

# 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 SGTracker 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	[WALK] ⟨sofa⟩ [SIT] ⟨sofa⟩ [STANDINGUP] ⟨sofa⟩ [WALK] ⟨bench⟩
Switch on and switch off a light switch	[WALK] ⟨lightswitch⟩ [SWITCHON] ⟨lightswitch⟩ [SWITCHOFF] ⟨lightswitch⟩ [WALK] ⟨chair⟩
Switch on and switch off a tv	[WALK] ⟨tv⟩ [SWITCHON] ⟨tv⟩ [SWITCHOFF] ⟨tv⟩ [WALK] ⟨table⟩
Open and close curtains	[WALK] ⟨curtains⟩ [OPEN] ⟨curtains⟩ [CLOSE] ⟨curtains⟩ [WALK] ⟨toilet⟩
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	[WALK] ⟨cupcake⟩ [GRAB] ⟨cupcake⟩ [WALK] ⟨fridge⟩ [OPEN] ⟨fridge⟩ [PUTIN] ⟨cupcake⟩ ⟨fridge⟩ [CLOSE] ⟨fridge⟩ [WALK] ⟨table⟩
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 “[ ]”, “⟨ ⟩”, separately.

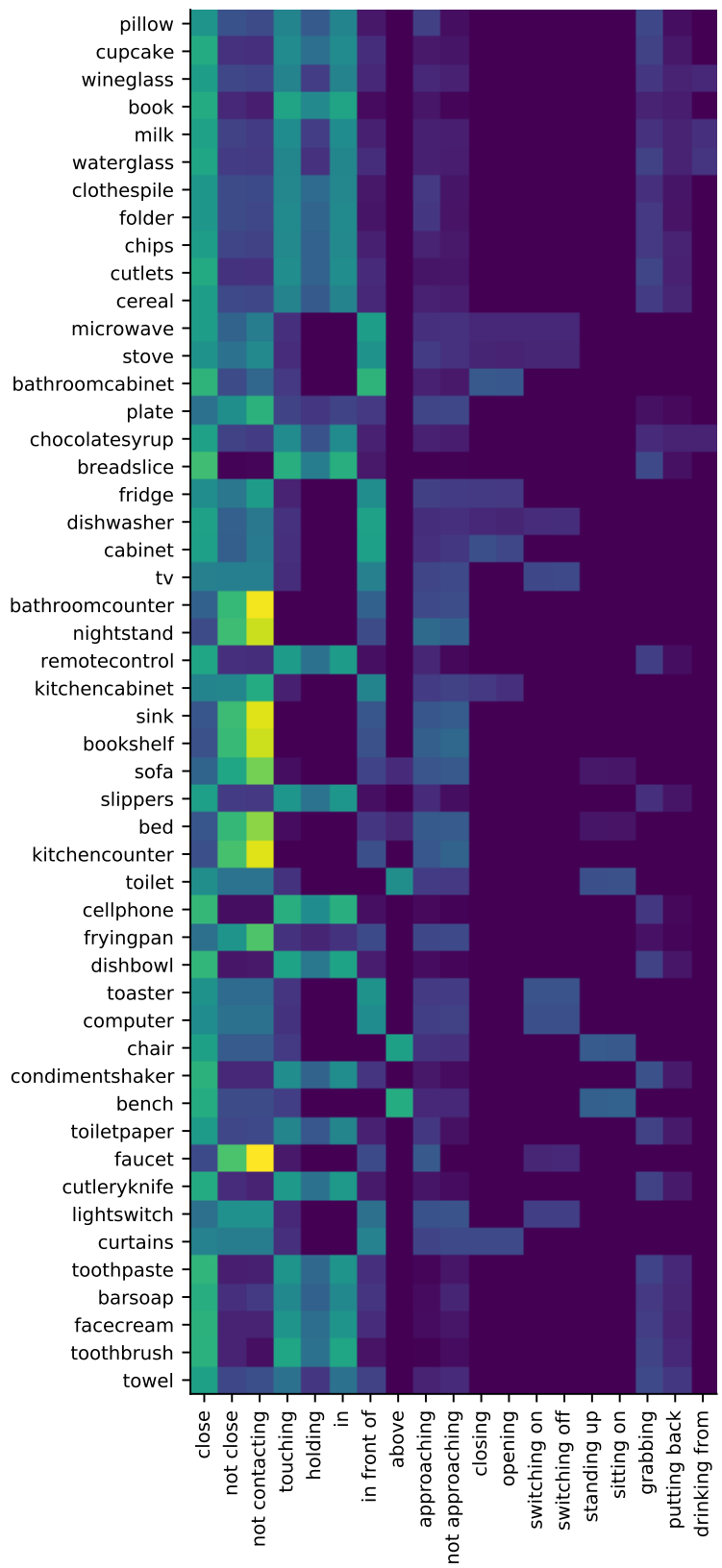


Figure 8. The co-occurrence statistics for objects and relationships in the VirtualHAG dataset.



Figure 9. Three examples sampled from the VirtualHAG dataset.

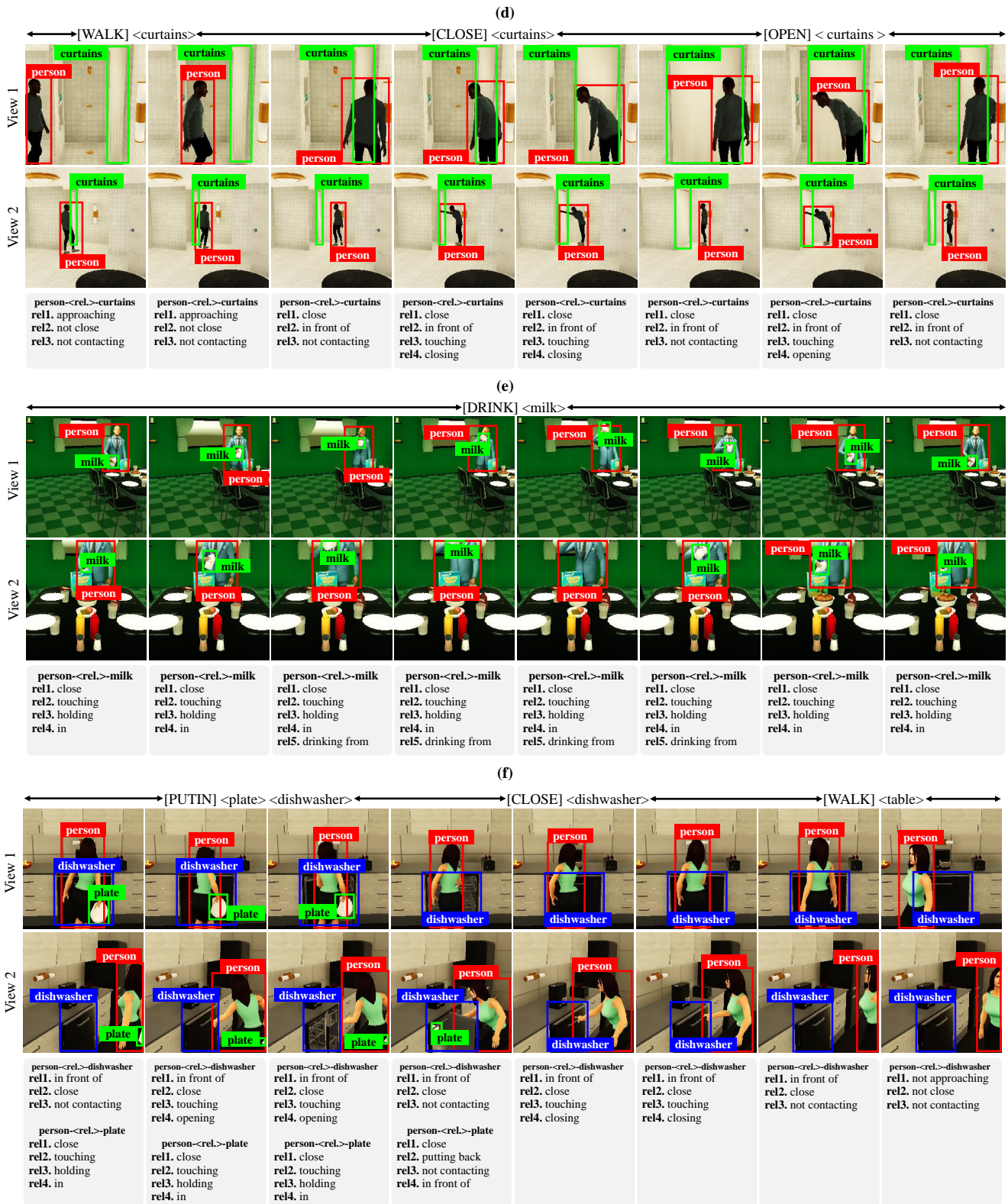


Figure 10. Three examples sampled from the VirtualHAG dataset.

(a)							
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
SGTracker: person-<rel.>-bed rel1. not close rel2. approaching rel3. not contacting	SGTracker: person-<rel.>-bed rel1. not close rel2. approaching rel3. not contacting	SGTracker: person-<rel.>-bed rel1. not close rel2. approaching rel3. not contacting	SGTracker: person-<rel.>-bed rel1. close rel2. above rel3. sitting on	SGTracker: person-<rel.>-bed rel1. close rel2. above rel3. standing up rel4. touching	SGTracker: person-<rel.>-bed rel1. not close rel2. not approaching rel3. not contacting	SGTracker: person-<rel.>-bed rel1. not close rel2. not approaching rel3. not contacting	SGTracker: person-<rel.>-bed rel1. not close rel2. not approaching rel3. not contacting
Ground truth: person-<rel.>-bed rel1. not close rel2. approaching rel3. not contacting	Ground truth: person-<rel.>-bed rel1. not close rel2. approaching rel3. not contacting	Ground truth: person-<rel.>-bed rel1. close rel2. above rel3. sitting on rel4. not contacting	Ground truth: person-<rel.>-bed rel1. close rel2. above rel3. sitting on rel4. touching	Ground truth: person-<rel.>-bed rel1. close rel2. above rel3. sitting on rel4. touching	Ground truth: person-<rel.>-bed rel1. close rel2. above rel3. standing up rel4. not contacting	Ground truth: person-<rel.>-bed rel1. not close rel2. not approaching rel3. not contacting	Ground truth: person-<rel.>-bed rel1. not close rel2. not approaching rel3. not contacting
(b)							
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
SGTracker: person-<rel.>-tv rel1. not close rel2. approaching rel3. not approaching rel4. not contacting	SGTracker: person-<rel.>-tv rel1. not close rel2. approaching rel3. not contacting	SGTracker: person-<rel.>-tv rel1. close rel2. in front of rel3. switching on rel4. switching off rel5. not contacting	SGTracker: person-<rel.>-tv rel1. close rel2. in front of rel3. switching on rel4. touching	SGTracker: person-<rel.>-tv rel1. close rel2. in front of rel3. switching on rel4. touching	SGTracker: person-<rel.>-tv rel1. close rel2. in front of rel3. switching off rel4. not contacting	SGTracker: person-<rel.>-tv rel1. close rel2. in front of rel3. switching on rel4. switching off rel5. not contacting	SGTracker: person-<rel.>-tv rel1. close rel2. in front of rel3. switching off rel4. not contacting
Ground truth: person-<rel.>-tv rel1. not close rel2. approaching rel3. not contacting	Ground truth: person-<rel.>-tv rel1. not close rel2. approaching rel3. not contacting	Ground truth: person-<rel.>-tv rel1. close rel2. in front of rel3. switching on rel4. not contacting	Ground truth: person-<rel.>-tv rel1. close rel2. in front of rel3. switching on rel4. touching	Ground truth: person-<rel.>-tv rel1. close rel2. in front of rel3. switching on rel4. touching	Ground truth: person-<rel.>-tv rel1. close rel2. in front of rel3. not contacting	Ground truth: person-<rel.>-tv rel1. close rel2. in front of rel3. switching off rel4. not contacting	Ground truth: person-<rel.>-tv rel1. close rel2. in front of rel3. switching off rel4. not contacting
(c)							
Frame 142, view 2	Frame 152, view 2	Frame 155, view 4	Frame 177, view 4	Frame 395, view 7	Frame 420, view 7	Frame 431, view 7	Frame 487, view 7
SGTracker: person-<rel.>-cupcake rel1. in rel2. close rel3. touching rel4. holding	SGTracker: person-<rel.>-cupcake rel1. in rel2. close rel3. touching rel4. holding	SGTracker: person-<rel.>-cellphone rel1. in rel2. close rel3. touching rel4. holding	SGTracker: person-<rel.>-cellphone rel1. in rel2. close rel3. touching rel4. holding	SGTracker: person-<rel.>-slippers rel1. in rel2. close rel3. touching rel4. holding	SGTracker: person-<rel.>-slippers rel1. in rel2. close rel3. touching rel4. holding	SGTracker: person-<rel.>-nightstand rel1. close rel2. in front of rel3. not contacting	SGTracker: person-<rel.>-nightstand rel1. not close rel2. not approaching rel3. not contacting
Ground truth: person-<rel.>-cellphone rel1. in rel2. close rel3. touching rel4. holding	Ground truth: person-<rel.>-cellphone rel1. in rel2. close rel3. touching rel4. holding	Ground truth: person-<rel.>-cellphone rel1. in rel2. close rel3. touching rel4. holding	Ground truth: person-<rel.>-cellphone rel1. in rel2. close rel3. touching rel4. holding	Ground truth: person-<rel.>-cellphone rel1. in rel2. close rel3. touching rel4. holding	Ground truth: person-<rel.>-cellphone rel1. in rel2. close rel3. touching rel4. holding	Ground truth: person-<rel.>-nightstand rel1. close rel2. in front of rel3. not contacting	Ground truth: person-<rel.>-nightstand rel1. not close rel2. not approaching rel3. not contacting

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