Action Scene Graphs for Long-Form Understanding of Egocentric Videos

Supplementary Material

A. Screenshots from the Annotation tool

Figure 6 reports some screenshots from the annotation procedure. The procedure follows different steps as described in the main paper and illustrated in the figure. An interface providing instructions is initially shown to the annotator (Figure 6a). The annotator can hence play a video clip sampled around the PNR frame (Figure 6b). They are then prompted to select among a set of possible relations (Figure 6c). The annotator can add indirect objects by selecting among a list of proposals extracted from narrations or searching from taxonomy (Figure 6d). They will hence ground each indirect object in the three PRE, PNR, and POST frames (Figure 6e). If the provided verb-noun pair is incorrect, the annotators can specify an alternative correct pair (Figure 6f).

B. EASG Examples

Figure 7 provides examples of graph sequences sampled from Ego4D-EASG dataset. The temporal nature of the graphs allows to model long-form relations between the objects in the scene and the camera wearer. For instance, indirect objects may become direct (compare Figure 7a with Figure 7b), and vice versa (compare Figure 7b with Figure 7c).

C. Verbs, prepositions and object nouns of the EASG dataset

Table 6 reports the extensive list of all verbs, relations and object nouns.

D. Annotation costs

We spent \$0.072 per annotator per 1 Human Intelligence Task, thus spending around \$1500 in total for both annotation stages. Note, that this cost does not include taxes and service fees, which may depend on the country and platform used.

E. Prompts for Anticipation and Summarization

All the prompts we are using contain one example of an input sequence and of completion. We provide these examples in order to ensure the correct format of output sequences for the downstream tasks of action anticipation and long-form summarization. Hence, every prompt consists of four parts: 1) Task description; 2) Input example; 3) Output example; 4) Input sequence for which the request is sent. Table 7 and table 8 summarize descriptions, input and output examples for the considered anticipation and summarization tasks respectively.

F. Qualitative Results for EASG Generation

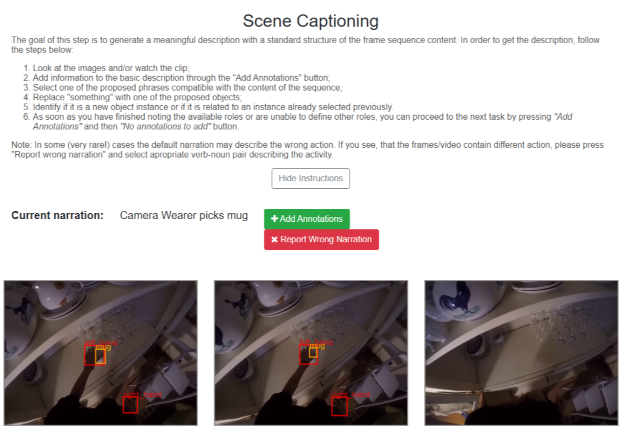
We provide qualitative results of our baseline model in Figure 8. We draw each graph using the top 10 predictions under the *No Contraint* setup. We can observe that the generated graphs for *EASG Cls* have more false positives than the other two tasks, indicating that action verbs play a significant role in EASG understanding.

G. Qualitative Results for Downstream Tasks

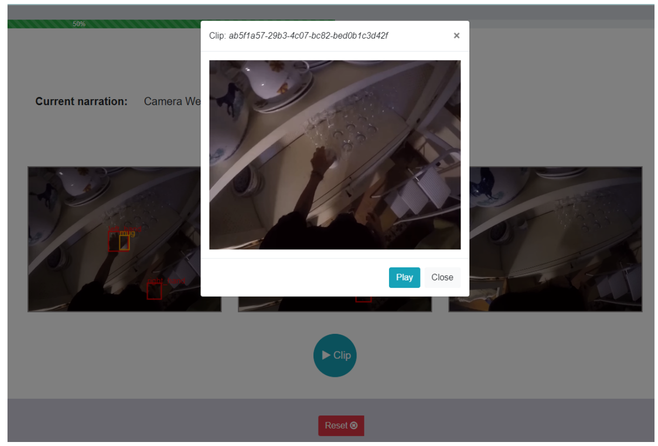
EASGs provide important context for better understanding the whole activity in general. Classic atomic verb-noun form action representations allow to focus on activities performed and human-object interactions, but they often miss important context required for long-form video understanding. In our dataset, there are examples of graph sequences where the important term mentioned in the clip summary does not appear as the active object.

Figure 9 and Figure 10 show qualitative anticipation and summarization examples respectively.

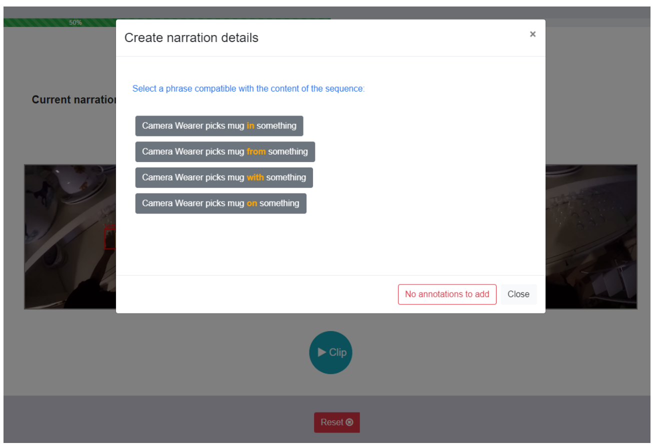
Acknowledgements. This research is supported by Intel Corporation. Research at the University of Catania is supported in part by the project Future Artificial Intelligence Research (FAIR) – PNRR MUR Cod. PE0000013 - CUP: E63C22001940006.



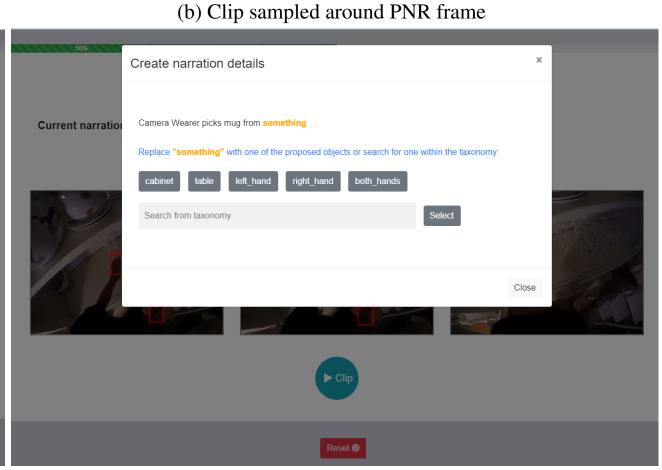
(a) Initial interface



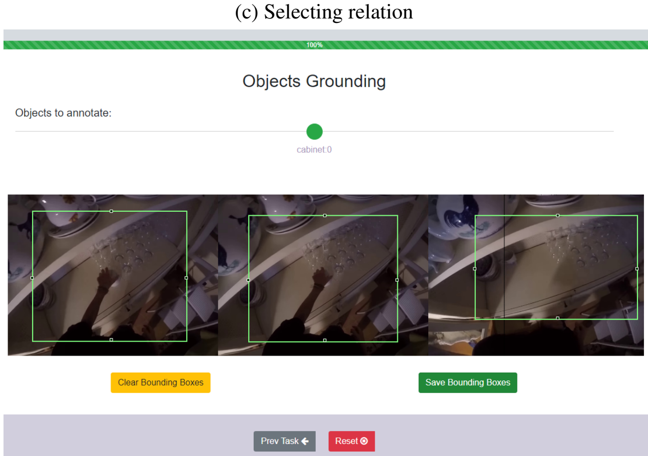
(b) Clip sampled around PNR frame



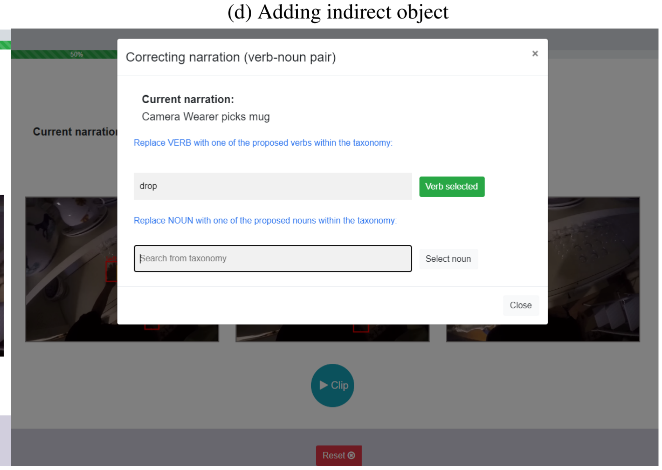
(c) Selecting relation



(d) Adding indirect object



(e) Providing grounding for indirect objects



(f) Interface for correcting initial verb-noun pair

Figure 6. The procedure which the annotators have to follow in order to provide scene graph annotations. (a) Initial interface providing instructions. (b) A video clip sampled around the PNR frame is shown. (c) The annotator can select among a set of possible relations. (d) Indirect objects are added by selecting among a list of proposals extracted from narrations or searching from taxonomy. (e) Each indirect object is grounded by the annotator in the three PRE, PNR, and POST frames. (f) In case the provided verb-noun pair is incorrect, the annotators can specify an alternative correct pair.

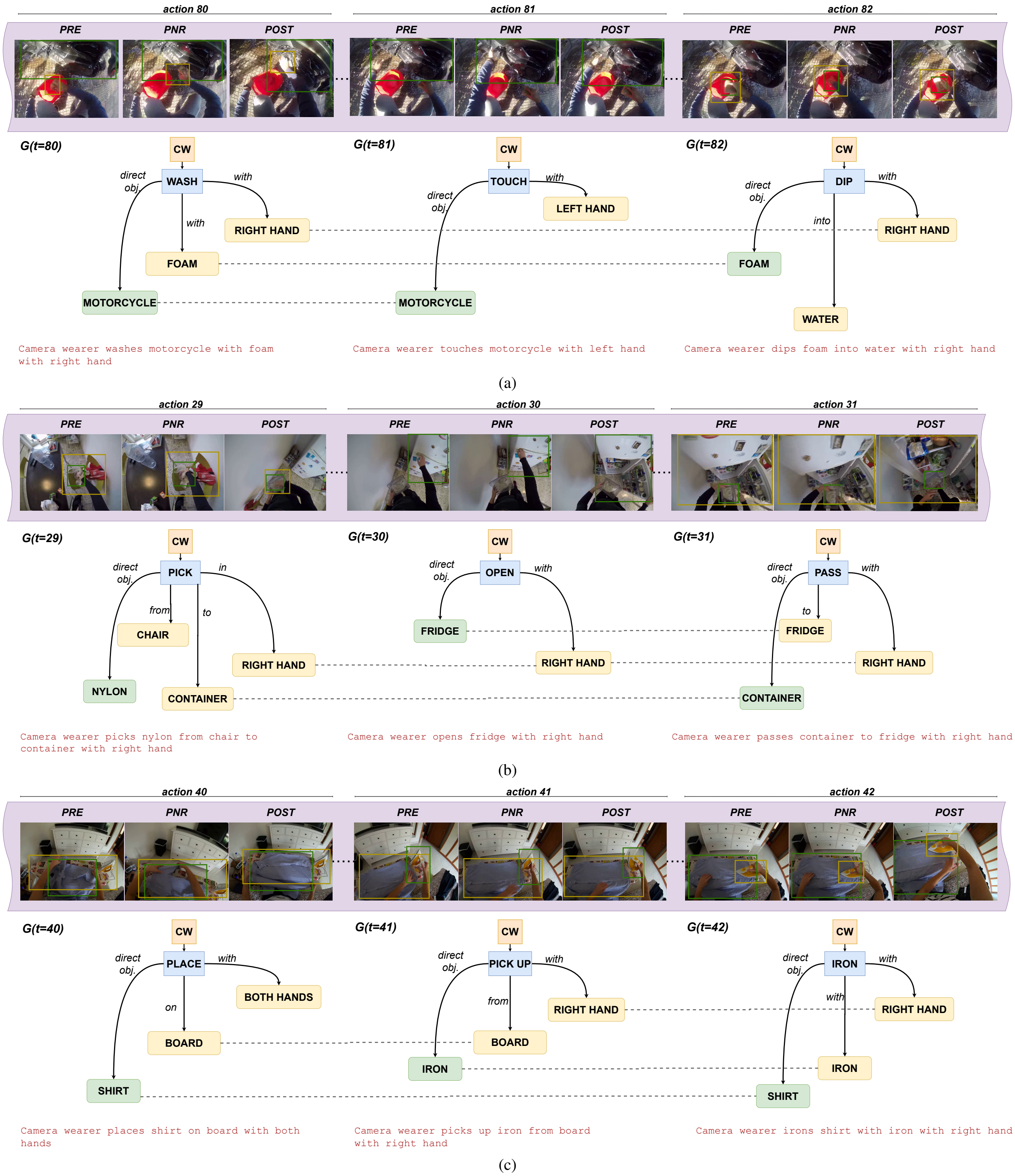


Figure 7. Sample subsequences from the Ego4D-EASG dataset. The temporal nature of Egocentric Action Scene Graphs allows to model long-form relations between the objects in the scene and the camera wearer: the indirect objects may become direct (a, b), and vice versa (b,c).

Table 6. Extensive list of all verbs, relations, and object nouns in Ego4D-EASG dataset

Verbs	add, adjust, align, apply, arrange, attach, beat, bend, break, bring, bring-out, brush, carry, carry-out, carry-up, carve, change, check, chop, clean, clean-off, clear, climb, close, collect, connect, cover, crumple, cut, cut-off, cut-out, detach, dip, dip in, disconnect, divide, drag, drill, drive, drop, drop-out, drop up, dry, dust, empty, examine, fasten, feel, fetch, fill-up, fit, fix, flap, flip, fold, force, glue, grab, grasp, grip, hammer, hang, hit, hold, hold-up, insert, inspect, iron, join, keep, knead, knit, lay, leave, lift, lift-up, loose, loosen, loosen out, losse, lower, mark, measure, mix, mount, move, move-off, move-up, open, operate, pack, paint, pass, peel, pet, pick, pick-out, pick-up, place, place down, plaster, play, point, position, pour, pour down, pour-in, pour-off, pour-out, press, pull, pull-out, push, push-down, push-in, put, put-away, put-down, put-in, put-off, put-on, put-out, raise, read, release, remove, reposition, rest, return, rinse, roll, rotate, rub, sand, scan, scoop, scoop-out, scrap, scrape, scratch, screw, screw-in, scrub, search, separate, seperate, set, sew, shake, shape, shave, shift, shuffle, slice, slide, smoothen, soak, spin, split, spray, spread, spread out, sprinkle, squeeze, squeeze out, stick, stir, store, straighten, straighten-out, streche, stretch, sweep, swing, swirl, switch, switch-off, take, take-off, take-out, take-up, tap, taste, tear-off, test, throw, throw-away, tie, tight, tighten, tilt, touch, transfer, trim, turn, turn off, turn out, turn over, twist, unfold, unhang, unlock, unplug, unscrew, untangle, untie, untighten, unwrap, uproot, use, wash, water, wear, wet, wipe, wipe-off, withdraw, wrap
Relations	direct object, verb, around, from, in, inside, into, on, onto, out, through, to, towards, under, up, verb, with
Objects	both hands, left hand, right hand,adapter, apron, art, bag, bar, basket, battery, bed, belt holder, bench, bicycle, bicycle wheel, bike, bike part, bin, board, bobbin, bolt, book, booklet, bookshelf, bottle, bowl, bowls, box, boxer, brake, brake shoe pack, branch, bread, brick, broom, brush, bucket, bulb, bunch, button, cabbage, cabinet, cable, cables, caliper, camera, can, cap, car, car part, carburetor, card, cardboard, carpet, carton, case, casing, cello, cement, chaff, chain, chair, charger, chips, chisel, chopping board, clamp, cleaner, clip, cloth, clothe, clothes, clothing material, compartment, computer, connector, container, control, cooker, cord, cot, counter, countertop, cover, cracker, craft, crumbs, cup, cupboard, cutter, cutting board, debris, derailleur, desk, detergent, dirt, dish, dog, door, dough, dough strip, dough strips, drawer, dress, drill, drill bit, driller, drink, driver, drum, dust, dustbin, dustpan, egg, engine, fabric, fabrics, faucet, fence, file, filter, finger, floor, flour, flower, foam, food, fork, fridge, fuel, furniture, gasket, gauge, gear, generator, glass, glasses, glove, glue, gouge, grass, grater, grease, grinder, ground, guitar, hammer, hand, handle, hanger, heap, hoe, hoes, holder, hose, ice, ice cubes, ice tray, insulator, iron, iron box, ironbox, ironing board, jack, jacket, jar, jug, keg, key, keyboard, knife, knob, knot, ladder, laptop, layer, leaf, leg, lever, lid, lift, light, liquid, lock, machine, manual, marker, mask, mat, matchstick, material, measure, meat, metal, metal board, milk, mirror, mixer, mixture, mop stick, motorbike, motorcycle, mouse, mouth, mower, mug, multimeter, nail, napkin, needle, net, newspaper, note, nozzle, nut, nylon, oil, onion, oven, pack, pad, paddle, paint, paint brush, paintbrush, palette, pan, pants, paper, papers, part, pastry, pedal, peel, peeler, pen, pencil, phone, photo, picture, piece, piece of cloth, pieces, pile, piler, pin, pipe, pizza, plank, plant, planter, plastic wraps, plate, plates, platform, plier, pliers, plug, pocket, pole, polythene, pot, potato, pruner, pruning sheer, pump, purse, rack, rag, rail, railing, rake, refrigerator, rim, ring, rod, rod metal, roller, room, root, rope, ropes, rubber, ruler, sachet, salt, sand, sandpaper, sauce, saucer, saw, scaffold, scale, scarf, scissors, scoop, scooter, scourer, scraper, scrapper, screw, screwdriver, seat, seed, sellotape, serviette, shaft, shear, shears, sheet, shelf, shelve, shirt, short, side, sieve, sink, sink faucet, slab, smartphone, soap, sock, socket, soil, spanner, spatula, spice, sponge, spoon, spring, stack, stairs, stand, steel, stick, stool, stove, strand, string, sugar, switch, table, table cloth, tablet, tag, tank, tap, tape, terminal, thread, tie, timber, timer, tin, tire, tissue, tomato, toolbox, tools, top, torch, towel, toy, train, trash, tray, trey, trimmer, trolley, trouser, trowel, trunk, tub, tube, tyre, umbrella, vacuum cleaner, valve, vase, vice, vine, waist, wall, wallet, wardrobe, washer, water, water hose, weed, wheel, windshield, wipe, wiper, wire, wire cutter, wires, with, wood, wood plank, wooden block, wooden stand, workbench, worktop, wrapper, wrench, yarn, yeast

Table 7. Examples of prompts and outputs for the anticipation task.

System prompt	Input Example	Completion Example
You are an assistant which models human behaviour very well. You'll be provided with a sequence of graphs (1..N-1) describing the actions retrieved from a first-person view video. Your task is to predict the next graph (N).	Example: Graph 1: Camera wearer - verb - take; take - direct object - flour; take - from - package; take - with - right hand Graph 2: Camera wearer - verb - add; add - direct object - flour; add - to - bowl; bowl - with - dough; add - with - right hand Graph 3: Camera wearer - verb - press; press - direct object - dough; press - with - both hands Graph 4: Camera wearer - verb - move; move - direct object - dough; move - from - bowl; move - to - scale Graph 5: Camera wearer - verb - move; move - direct object - dough; move - from - bowl; move - to - scale	Prediction: Graph 6: Camera wearer - verb - remove; remove - direct object - dough; remove - from - scale; remove - to - bowl
You are an assistant which models human behaviour very well. You'll be provided with a sequence of verb-noun pairs (1..N-1) describing the actions retrieved from a first-person view video. Your task is to predict the next action (N).	Example: Action 1: take flour Action 2: add flour Action 3: press dough Action 4: move dough Action 5: put dough	Prediction: Action 6: remove dough

Table 8. Examples of prompts and outputs for the summarization task.

System prompt	Input Example	Completion Example
You are an assistant who can model human behaviour very well. You'll be provided with a sequence of actions retrieved from a first-person view video. Your task is to understand the general activity and describe it in one sentence. Please, provide a very general summary and try to avoid listing all the "atomic" activities.	Example: Action 1: Camera wearer pick up hose Action 2: Camera wearer point hose towards car Action 3: Camera wearer spray car with water hose Action 4: Camera wearer wash car Action 5: Camera wearer raise wiper Action 6: Camera wearer wash car Action 7: Camera wearer push down wiper	Summary: Camera wearer is washing and cleaning a car with a water hose and wiper.
You are an assistant which can model human behaviour very well. You'll be provided with a sequence of verb-noun pairs describing the actions retrieved from a first-person view video. Your task is to understand the general activity and describe it in one sentence. Please, provide a very general summary and try to avoid listing all the "atomic" activities.	Example: Action 1: pick up hose Action 2: point hose Action 3: spray car Action 4: wash car Action 5: raise wiper Action 6: wash car Action 7: push down wiper	Summary: Camera wearer is washing and cleaning a car with a water hose and wiper.

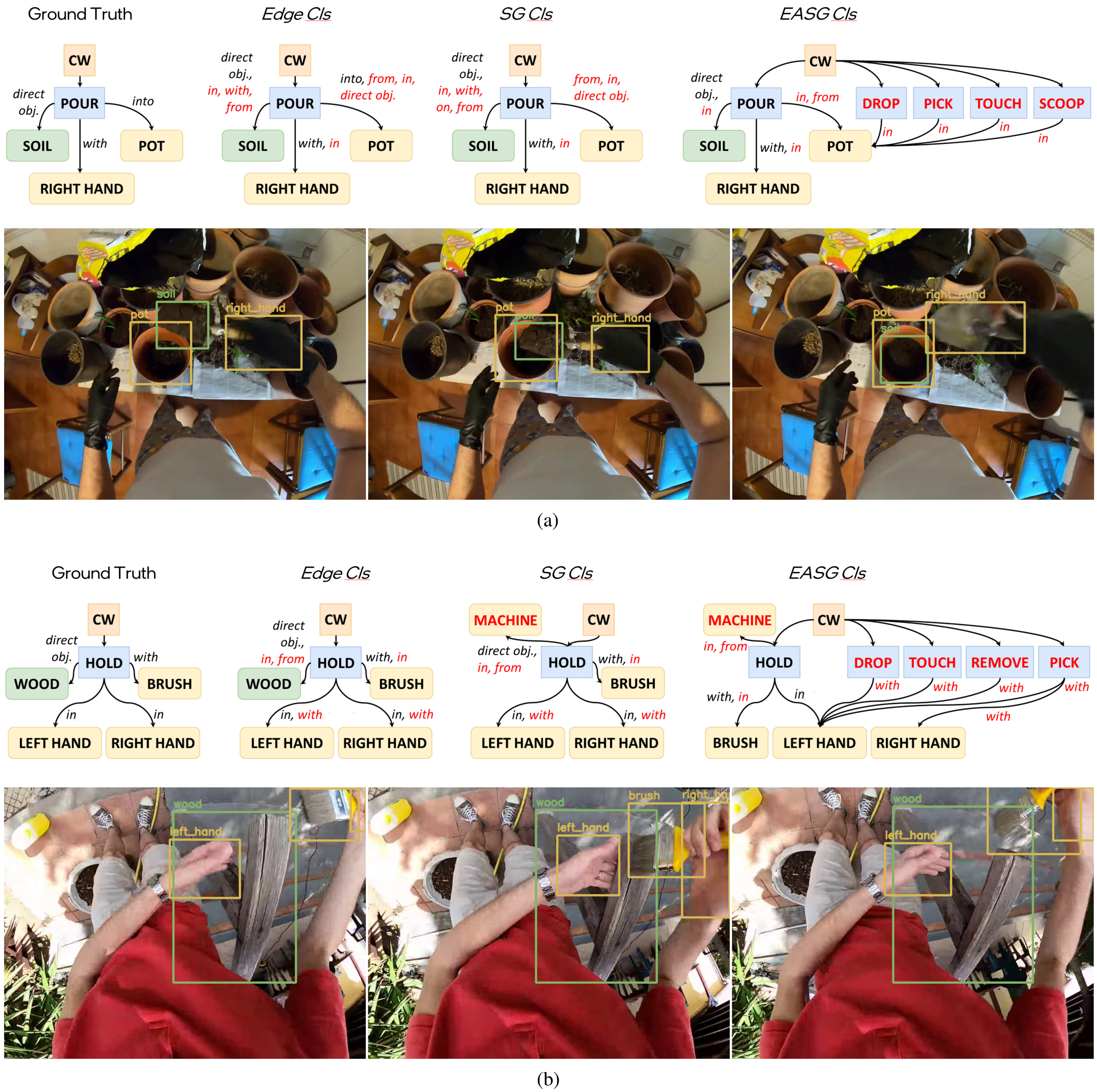


Figure 8. Qualitative results of our baseline model for the three EASG generation tasks (i.e. *Edge Cls*, *SG Cls*, and *EASG Cls*). (a): An example clip of "Camera wearer pours soil into the pot with right hand.", (b): An example clip of "Camera wearer holds wood with the brush in left and right hands." Only the top 10 predictions are illustrated in each graph, and texts in red color denote the false positives. We can observe that the generated graphs for the *EASG Cls* task have many false positives when compared to the other two tasks.

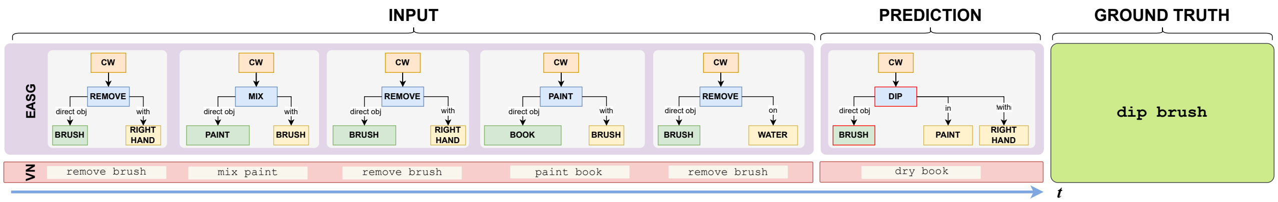


Figure 9. Qualitative example of input sequences and outputs produced using the EASG (top) and verb-noun (bottom) representations for action anticipation. The additional context provided by indirect objects and relations allows the model to predict a more meaningful future action.

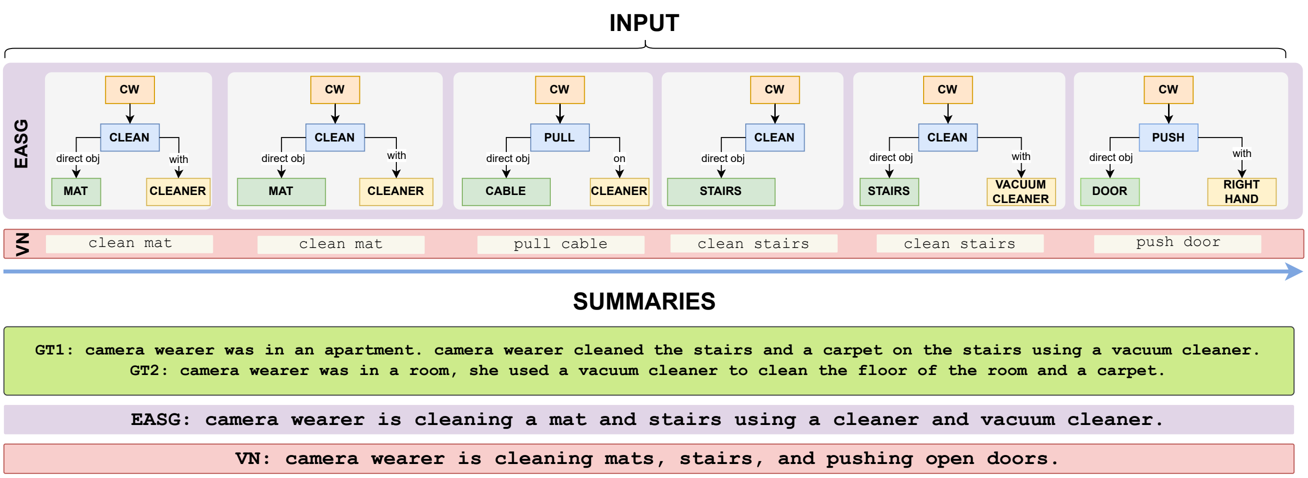


Figure 10. Qualitative example of input sequences and outputs produced using the EASG (top) and verb-noun (bottom) representations for video summarization, along with the reference summaries (in green). Even a single node in EASG (*vacuum cleaner*) may provide an important context for a better understanding of the whole activity.