VirtualHome Action Genome: A Simulated Spatio-Temporal Scene Graph Dataset with Consistent Relationship Labels (Supplementary Material)

This supplementary material provides additional details of the proposed dataset VirtualHAG, including scripts for video generation, co-occurrence statistics for objects and relationships, and dataset examples. We also provide more experimental results on the proposed dataset VirtualHAG.

A. Additional Details on VirtualHAG Dataset

Scripts for Video Generation. We manually designed 108 unique scripts to generate videos in the VirtualHome simulator. We show eight script examples in Table 6. Each script consists of a sequence of actions, and each action contains an action type and related objects. During the execution of scripts, object instances are randomly instantiated.

Co-occurrence Statistics for Objects and Relationships. We show the co-occurrence statistics for objects and relationships of the VirtualHAG dataset in Figure 8. Relationship distribution tends to show relevance with object affordance. Such as, for small objects (*e.g.* "cupcake" class), the "touching", "holding", and "in" relationships dominate, while for "chair" and "bench", the "above" relationship is relatively more dominant than other relationships.

Dataset Examples. We show six dataset examples sampled from the VirtualHAG dataset in Figure 9 and Figure 10 ((a) to (f)). In each example, we show eight frames sampled from two different observation viewpoints.

B. Additional Experimental Results on VirtualHAG Dataset

We show three example results for frames and viewpoints sampled from the VirtualHAG dataset in Figure 11. The SGTracker exhibited high performance in localizing objects and humans and determining human-object relationships in these three examples. However, we also found that SGTracker is less accurate in predicting correct relationships around the action change moment (*e.g.* Figure 11 example (a) Frames 250 and 384), and there is room for SG-Tracker to improve in detecting small objects (*e.g.* Figure 11 example (c) Frames 142, 152, 395 and 420).

Script abstract	Scripts
Sit and stand up from a chair	[WALK] ⟨chair⟩ [SIT] ⟨chair⟩ [STANDINGUP] ⟨chair⟩ [WALK] ⟨bed⟩
Sit and stand up from a sofa	$ \begin{array}{ l l l l l l l l l l l l l l l l l l l$
Switch on and switch off a light switch	[WALK] (lightswitch) [SWITCHON] (lightswitch) [SWITCHOFF] (lightswitch) [WALK] (chair)
Switch on and switch off a tv	$ \begin{array}{l} \left[\text{WALK} \right] \left\langle tv \right\rangle \\ \left[\text{SWITCHON} \right] \left\langle tv \right\rangle \\ \left[\text{SWITCHOFF} \right] \left\langle tv \right\rangle \\ \left[\text{WALK} \right] \left\langle table \right\rangle \\ \end{array} $
Open and close curtains	$\begin{array}{l} \mbox{[WALK]} \langle curtains \rangle \\ \mbox{[OPEN]} \langle curtains \rangle \\ \mbox{[CLOSE]} \langle curtains \rangle \\ \mbox{[WALK]} \langle toilet \rangle \end{array}$
Grab, drink, and put a cup of milk back	[WALK] ⟨milk⟩ [GRAB] ⟨milk⟩ [DRINK] ⟨milk⟩ [PUTOBJBACK] ⟨milk⟩ [WALK] ⟨bed⟩
Grab and put a cupcake in a fridge, and close the fridge	$ \begin{array}{l} \mbox{[WALK]} \langle cupcake \rangle \\ \mbox{[GRAB]} \langle cupcake \rangle \\ \mbox{[WALK]} \langle fridge \rangle \\ \mbox{[OPEN]} \langle fridge \rangle \\ \mbox{[PUTIN]} \langle cupcake \rangle \langle fridge \rangle \\ \mbox{[CLOSE]} \langle fridge \rangle \\ \mbox{[WALK]} \langle table \rangle \end{array} $
Grab a breadslice, heat the breadslice, and put the breadslice in a plate	WALK] (breadslice) [GRAB] (breadslice) [WALK] (toaster) [PUTIN] (breadslice) (toaster) [SWITCHON] (toaster) [SWITCHOFF] (toaster) [GRAB] (breadslice) [WALK] (plate) [PUTIN] (breadslice) (plate)

Table 6. Script examples used in the VirtualHAG dataset. The action types and objects are circle in "[]", " $\langle \rangle$ ", separately.

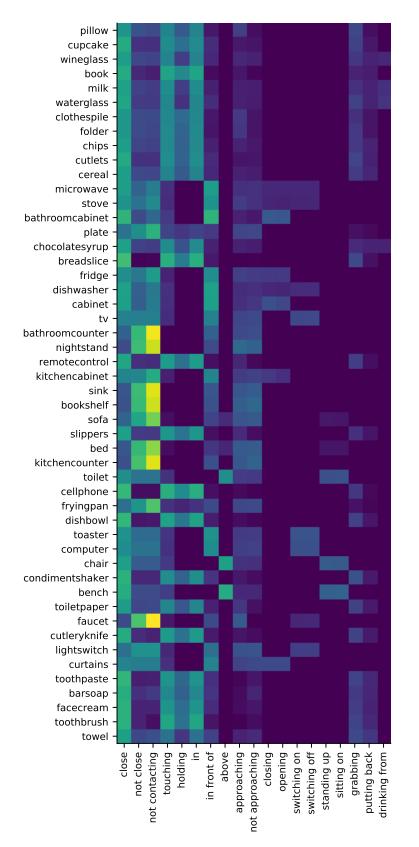


Figure 8. The co-occurence statistics for objects and relationships in the VirtualHAG dataset.



Figure 9. Three examples sampled from the VirtualHAG dataset.

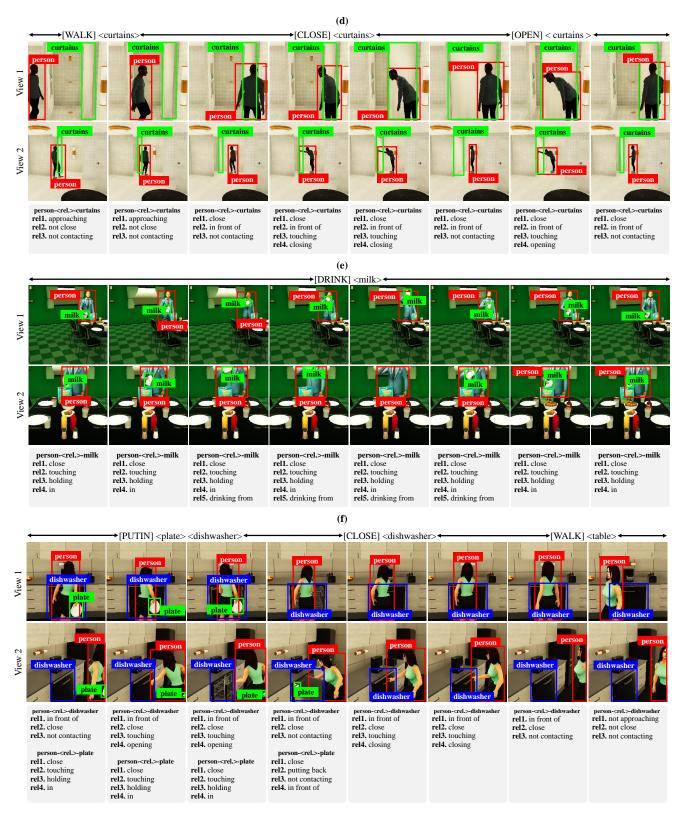


Figure 10. Three examples sampled from the VirtualHAG dataset.

			(;				
Frame 202, view 2	Frame 222, view 2	Frame 250, view 2	Frame 283, view 3	Frame 314, view 2	Frame 384, view 2	Frame 386, view 2	Frame 401, view 2
person	person A.	person at	person	person :	person A	person e	person
			bed				BAS
bed	bed	bed		bed	bed	bed	bed
SGTracker: person- <rel.>-bed</rel.>	SGTracker: person- <rel.>-bed</rel.>	SGTracker: person- <rel.>-bed</rel.>	SGTracker: person- <rel.>-bed</rel.>	SGTracker: person- <rel.>-bed</rel.>	SGTracker: person- <rel.>-bed</rel.>	SGTracker: person- <rel.>-bed</rel.>	SGTracker: person- <rel.>-bed</rel.>
rel1. not close rel2. approaching	rel1. not close rel2. approaching rel3. not contacting	rel1. not close rel2. approaching	rel1. close rel2. above	rel1. close rel2. above	rel1. not close rel2. not approaching	rel1. not close rel2. not approaching	rel1. not close rel2. not approaching
rel3. not contacting		rel3. not contacting	rel3. sitting on	rel3. standing up rel4. touching	rel3. not contacting	rel3. not contacting	rel3. not contacting
Ground truth: person- <rel.>-bed rel1. not close</rel.>	Ground truth: person- <rel.>-bed rel1. not close</rel.>	Ground truth: person- <rel.>-bed rel1. close</rel.>	Ground truth: person- <rel.>-bed rel1. close</rel.>	Ground truth: person- <rel.>-bed rel1. close</rel.>	Ground truth: person- <rel.>-bed rel1. close</rel.>	Ground truth: person- <rel.>-bed rel1. not close</rel.>	Ground truth: person- <rel.>-bed rel1. not close</rel.>
rel2. approaching rel3. not contacting	rel2. approaching rel3. not contacting	rel2. above rel3. sitting on	rel2. above rel3. sitting on	rel2. above rel3. sitting on	rel2. above rel3. standing up	rel2. not approaching rel3. not contacting	rel2. not approaching rel3. not contacting
		rel4. not contacting	rel4. touching	rel4. touching	rel4. not contacting		
Frame 136, view 2	Frame 159, view 5	Frame 204, view 5	Frame 219, view 2	Frame 219, view 5	Frame 240, view 5	Frame 246, view 5	Frame 252, view 5
			i w i			tv -	tv
	person	person		person	person	person	person
person	ara wa kata was	1992 and Kolle west	person	ara wa kata was	organit kain waa	organit kala was	are and water and
SGTracker: person- <rel.>-tv</rel.>	SGTracker: person- <rel.>-tv</rel.>	SGTracker: person- <rel.>-tv</rel.>	SGTracker: person- <rel.>-tv</rel.>	SGTracker: person- <rel.>-tv</rel.>	SGTracker: person- <rel.>-tv</rel.>	SGTracker: person- <rel.>-tv</rel.>	SGTracker: person- <rel.>-tv</rel.>
rel1. not close rel2. approaching	rel1. not close rel2. approaching	rel1. close rel2. in front of	rel1. close rel2. in front of	rel1. close rel2. in front of	rel1. close rel2. in front of	rel1. close rel2. in front of	rel1. close rel2. in front of
rel3. not approaching rel4. not contacting	rel3. not contacting	rel3. switching on rel4. switching off rel5. not contacting	rel3. switching on rel4. touching	rel3. switching on rel4. touching	rel3. switching off rel4. not contacting	rel3. switching on rel4. switching off rel5. not contacting	rel3. switching off rel4. not contacting
Ground truth:	Ground truth:	Ground truth:	Ground truth:	Ground truth:	Ground truth:	Ground truth:	Ground truth:
person- <rel.>-tv rel1. not close rel2. approaching</rel.>	person- <rel.>-tv rel1. not close rel2. approaching</rel.>	person- <rel.>-tv rel1. close rel2. in front of</rel.>	person- <rel.>-tv rel1. close rel2. in front of</rel.>	person- <rel.>-tv rel1. close rel2. in front of</rel.>	person- <rel.>-tv rel1. close rel2. in front of</rel.>	person- <rel.>-tv rel1. close rel2. in front of</rel.>	person- <rel.>-tv rel1. close rel2. in front of</rel.>
rel3. not contacting	rel3. not contacting	rel3. switching on rel4. not contacting	rel3. switching on rel4. touching	rel3. switching on rel4. touching	rel3. not contacting	rel3. switching off rel4. not contacting	rel3. switching off rel4. not contacting
Frame 142, view 2	Frame 152, view 2	Frame 155, view 4	(e Frame 177, view 4	c) Frame 395, view 7	Frame 420, view 7	Frame 431, view 7	Frame 487, view 7
person	person						person
cupcake	cupcake	cellphone person	cellphone	nightstand slippers	slippers nightstand	nightstand	
			person	person	person	person	nightstand
SC/Track-	SC/Track	SCTure	SC/Treel	SCTmal	SC/Track-	SCITural	SCTmal
SGTracker: person- <rel.>-cupcake rel1. in</rel.>	SGTracker: person- <rel.>- cupcake rel1. in</rel.>	SGTracker: person- <rel.>-cellphone rel1. in</rel.>	SGTracker: person- <rel.>-cellphone rel1. in</rel.>	SGTracker: person- <rel.>-slippers rel1. in</rel.>	SGTracker: person- <rel.>-slippers rel1. in</rel.>	SGTracker: person- <rel.>-nightstand rel1. close</rel.>	SGTracker: person- <rel.>-nightstand rel1. not close</rel.>
rel2. close rel3. touching	rel2. close rel3. touching	rel2. close rel3. touching	rel2. close rel3. touching	rel2. close rel3. touching	rel2. close rel3. touching	rel2. in front of rel3. not contacting	rel2. not approaching rel3. not contacting
rel4. holding	rel4. holding	rel4. holding	rel4. holding	rel4. holding person- <rel.>-nightstand</rel.>	rel4. holding person- <rel.>-nightstand</rel.>		
				rel1. not close rel2. approaching	rel1. not close rel2. approaching		
				rel3. not contacting	rel3. not contacting rel4. in front of		
Ground truth: person- <rel.>-cellphone rel1. in</rel.>	Ground truth: person- <rel.>-cellphone rel1. in</rel.>	Ground truth: person- <rel.>-cellphone rel1. in</rel.>	Ground truth: person- <rel.>-cellphone rel1. in</rel.>	Ground truth: person- <rel.>-cellphone rel1. in</rel.>	Ground truth: person- <rel.>-cellphone rel1. in</rel.>	Ground truth: person- <rel.>-nightstand rel1. close</rel.>	Ground truth: person- <rel.>-nightstand rel1. not close</rel.>
rel2. close rel3. touching	rel2. close rel3. touching	rel2. close rel3. touching	rel2. close rel3. touching	rel2. close rel3. touching	rel2. close rel3. touching	rel2. in front of rel3. not contacting	rel2. not approaching rel3. not contacting
rel4. holding	rel4. holding	rel4. holding	rel4. holding	rel4. holding	rel4. holding		
				rel1. not close rel2. approaching	rel1. not close rel2. approaching		
				rel3. not contacting	rel3. not contacting		

Figure 11. Example results on the VirtualHAG dataset. The incorrect predictions are highlighted in red.