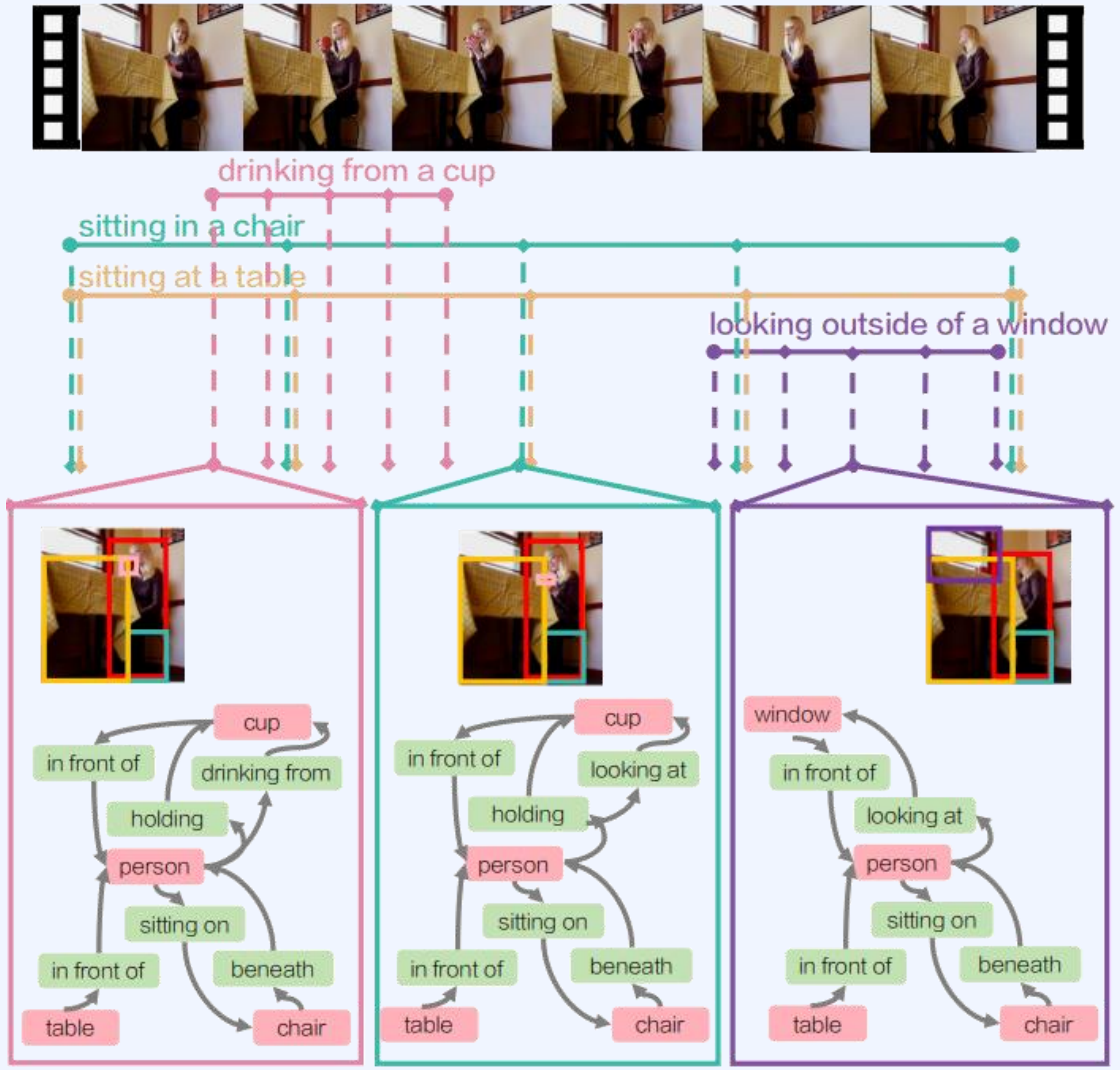
Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding



Wikimedia Commons

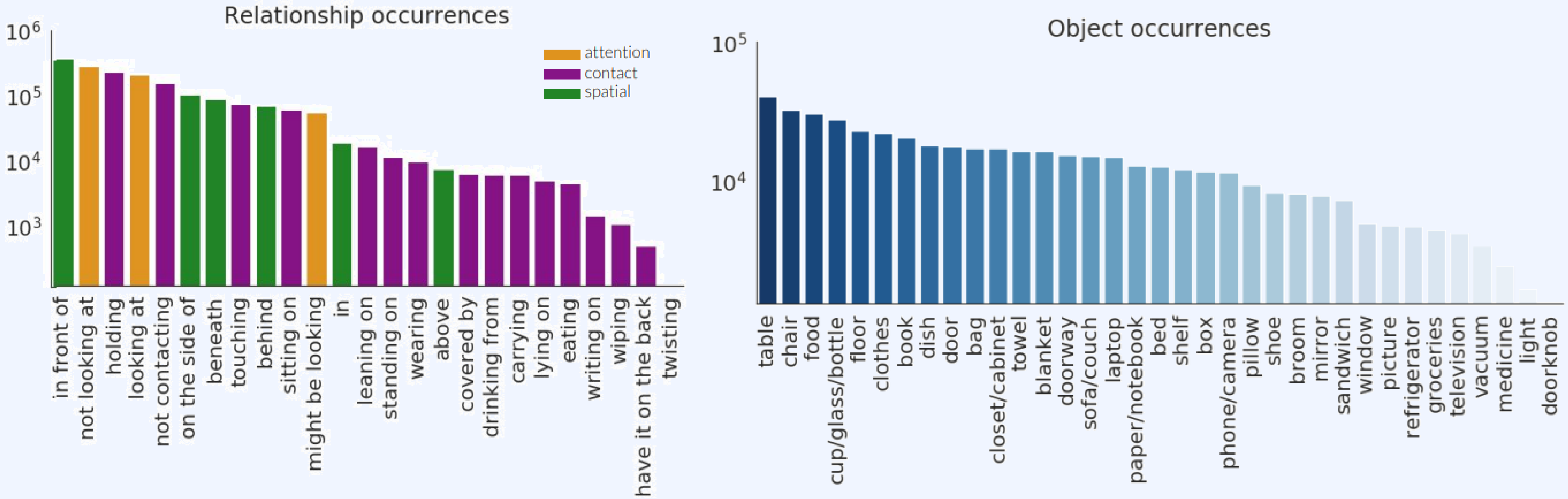
Video Understanding
Action Genome

J. Ji et al. Action Genome: Actions as Composition of Spatio-temporal Scene Graphs. CVPR



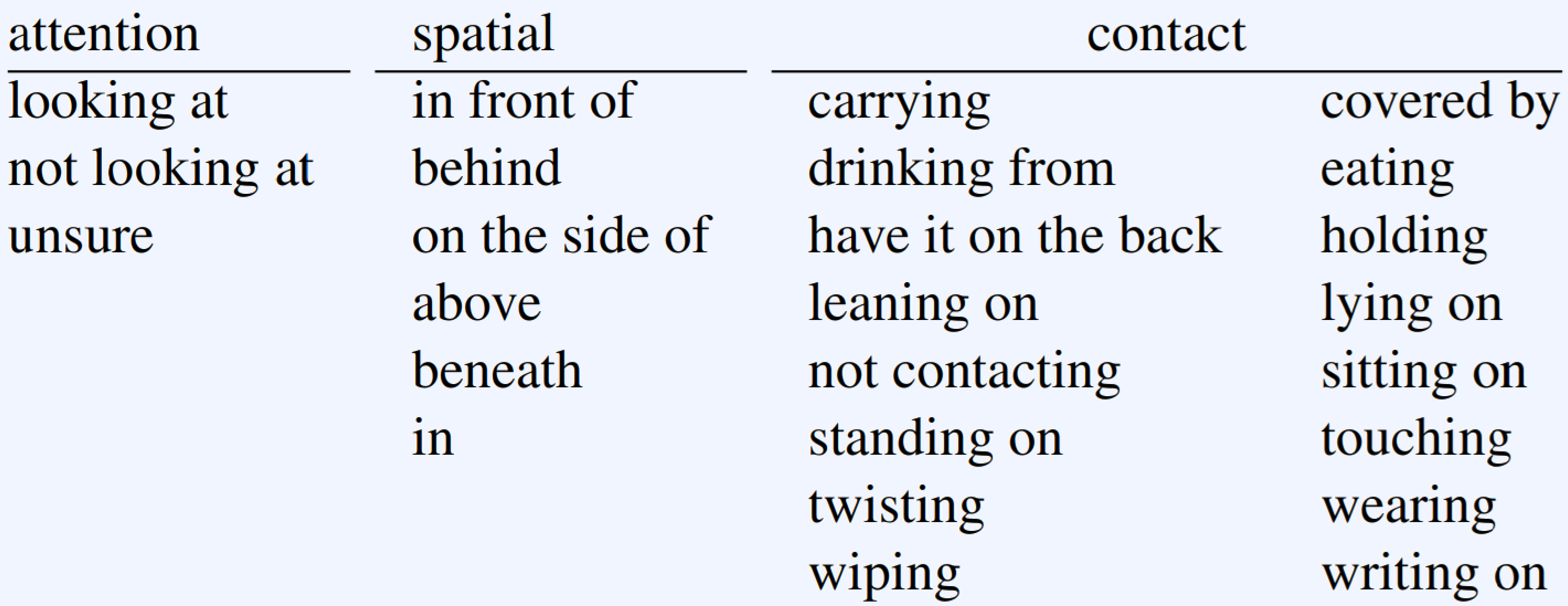
Video Understanding
Action Genome

J. Ji et al. Action Genome: Actions as Composition of Spatio-temporal Scene Graphs. CVPR



Video Understanding
Action Genome

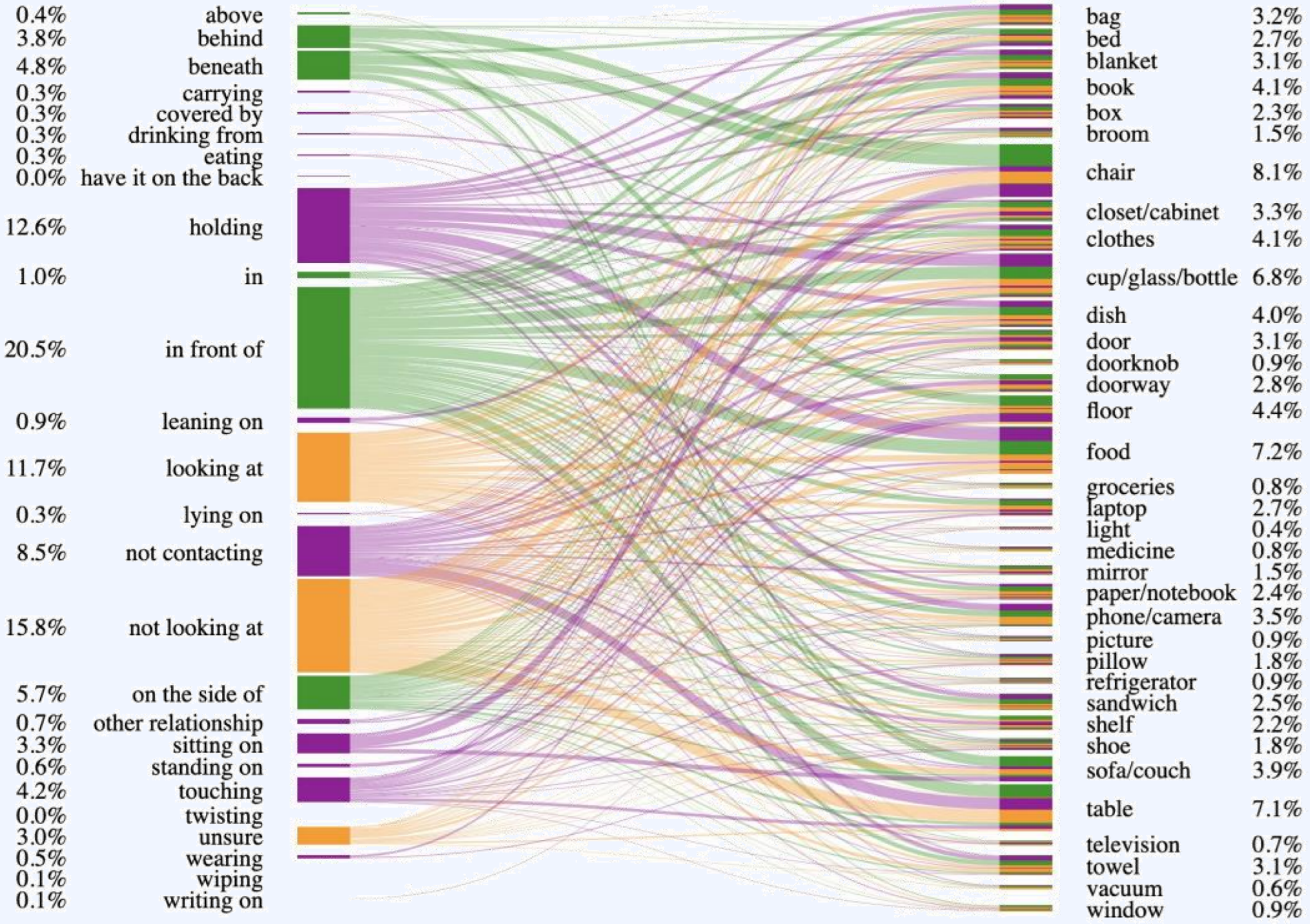
J. Ji et al. Action Genome: Actions as Composition of Spatio-temporal Scene Graphs. CVPR



attention	spatial	contact	
looking at	in front of	carrying	covered by
not looking at	behind	drinking from	eating
unsure	on the side of	have it on the back	holding
	above	leaning on	lying on
	beneath	not contacting	sitting on
	in	standing on	touching
		twisting	wearing
		wiping	writing on

Video Understanding
Action Genome

J. Ji et al. Action Genome: Actions as Composition of Spatio-temporal Scene Graphs. CVPR

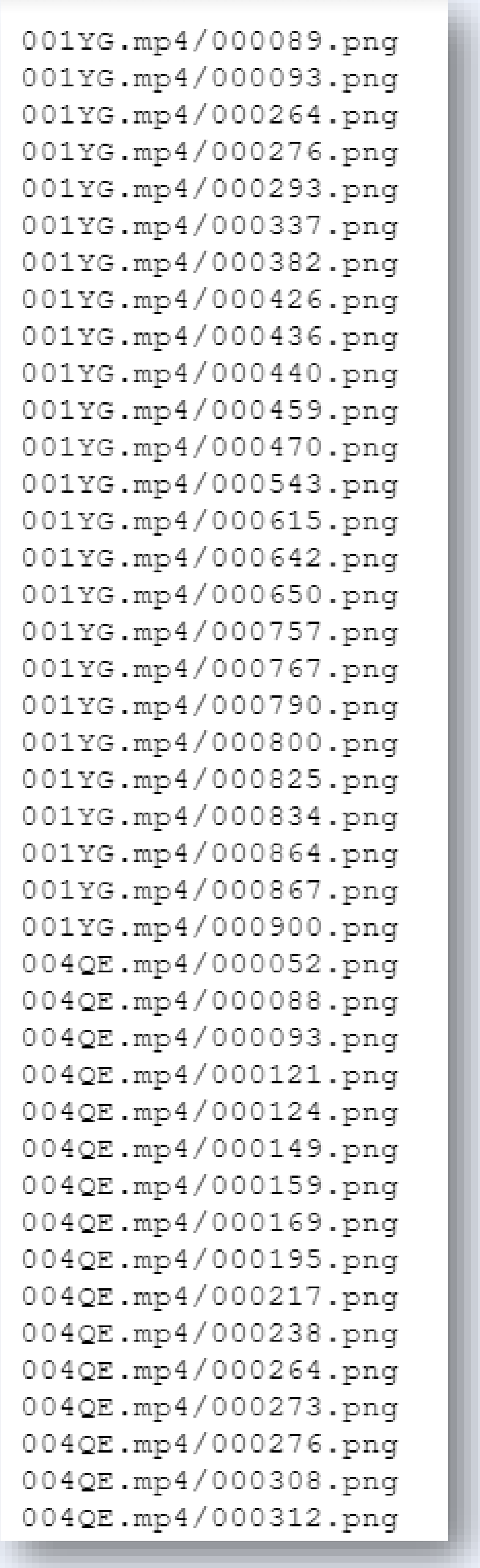


attention	spatial	contact	
looking at	in front of	carrying	covered by
not looking at	behind	drinking from	eating
unsure	on the side of	have it on the back	holding
	above	leaning on	lying on
	beneath	not contacting	sitting on
	in	standing on	touching
		twisting	wearing
		wiping	writing on

Video Understanding
Action Genome

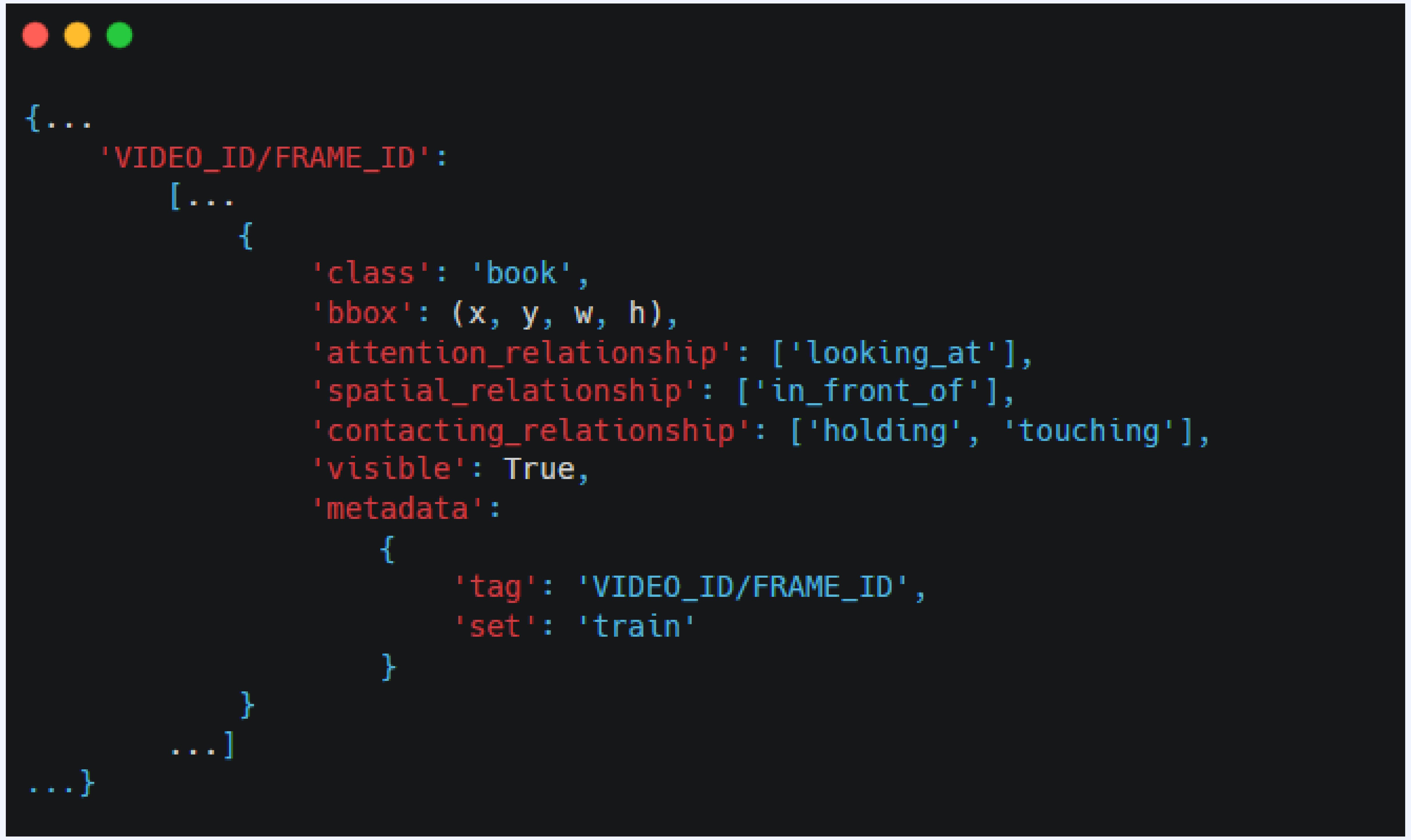
J. Ji et al. Action Genome: Actions as Composition of Spatio-temporal Scene Graphs. CVPR

frame_list.txt



```
001YG.mp4/000089.png
001YG.mp4/000093.png
001YG.mp4/000264.png
001YG.mp4/000276.png
001YG.mp4/000293.png
001YG.mp4/000337.png
001YG.mp4/000382.png
001YG.mp4/000426.png
001YG.mp4/000436.png
001YG.mp4/000440.png
001YG.mp4/000459.png
001YG.mp4/000470.png
001YG.mp4/000543.png
001YG.mp4/000615.png
001YG.mp4/000642.png
001YG.mp4/000650.png
001YG.mp4/000757.png
001YG.mp4/000767.png
001YG.mp4/000790.png
001YG.mp4/000800.png
001YG.mp4/000825.png
001YG.mp4/000834.png
001YG.mp4/000864.png
001YG.mp4/000867.png
001YG.mp4/000900.png
004QE.mp4/000052.png
004QE.mp4/000088.png
004QE.mp4/000093.png
004QE.mp4/000121.png
004QE.mp4/000124.png
004QE.mp4/000149.png
004QE.mp4/000159.png
004QE.mp4/000169.png
004QE.mp4/000195.png
004QE.mp4/000217.png
004QE.mp4/000238.png
004QE.mp4/000264.png
004QE.mp4/000273.png
004QE.mp4/000276.png
004QE.mp4/000308.png
004QE.mp4/000312.png
```

object_bbox_and_relationship.pkl



```
{...
    'VIDEO_ID/FRAME_ID':
        [...
            {
                'class': 'book',
                'bbox': (x, y, w, h),
                'attention_relationship': ['looking_at'],
                'spatial_relationship': ['in_front_of'],
                'contacting_relationship': ['holding', 'touching'],
                'visible': True,
                'metadata':
                    {
                        'tag': 'VIDEO_ID/FRAME_ID',
                        'set': 'train'
                    }
            }
        ...]
...}
```

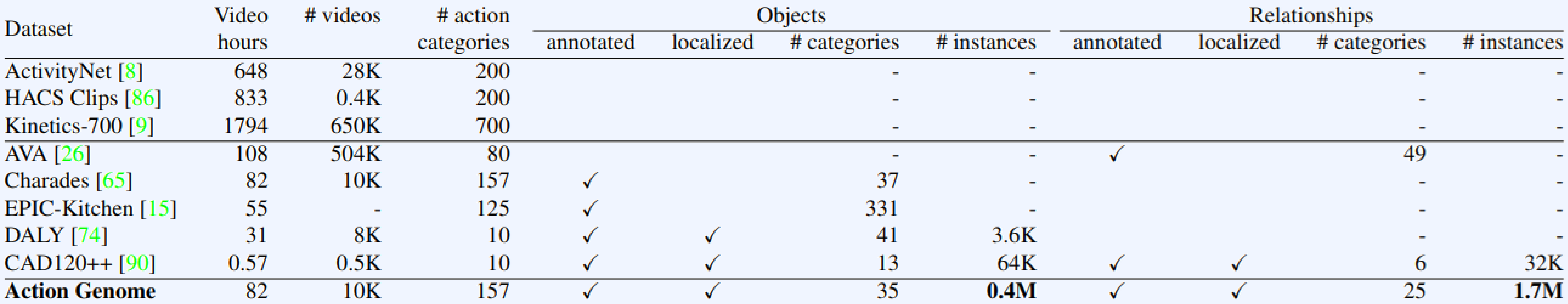
person_bbox.pkl

object_classes.txt

relationship_classes.txt

Video Understanding
Action Genome

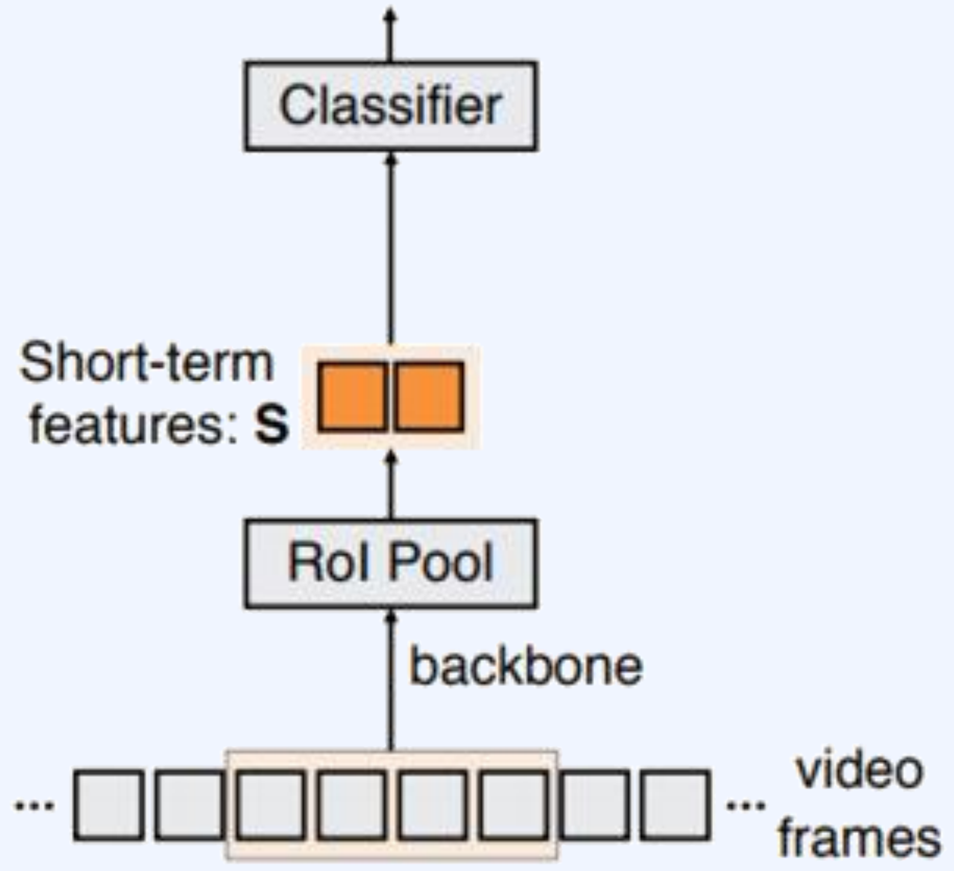
J. Ji et al. Action Genome: Actions as Composition of Spatio-temporal Scene Graphs. CVPR



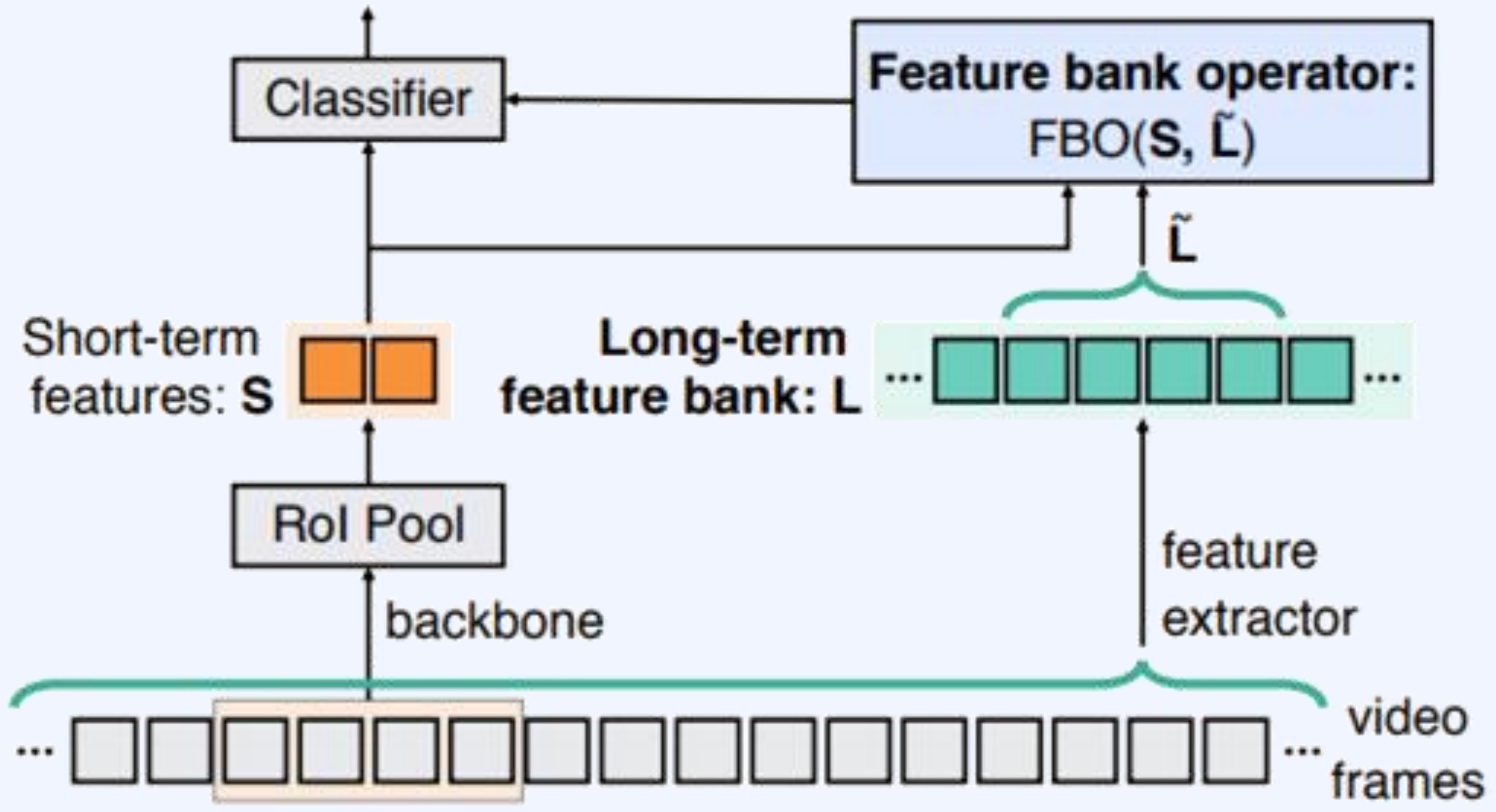
Dataset	Video	# videos	# action	Objects				Relationships			
	hours		categories	annotated	localized	# categories	# instances	annotated	localized	# categories	# instances
ActivityNet [8]	648	28K	200			-	-			-	-
HACS Clips [86]	833	0.4K	200			-	-			-	-
Kinetics-700 [9]	1794	650K	700			-	-			-	-
AVA [26]	108	504K	80			-	-	✓		49	-
Charades [65]	82	10K	157	✓		37	-			-	-
EPIC-Kitchen [15]	55	-	125	✓		331	-			-	-
DALY [74]	31	8K	10	✓	✓	41	3.6K			-	-
CAD120++ [90]	0.57	0.5K	10	✓	✓	13	64K	✓	✓	6	32K
Action Genome	82	10K	157	✓	✓	35	**0.4M**	✓	✓	25	**1.7M**

Video Understanding
Action Genome

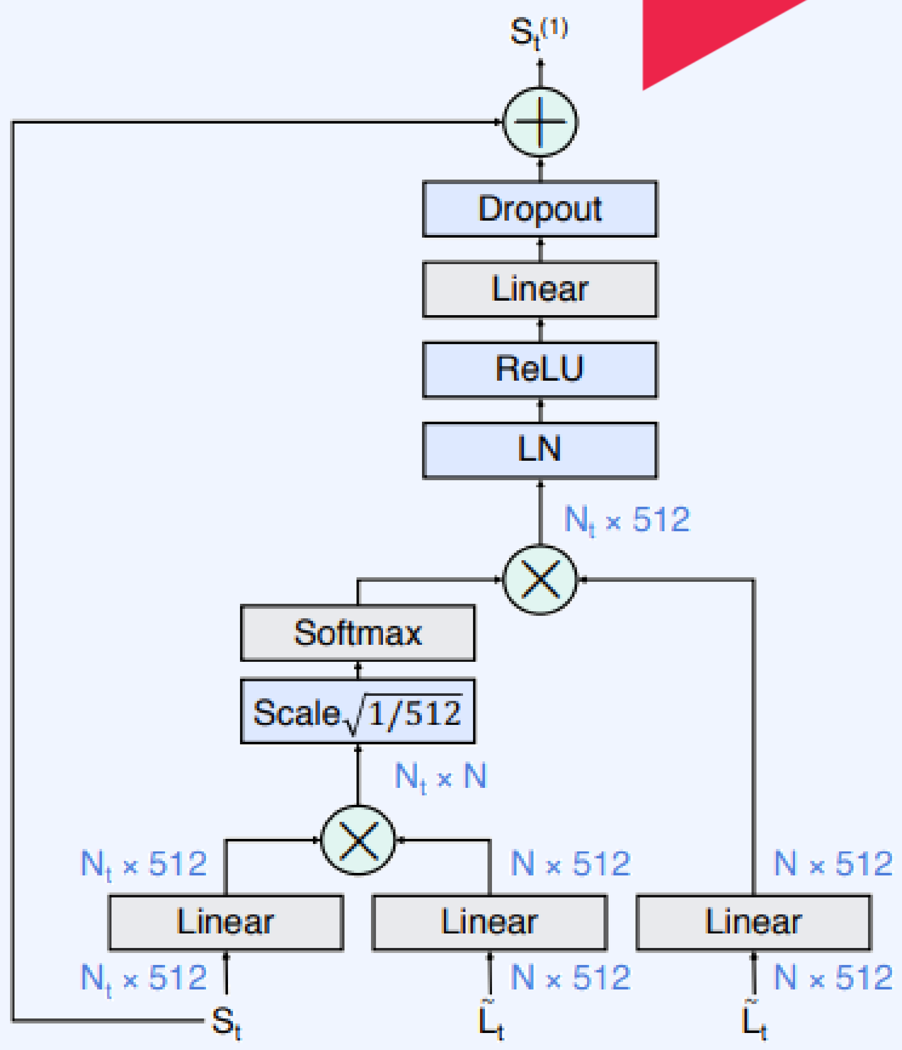
C. Wu et al. Long-Term Feature Banks for Detailed Video Understanding. CVPR



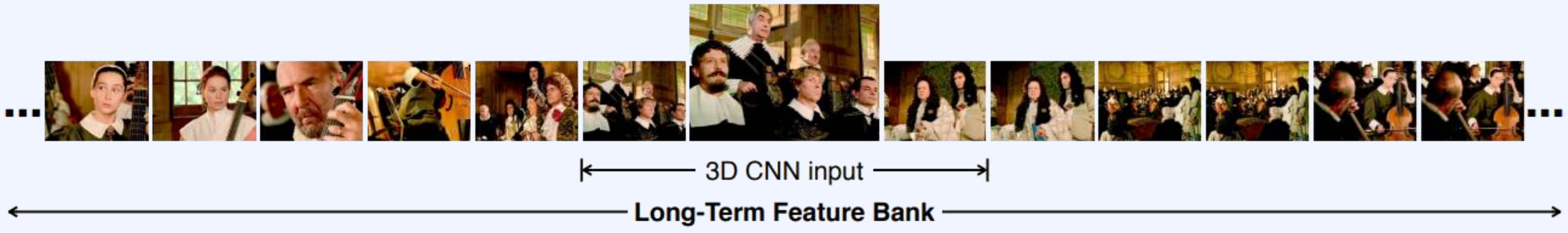
3D CNN



3D CNN + LFB

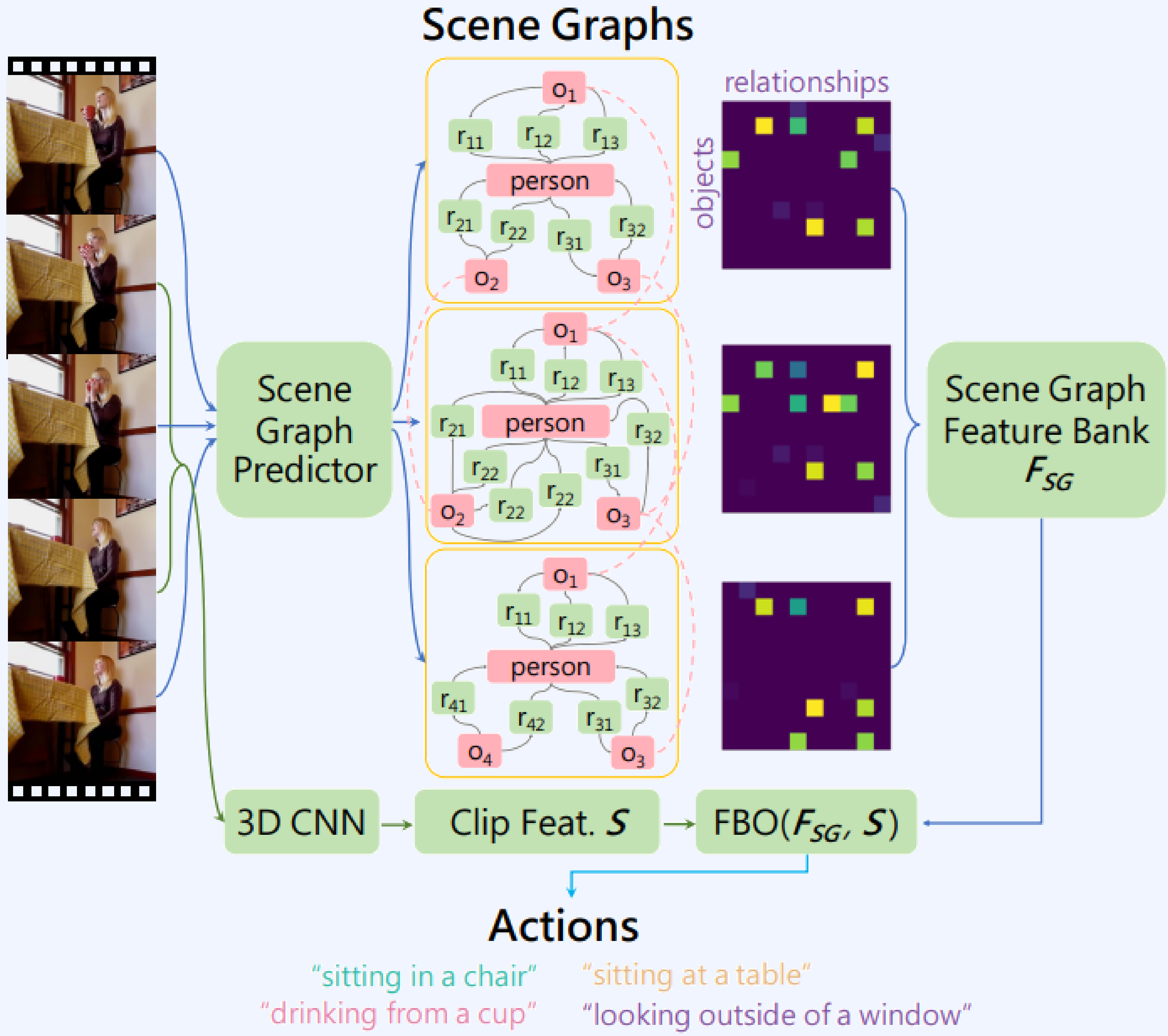


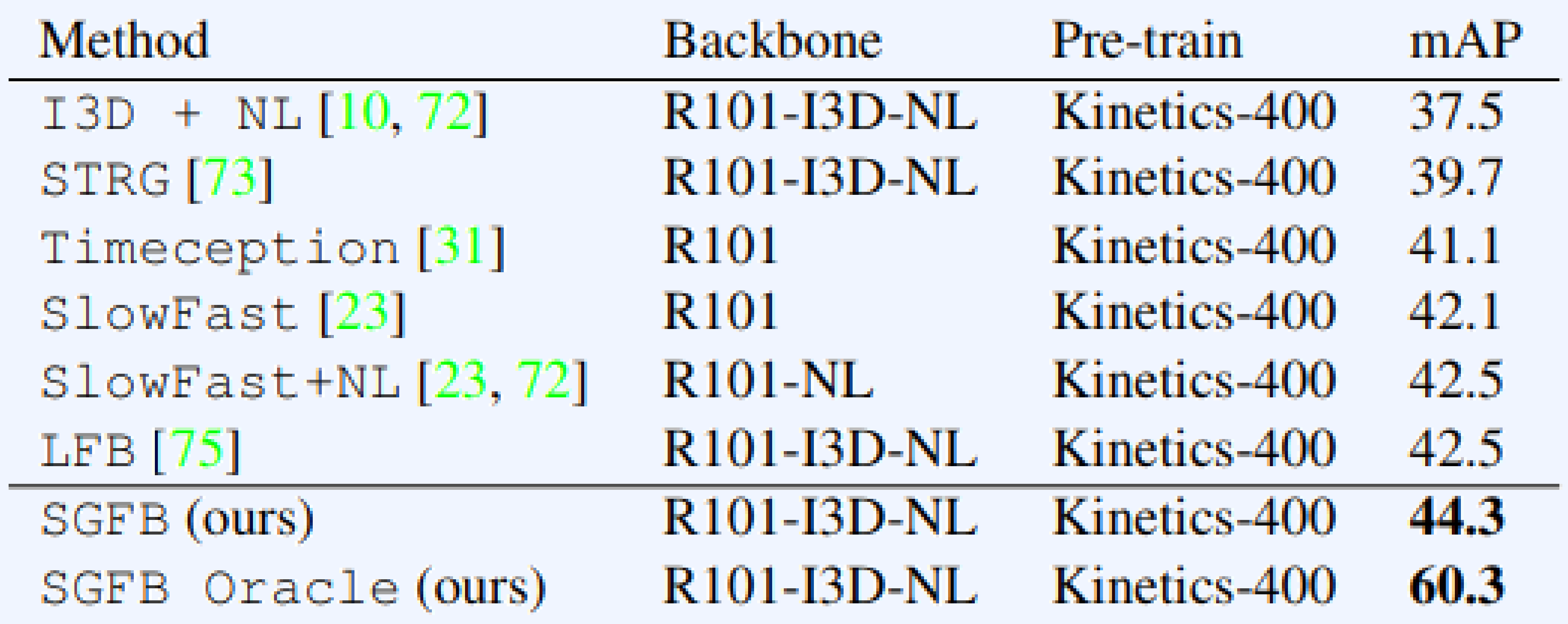
non-local block



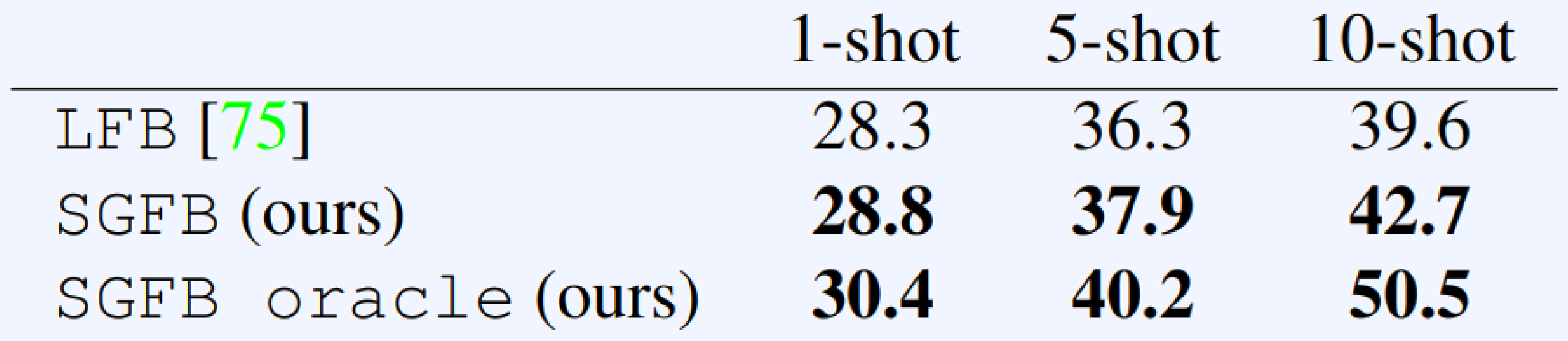
Video Understanding
Action Genome

J. Ji et al. Action Genome: Actions as Composition of Spatio-temporal Scene Graphs. CVPR





Method	Backbone	Pre-train	mAP
I3D + NL [10, 72]	R101-I3D-NL	Kinetics-400	37.5
STRG [73]	R101-I3D-NL	Kinetics-400	39.7
Timeception [31]	R101	Kinetics-400	41.1
SlowFast [23]	R101	Kinetics-400	42.1
SlowFast+NL [23, 72]	R101-NL	Kinetics-400	42.5
LFB [75]	R101-I3D-NL	Kinetics-400	42.5
SGFB (ours)	R101-I3D-NL	Kinetics-400	**44.3**
SGFB Oracle (ours)	R101-I3D-NL	Kinetics-400	**60.3**



	1-shot	5-shot	10-shot
LFB [75]	28.3	36.3	39.6
SGFB (ours)	**28.8**	**37.9**	**42.7**
SGFB oracle (ours)	**30.4**	**40.2**	**50.5**