Supplementary Material for
Egocentric Action Recognition by Capturing Hand-Object Contact and Object State

Tsukasa Shiota Motohiro Takagi Kaori Kumagai Hitoshi Seshimo Yushi Aono
NTT Human Informatics Laboratories, NTT Corporation
{tsukasa.shiota, motohiro.takagi, kaori.kumagai, hitoshi.seshimo, yushi.aono}@ntt.com

1. Extended Qualitative Analysis

Additional qualitative analysis was done on the EGTEA and MECCANO dataset, *i.e.*, in a cooking and industrial-like domain.

Figure A1 shows example visualizations in a cooking domain. The SlowFast model trained on both the HOCL and OSL methods output the correct action labels; the results show that the model focused on very different aspects of the video depending on whether it was trained on action learning or on HOCL and OSL. Specifically, SlowFast trained on only action learning output a label in response to the food bag, which is unrelated to the fridge or person's hands. This is evidence that the model was trained to capture irrelevant information without considering hand-object interaction. On the other hand, when SlowFast was constrained to focus on the hands and objects due to being trained on HOCL and OSL, it focused on the fridge and hands to determine the action label. Figure A2 shows additional example visualizations in which the SlowFast correctly predicted the action label under all training conditions. The model without being trained on HOCL and OSL focused only on the point where the left hand holds the condiment container in the second frame of the video, whereas when it was trained on both HOCL and OSL, it focused on all the places where the condiment container is placed.

Figure A3 and A4 show example visualizations in an industrial-like domain. In Figure A3, SlowFast with only action learning incorrectly recognized the action, focusing on "partial_model," which is unrelated to "put_screwdriver." On the other hand, SlowFast trained on both HOCL and OSL precisely captures the interaction between right hand and screwdriver. Figure A4 shows example visualizations in which the SlowFast correctly predicted the action label under all training conditions. SlowFast with only action learning output a label only focusing on the last frame of the video, which means it does not observe the action of the screw being tightened. When trained on HOCL and OSL, SlowFast understood the action by focusing on the screw-tightening movements captured throughout the video.

Figure A5 shows example visualizations in which the SlowFast model incorrectly predicted the action label under all training conditions. The ground truth action for this video is "Take eating_utensil." SlowFast without HOCL and OSL incorrectly predicted the action as "Take lettuce" due to a strong response to the lettuce in the third frame. SlowFast with OSL output "Divide/Pull Apart lettuce" because it made an incorrect prediction in the second frame, judging that the lettuce and the right hand are in contact even though they are not. The label was predicted to be "Cut lettuce" when HOCL was used alone and when HOCL and OSL were both used. This is because the model recognized that the knife and lettuce interact in the video. Future work includes addressing two challenging issues: (1) making contact determination more accurate by using other modalities and (2) clarifying object-object interaction.

2. Extended Discussion of Limitations

Neither an untrained model nor one trained on the proposed methods can prioritize the actions to be recognized on the current datasets. We often perform multiple actions simultaneously. For example, we might lift a loaf of bread with our left hand and simultaneously grasp a knife with our right hand to slice it. In this situation, the video shows both "Take bread" and "Take eating_utensil"; therefore, there are two ground truth actions. However, recognition models cannot determine which action is salient. We show an example of this in Figure A6. In the video, a person holds a knife in his/her right hand while simultaneously lifting a loaf of bread with his/her left hand. The video captures both actions, namely "Take bread" and "Take eating_utensil"; therefore, two actions are recognized. However, the ground truth label is defined as only "Take eating_utensil" on the EGTEA dataset [1]. The SlowFast model under all training conditions made incorrect predictions on the action recognition task even though it correctly output the other label, "Take bread." This may result in in-

correct performance evaluation. Here, we discuss this problem from two perspectives: (1) models and (2) task definition.

(1) If we expect models to correctly identify the most salient action among multiple actions (*e.g.*, "Take eating_utensil" is more salient than "Take bread"), we need to define saliency and train models to learn the meaning. There are several possible definitions, such as actions that are more centrally imaged in the spatio-temporal direction or hand actions with large movements.

(2) In action recognition, the ground truth label is ultimately determined to be one specific label due to the nature of the action recognition task. With this setting, the performance of the model might not be correctly and sufficiently evaluated. To overcome this limitation, we suggest annotating action labels for the right hand, the left hand, and both hands and redefining them as tasks to predict each action, for example.

3. Annotated Adjective Labels

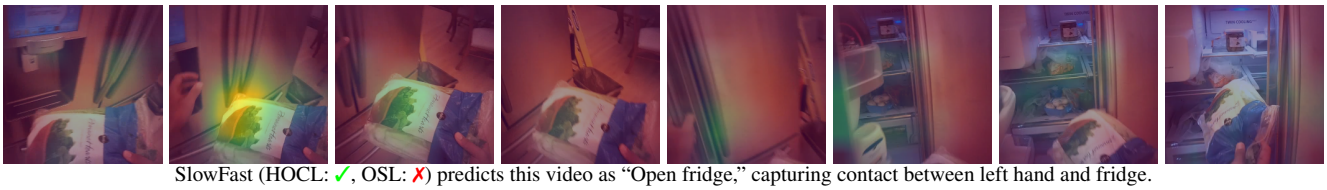
Tables A1, A2, A3, and A4 list the adjectives describing the initial and final states assigned to the action labels in the EGTEA, MECCANO, and EPIC-100 datasets. Due to the enormous number of verb-noun combinations in EPIC-100, we use asterisks to denote noun labels in Tables A3 and A4. Actions other than hand-related actions, such as those using the eyes (*e.g.*, check_booklet in MECCANO) and actions expressed with an intransitive verb (e.g., walk * in EPIC-100), do not involve a change in object state. In those cases, we define the initial and final states as "none." Note that the cost of annotating adjective labels is lower than that of annotating other labels, such as action labels, because they are allocated to each action label rather than to each instance.

References

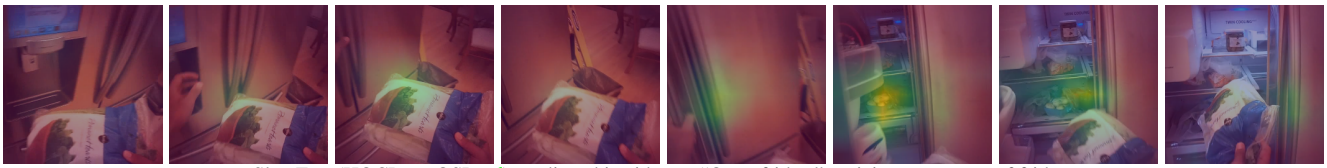
[1] Yin Li, Miao Liu, and James M. Rehg. In the Eye of Beholder: Joint Learning of Gaze and Actions in First Person Video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1

[2] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 3, 4, 5

SlowFast (HOCL: ✗, OSL: ✗) predicts this video as "Open fridge," focusing on irrelevant object.



SlowFast (HOCL: ✓, OSL: ✗) predicts this video as "Open fridge," capturing contact between left hand and fridge.



SlowFast (HOCL: ✗, OSL: ✓) predicts this video as "Open fridge," noticing open state of fridge.

SlowFast (HOCL: ✓, OSL: ✓) predicts this video as "Open fridge," being aware of action.

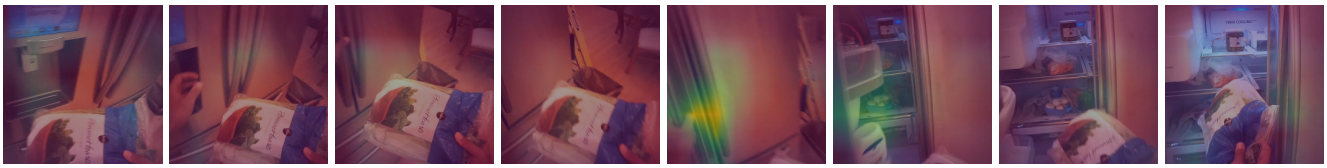


Figure A1. Visualization using GradCAM [2]. This video shows the action "Open fridge," in which a person opens a fridge with his/her left hand.

SlowFast (HOCL: ✗, OSL: ✗) predicts this video as "Put condiment_container," capturing only grasped condiment container.

SlowFast (HOCL: ✓, OSL: ✗) predicts this video as "Put condiment_container," focusing on hand-object contact.

SlowFast (HOCL: ✗, OSL: ✓) predicts this video as "Put condiment_container," being aware that condiment container has been "put".

SlowFast (HOCL: ✓, OSL: ✓) predicts this video as "Put condiment_container," capturing the action.

Figure A2. Visualization using GradCAM [2]. This video shows the action "Put condiment_container," in which a person puts a condiment container in the storage compartment of the fridge door with his/her left hand.

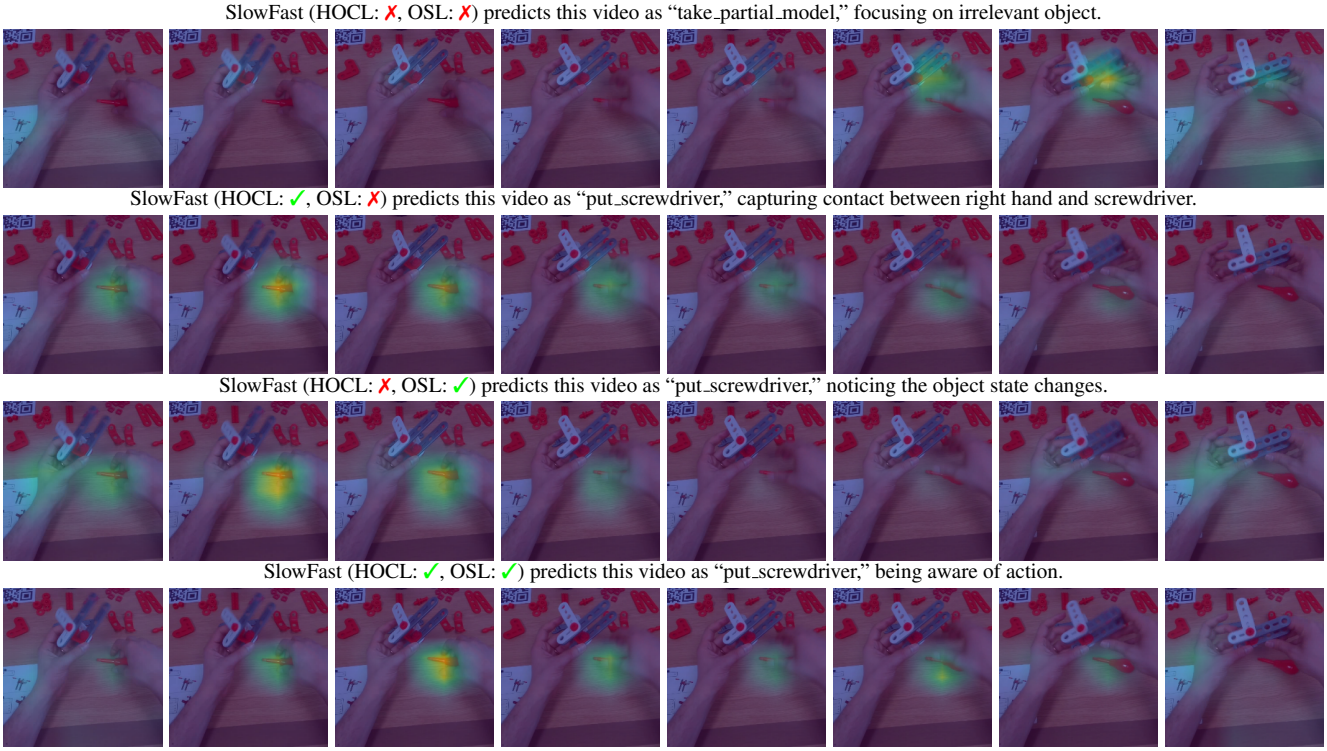


Figure A3. Visualization using GradCAM [2]. This video shows the action “put_screwdriver,” in which a person puts a screwdriver with his/her right hand.

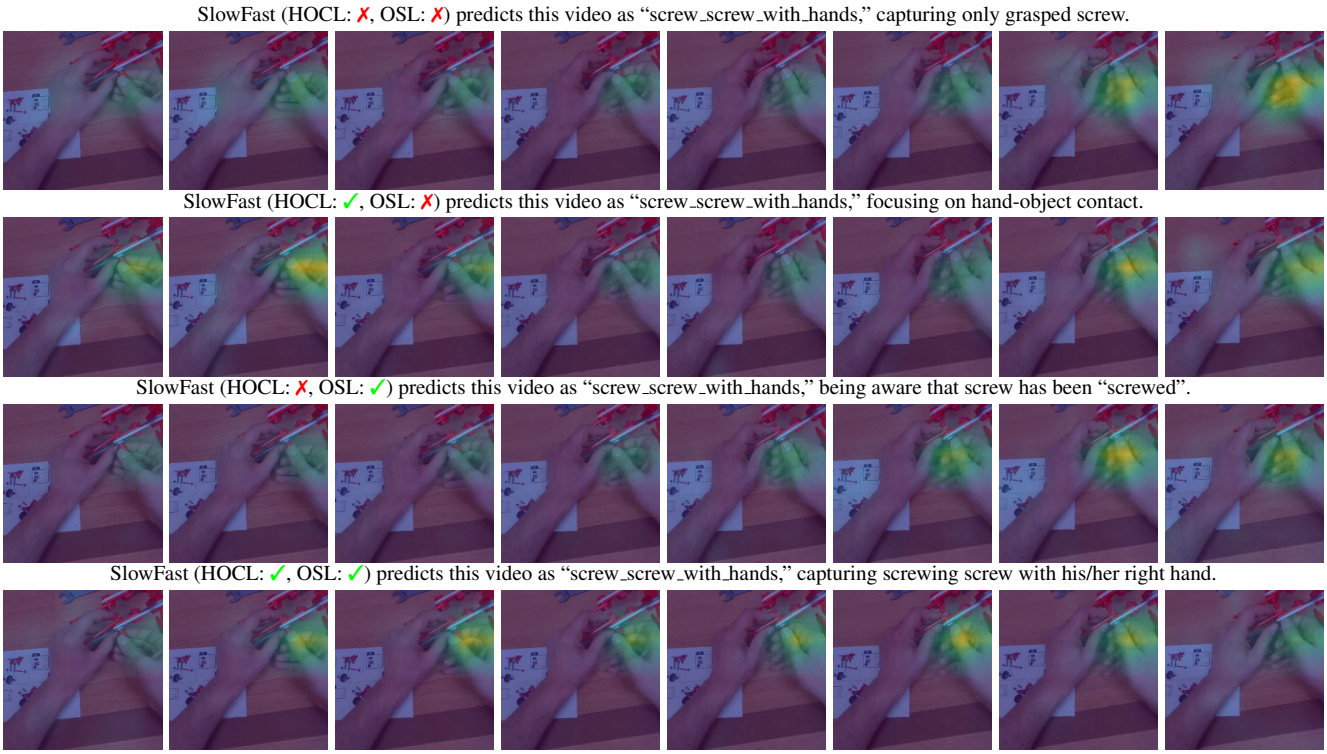


Figure A4. Visualization using GradCAM [2]. This video shows the action “screw_screw_with_hands,” in which a person screws a screw with his/her right hand.

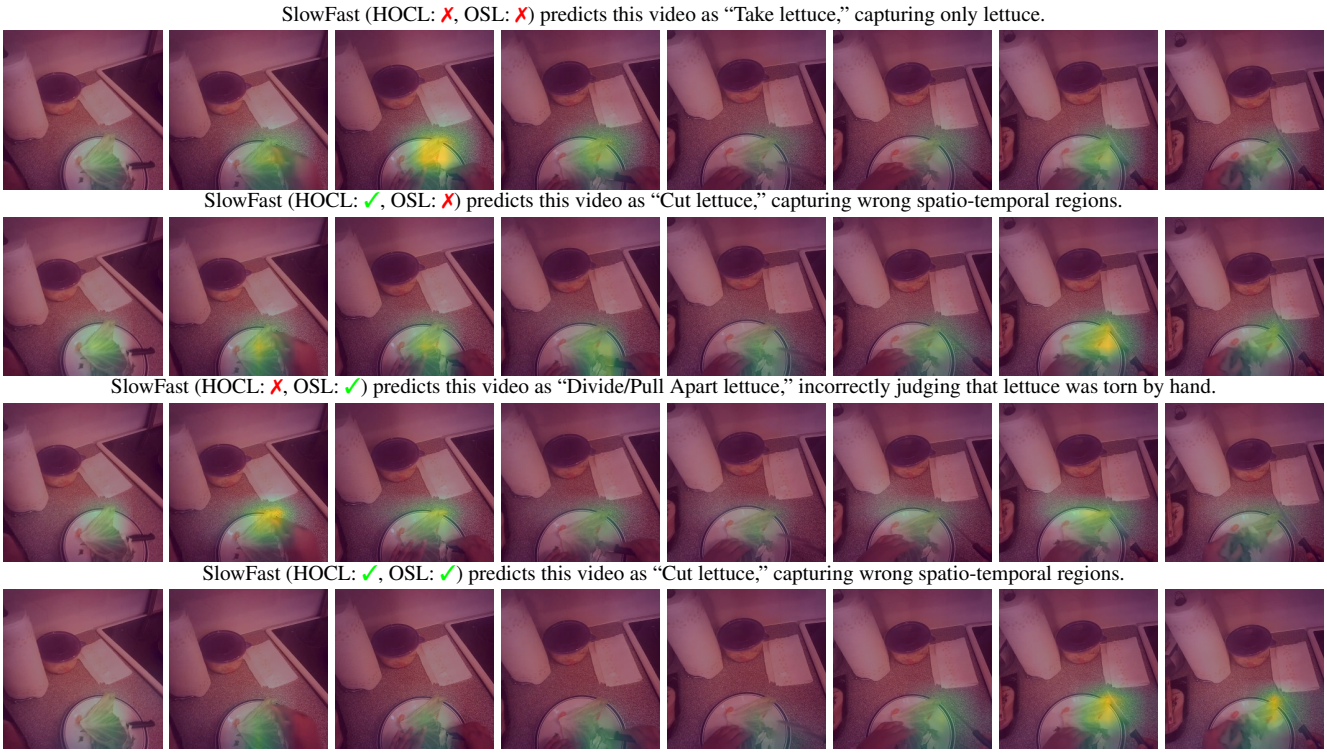


Figure A5. Visualization using GradCAM [2]. The video shows the action “Take eating_utensil,” in which a person pick up knife with his/her right hand.

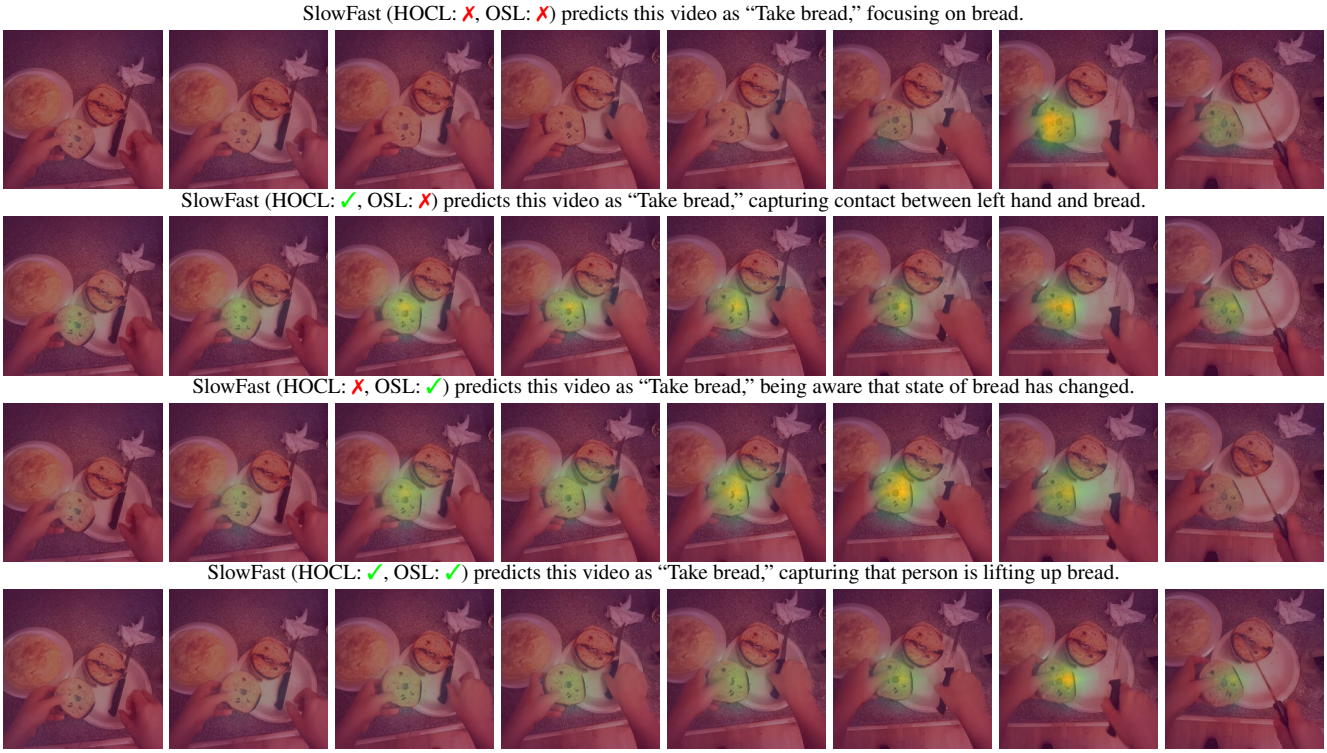


Figure A6. Visualization using GradCAM [2]. The video shows two actions, “Take bread” and “Take eating_utensil,” in which a person simultaneously picks up loaf of bread with his/her left hand and knife with his/her right hand.

Table A1. Annotated adjective labels for action labels in EGTEA dataset.

Initial state	**Final state**	**Action labels**
closed	open	Open fridge, Open cabinet, Open drawer, Open condiment_container, Open fridge_drawer, Open dishwasher, Open bread_container, Open oil_container, Open cheese_container, Turn on faucet
open	closed	Close fridge, Close cabinet, Close condiment_container, Close drawer, Close fridge_drawer, Close oil_container, Turn off faucet
put	grasped	Take eating_utensil, Take condiment_container, Take plate, Take bowl, Take paper_towel, Take cooking_utensil, Take bread, Take seasoning_container, Take cup, Take pot, Take bread_container, Take tomato, Take pan, Take sponge, Take lettuce, Take onion, Take cheese_container, Take oil_container, Take cucumber, Take bell_pepper, Take tomato_container, Take pasta_container, Take grocery_bag, Take cheese, Take cutting_board, Take egg
grasped	put	Put eating_utensil, Put condiment_container, Put bowl, Put "trash, trash_container," Put plate, Put cooking_utensil, Put pan, Put lettuce, Put pot, Put bread, Put tomato, Put cup, Put bread_container, Put sponge, Put seasoning_container, Put cutting_board, Put cheese, Put bell_pepper, Put tomato_container, Put paper_towel, Put cucumber, Put cheese_container, Put onion, Put grocery_bag, Put oil_container
unsqueezed	squeezed	Squeeze washing_liquid, Squeeze sponge
whole	cut	Cut tomato, Cut cucumber, Cut carrot, Cut onion, Cut bell_pepper, Cut lettuce, Cut olive
empty	filled	Pour "oil, oil_container, pan," Pour "condiment, condiment_container, salad," Pour "seasoning, seasoning_container, salad," Pour "water, faucet, pot"
whole	mixed	Mix egg
dirty	clean	Wash eating_utensil, Wash hand, Wash cutting_board, Wash pan, Wash bowl, Wash pot, Wash strainer, Clean/Wipe counter
gathered	spread	Spread "condiment, bread, eating_utensil"
separated	mixed	Mix "mixture, eating_utensil," Mix pasta
separated	compressed	Compress sandwich
grasped	grasped	Move Around bacon, Move Around patty, Move Around pan, Move Around eating_utensil, Move Around bowl, Move Around pot
non-operated	operated	Operate stove, Operate microwave
whole	separated	Divide/Pull Apart lettuce, Divide/Pull Apart paper_towel, Divide/Pull Apart onion, Crack egg
none	none	Inspect/Read recipe

Table A2. Annotated adjective labels for action labels in MECCANO dataset.

Initial state	**Final state**	**Action labels**
put	grasped	take_red_perforated_bar, take_screw, take_bolt, take_screwdriver, take_red_angled_perforated_bar, take_rod, take_red_perforated_junction_bar, take_partial_model, take_washer, take_tire, take_rim, take_roller, take_gray_perforated_bar, take_gray_angled_perforated_bar, take_white_angled_perforated_bar, take_booklet, take_wheels_axle, take_red_4_perforated_junction_bar, take_handlebar, take_wrench, take_objects
unplugged	plugged	plug_screw, plug_rod, plug_handlebar
separated	aligned	align_objects, align_screwdriver_to_screw, align_wrench_to_bolt
grasped	put	put_booklet, put_screw, put_screwdriver, put_partial_model, put_roller, put_washer, put_red_4_perforated_junction_bar, put_wheels_axle, put_gray_perforated_bar, put_handlebar, put_red_perforated_bar, put_red_angled_perforated_bar, put_bolt, put_white_angled_perforated_bar, put_rod, put_gray_angled_perforated_bar, put_red_perforated_junction_bar, put_rim, put_wrench, put_tire, put_objects
loosed	tightened	tighten_bolt_with_hands, screw_screw_with_screwdriver, screw_screw_with_hands, tighten_bolt_with_wrench
attached	detached	pull_partial_model, pull_screw, pull_rod
unflipped	flipped	browse_booklet
tightened	loosed	unscrew_screw_with_hands, unscrew_screw_with_screwdriver, loosen_bolt_with_hands
detached	attached	fit_rim_tire
none	none	check_booklet

Table A3. Annotated adjective labels for action labels in EPIC-100 dataset (1/2).

Initial state	Final state	Action labels
put	grasped	take *
dirty	clean	scrub *, filter *, wash *, brush *
closed	open	open *
open	closed	close *
turned-off	turned-on	turn-on *
whole	cut	cut *
turned-on	turned-off	turn-off *
empty	filled	pour *, fill *
separated	mix	mix *
grasped	grasped	use *, pull *, shake *, stretch *, carry *, slide *, lift *, move *, pat *, hold *
grasped	put	eat *, put *, drink *, insert *, let-go *
gathered	separated	remove *
grasped	thrown	throw *
wet	dried	dry *
unscooped	scooped	scoop *
unoperated	operated	adjust *
unsqueezed	squeezed	squeeze *
unpeeled	peeled	peel *
filled	empty	empty *
unpressed	pressed	press *
unflipped	flipped	flip *
unturned	turned	turn *
unscraped	scraped	scrape *
gathered	spread	apply *
unfolded	folded	fold *, bend *
whole	separated	rip *, divide *, break *
unwrapped	wrapped	wrap *
unchecked	checked	look *
folded	unfolded	unroll *
unarranged	arranged	form *, sort *
grasped	hanged	hang *
unsprinkled	sprinkled	sprinkle *
unsprayed	sprayed	spray *

Table A4. Annotated adjective labels for action labels in EPIC-100 dataset (2/2).

Initial state	Final state	Action labels
unheated	heated	cook *, bake *
separated	mixed	add *
unrolled	rolled	roll *
uncrushed	crushed	crush *
mixed	mixed	knead *
unset	set	set *
untouched	touched	feel *
unrubbed	rubbed	rub *
unimmersed	immersed	soak *
unsharpened	sharpened	sharpen *
grasped	dropped	drop *
dried	wet	water *
spread	gathered	gather *
unattached	attached	attach *
strong	weak	turn-down *, lower *
undipped	dipped	coat *
unequipped	equipped	wear *
weak	strong	increase *
tightened	loosed	unscrew *
whole	grated	grate *
loosed	tightened	screw *
worked	stopped	finish *
unpierced	pierced	stab *
unserved	served	serve *
wrapped	unwrapped	uncover *, unwrap *
unlocked	locked	lock *
unflatten	flatten	flatten *
unseasoned	seasoned	season *
locked	unlocked	unlock *
unmarked	marked	mark *
frozen	unfrozen	unfreeze *
none	none	switch *, prepare *, choose *, wait *, search *, transition *, smell *, measure *, check *