A. Dataset and Metrics

We test on four datasets as described in the main paper. *EpicKitchens-100 (EK100)* [14] is the largest egocentric (first-person) video dataset with 700 long unscripted videos of cooking activities totalling 100 hours. It contains 89,977 segments labeled with one of 97 verbs, 300 nouns, and 3807 verb-noun combinations (or "actions"), and uses τ_a=1s. The dataset is split in 75:10:15 ratio into train/val/test sets, and the test set evaluation requires submission to the CVPR'21 challenge server. The evaluation metric used is class-mean recall@5 [22], which evaluates if the correct future class is within the top-5 predictions, and equally weights all classes by averaging the performance computed individually per class. The top-5 criterion also takes into account the multi-modality in the future predictions. Entries are ranked according to performance on *actions*.

EpicKitchens-55 (EK55) [13] is an earlier version of the EK100, with 39,596 segments labeled with 125 verbs, 352 nouns, and 2,513 combinations (actions), totalling 55 hours, and $\tau_a = 1s$. We use the standard splits and metrics from [24]. For anticipation, [24] splits the public training set into 23,493 training and 4,979 validation segments from 232 and 40 videos respectively. The test evaluation is similarly performed on the challenge server. The evaluation metrics used are top-1/top-5 accuracies and class-mean recall@5 over verb/noun/action predictions at anticipation time $\tau_a = 1s$. Unlike EK100, the recall computation on EK55 is done over a subset of 'many-shot' classes as defined in [23]. While EK55 is a subset of EK100, we use it to compare to a larger set of baselines, which have not yet been reported on EK100.

EGTEA Gaze+ [53] is another popular egocentric action anticipation dataset, consisting of 10,325 action annotations with 106 unique actions. To be comparable to prior work [55], we report performance on the split 1 [53] of the dataset at $\tau_a = 0.5s$ using overall top-1 accuracy and mean over top-1 class accuracies (class mean accuracy).

Finally, we also experiment with a popular third-person action anticipation dataset: *50-Salads (50S)* [79]. It contains $50 \sim 40s$ long videos, with 900 segments labeled with one of 17 action classes. We report top-1 accuracy averaged over the pre-defined 5 splits for an anticipation time $\tau_a = 1s$, following prior work [2, 74].

B. Baselines Details

RULSTM leverages a 'rolling' LSTM to encode the past, and an 'unrolling' LSTM to predict the future, from different points in the past. It was ranked first in the EK55 challenge in 2019, and is currently the best reported method on EK100. ActionBanks [74] improves over RULSTM through a carefully designed architecture leveraging non-local [90] and long-term feature aggregation [92] blocks over different lengths of past features, and was one of the winners of the CVPR'20 EK55 anticipation challenge. Forecasting HOI [55] takes an alternate approach, leveraging latest spatio-temporal convnets [84] jointly with hand motion and interaction hotspot prediction.

	Verb		Noun		Action	
Method	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
DMR [87]	-	73.7	-	30.0	-	16.9
ATSN [13]	-	77.3	-	39.9	-	16.3
ED [26]	-	75.5	-	43.0	-	25.8
MCE [22]	-	73.4	-	38.9	-	26.1
FN [15]	-	74.8	-	40.9	-	26.3
RL [59]	-	76.8	-	44.5	-	29.6
EL [38]	-	75.7	-	43.7	-	28.6
FHOI (I3D) [55]	30.7	76.5	17.4	42.6	10.4	25.5
RULSTM [23, 24]	32.4	79.6	23.5	51.8	15.3	35.3
ImagineRNN [93]	-	-	-	-	-	35.6
ActionBanks [74]	**35.8**	**80.0**	23.4	52.8	15.1	35.6
AVT+	32.5	79.9	**24.4**	**54.0**	**16.6**	**37.6**

Table 9: EK55 (val) results reported in top-1/5 (%) at $\tau_a = 1.0s$ as summarized in in Table 5 (left) in the main paper.

C. Results

C.1. EpicKitchens-55 Full Results

Table 9 and Table 10 report the full comparison to the state-of-the-art on EpicKitchens-55 validation and test sets respectively for all label spaces: verb/noun/actions, which was summarized in Table 5 in the main paper. Note that our models are only trained for action prediction, and individual verb/noun predictions are obtained by marginalizing over the other. We outperform all prior work on on seen test set (S1), and are only second to concurrent work [16] on unseen (S2) for top-1 actions. It is worth noting that [16] uses transductive learning, leveraging the test set. AVT is also capable of similarly leveraging the test data with unsupervised objectives ($\mathcal{L}_{feat}$), which could potentially further improve in performance. We leave that exploration to future work.

D. Analysis

D.1. Per-class Gains

To better understand the source of these gains, we analyze the class-level gains with anticipative training in Figure 6. We notice certain verb classes show particularly large gains across the backbones, such as 'cook' and 'choose'. We posit that is because predicting the person will cook an item would often require understanding the sequence of ac-

	Seen test set (S1)						Unseen test set (S2)					
	Top-1 Accuracy%			Top-5 Accuracy%			Top-1 Accuracy%			Top-5 Accuracy%		
	Verb	Noun	Act.	Verb	Noun	Act.	Verb	Noun	Act.	Verb	Noun	Act.
2SCNN [13]	29.76	15.15	4.32	76.03	38.56	15.21	25.23	9.97	2.29	68.66	27.38	9.35
ATSN [13]	31.81	16.22	6.00	76.56	42.15	28.21	25.30	10.41	2.39	68.32	29.50	6.63
ED [26]	29.35	16.07	8.08	74.49	38.83	18.19	22.52	7.81	2.65	62.65	21.42	7.57
MCE [22]	27.92	16.09	10.76	73.59	39.32	25.28	21.27	9.90	5.57	63.33	25.50	15.71
P+D [61]	30.70	16.50	9.70	76.20	42.70	25.40	28.40	12.40	7.20	69.80	32.20	19.30
RULSTM [23, 24]	33.04	22.78	14.39	79.55	50.95	33.73	27.01	15.19	8.16	69.55	34.38	21.10
ActionBanks [74]	**37.87**	24.10	16.64	79.74	**53.98**	36.06	29.50	16.52	10.04	70.13	37.83	23.42
FHOI [55]	34.99	20.86	14.04	77.05	46.45	31.29	28.27	14.07	8.64	70.67	34.35	22.91
FHOI+obj [55]	36.25	23.83	15.42	79.15	51.98	34.29	29.87	16.80	9.94	71.77	38.96	23.69
ImagineRNN [93]	35.44	22.79	14.66	79.72	52.09	34.98	29.33	15.50	9.25	70.67	35.78	22.19
Ego-OMG [16]	32.20	**24.90**	16.02	77.42	50.24	34.53	27.42	**17.65**	**11.81**	68.59	37.93	23.76
AVT+	34.36	20.16	**16.84**	**80.03**	51.57	**36.52**	**30.66**	15.64	10.41	**72.17**	**40.76**	**24.27**

Table 10: EK55 test set results obtained from the challenge server, as summarized in Table 5 (right) in the main paper. AVT outperforms all published work on this dataset on top-5 metric, and is only second to [16] on S2 on top-1. Note that [16] leverages transductive learning (using the test set for initial graph representation learning), whereas AVT only uses the train set.

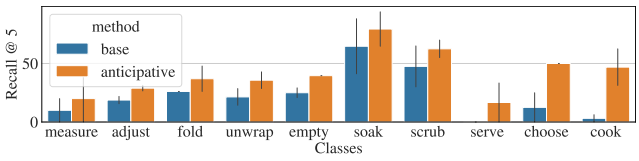

Figure 6: Verb classes that gain the most with causal modeling, averaged over the TSN and AVT-b backbones. Actions such as 'cook' and 'choose' show particularly significant gains.

tions so far, such as preparing ingredients, turning on the stove *etc*., which the anticipative training setting encourages.

D.2. Attention Visualizations

In Figure 9 we show additional visualizations of the spatial and temporal attention, similar to Figure 1. We also provide an attached video to visualize predicted future classes along with the ground truth (GT) future prediction in a video form for EK100 and EGTEA Gaze+, at each time step (as opposed to only 2 shown in these figures).

D.3. Long-term Anticipation

In Figure 8 we show additional visualizations of the long-term anticipation, similar to Figure 5.

D.4. $\mathcal{L}_{feat}$ Formulation

In Figure 7 we show the performance of AVT with both AVT-b and TSN backbones, using two different loss functions for $\mathcal{L}_{feat}$: L_2 as used in paper, and InfoNCE [64] objective as in some recent work [36, 93], at different weights used on that loss during training. We find that L_2 is as effective or better for both backbones, and hence we use it with weight=1.0 for all experiments. While further hyperparameter tuning can potentially lead to further improvements for InfoNCE as observed in some concurrent work [93], we leave that exploration to future work.

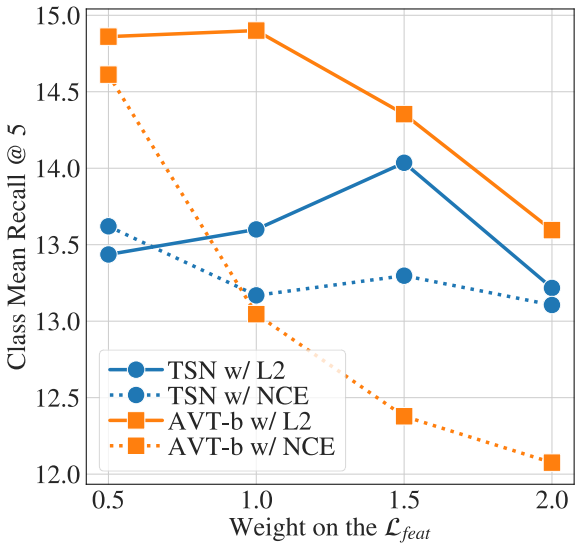

Figure 7: Different $\mathcal{L}_{feat}$ functions and weights. We found similar or better performance of the simpler L_2 metric over NCE and use it for all experiments in the paper. The graph here shows performance on EK100 (validation, RGB) at $\tau_a = 1s$, at different scalar weights used on this loss during optimization.

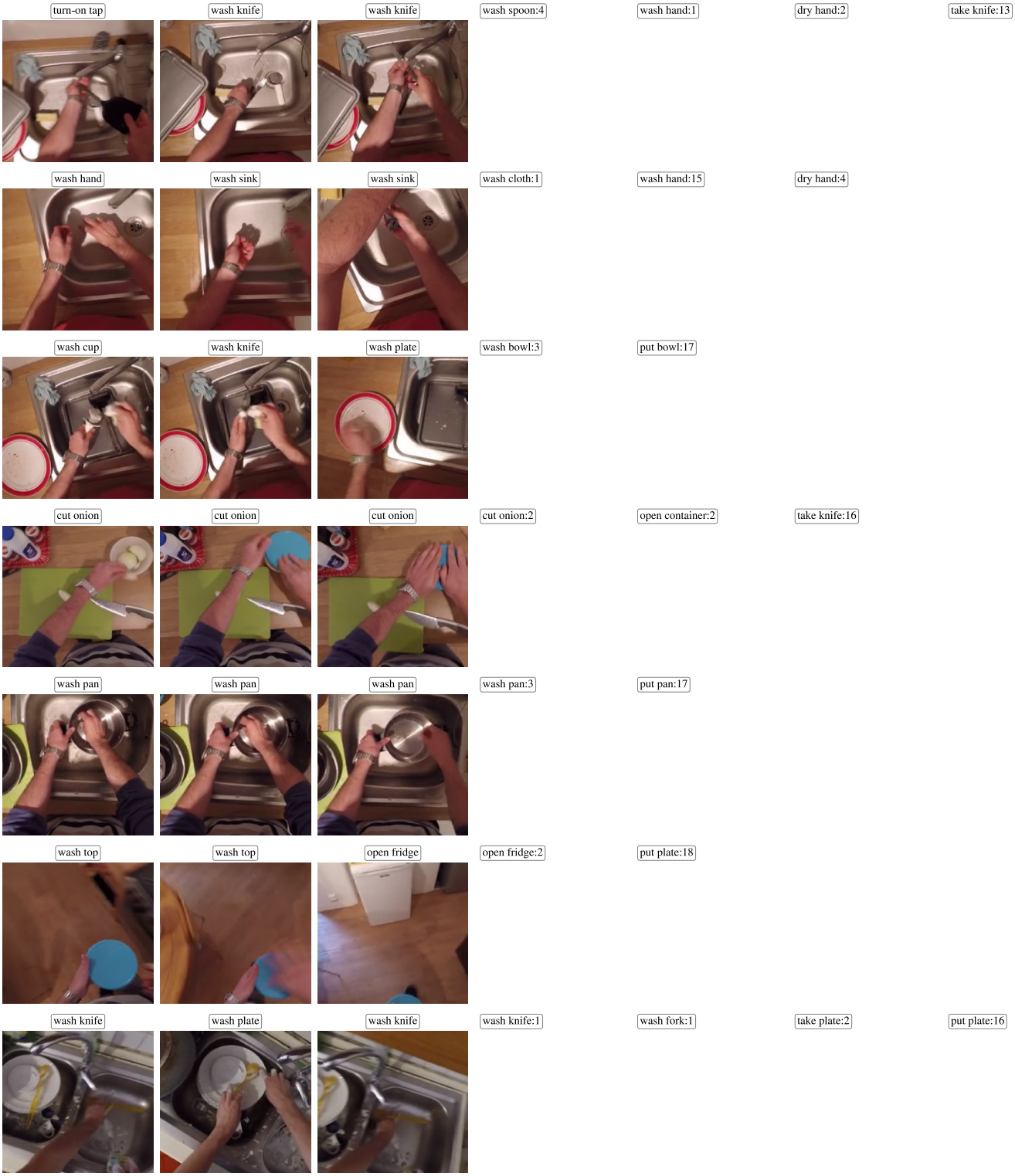

Figure 8: Long-term anticipation. Additional results continued from Figure 5 on EK100. On top of each frame, we show the *future* prediction at that frame (not the action that is happening in the frame, but what the model predicts will happen next). The following text boxes show the future predictions made by the model by rolling out autoregressively, using the predicted future feature. The number next to the rolled out predictions denotes for how many time steps that specific action would repeat, according to the model. For example, 'wash spoon: 4' means the model anticipates the 'wash spoon' action to continue for next 4 time steps. As we can observe, AVT makes reasonable future predictions, such as 'put pan' would follow 'wash pan'; 'dry hand' would follow 'wash hand' *etc*. This suggests the model has picked up on action schemas [66].

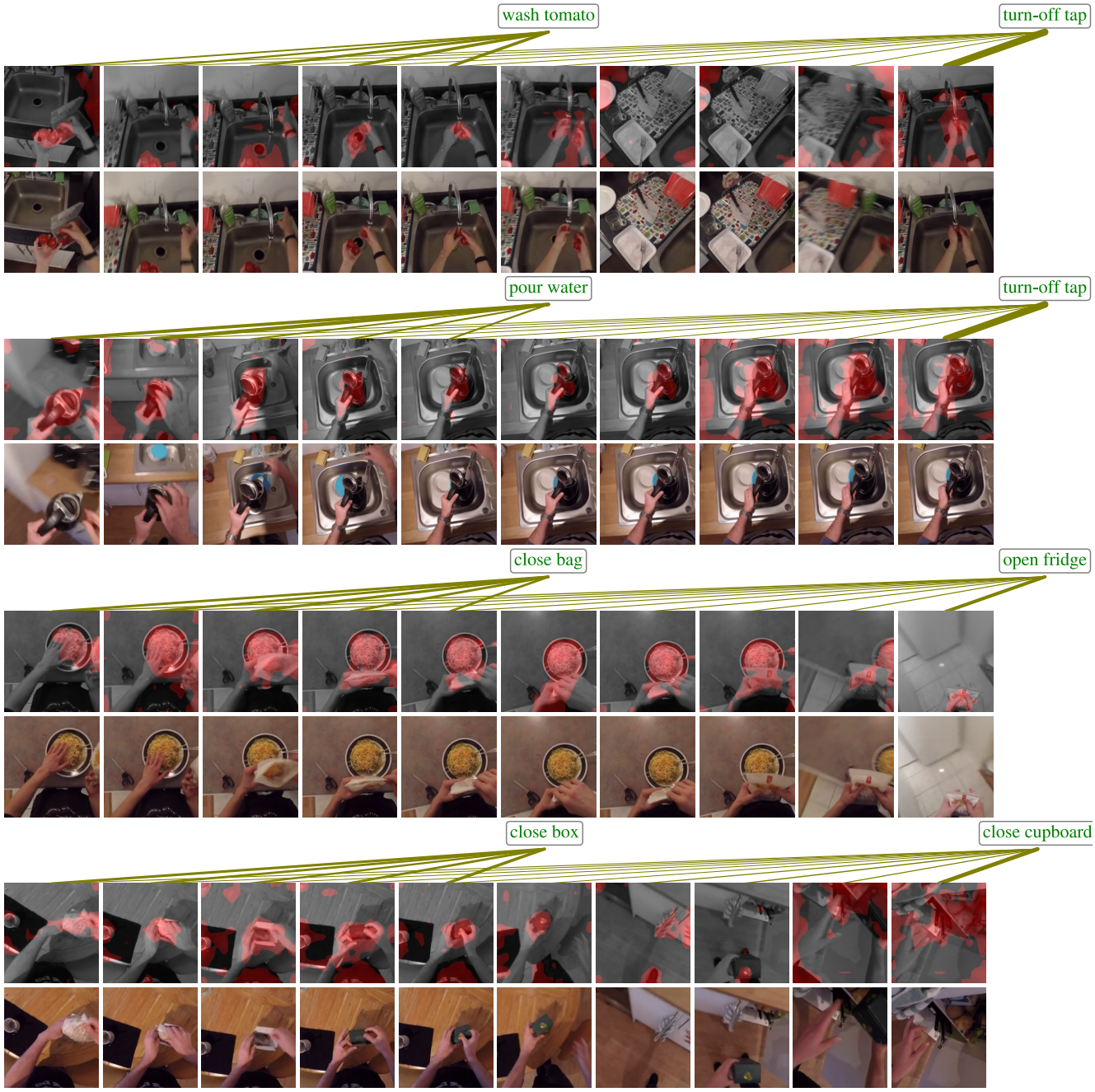

Figure 9: More Qualitative Results. The spatial and temporal attention visualization in EK100, similar to Figure 1. For each input frame, we visualize the effective spatial attention by AVT-b using attention rollout [1]. The red regions represent the regions of highest attention, which we find to often correspond to hands+objects in the egocentric EpicKitchens-100 videos. The text on the top show future predictions at 2 points in the video, along with the temporal attention (last layer of AVT-h averaged over heads) visualized using the width of the lines. The green color of text indicates that it matches the GT action at that future frame (or that nothing is labeled at that frame). As seen in Figure 1, spatial attention focuses on hands and objects. The temporal attention focuses on the last frame when predicting actions like 'turn-off tap', whereas more uniformly on all frames when predicting 'open fridge' (as an action like that usually follows a sequence of actions involving packing up food items and moving towards the fridge).

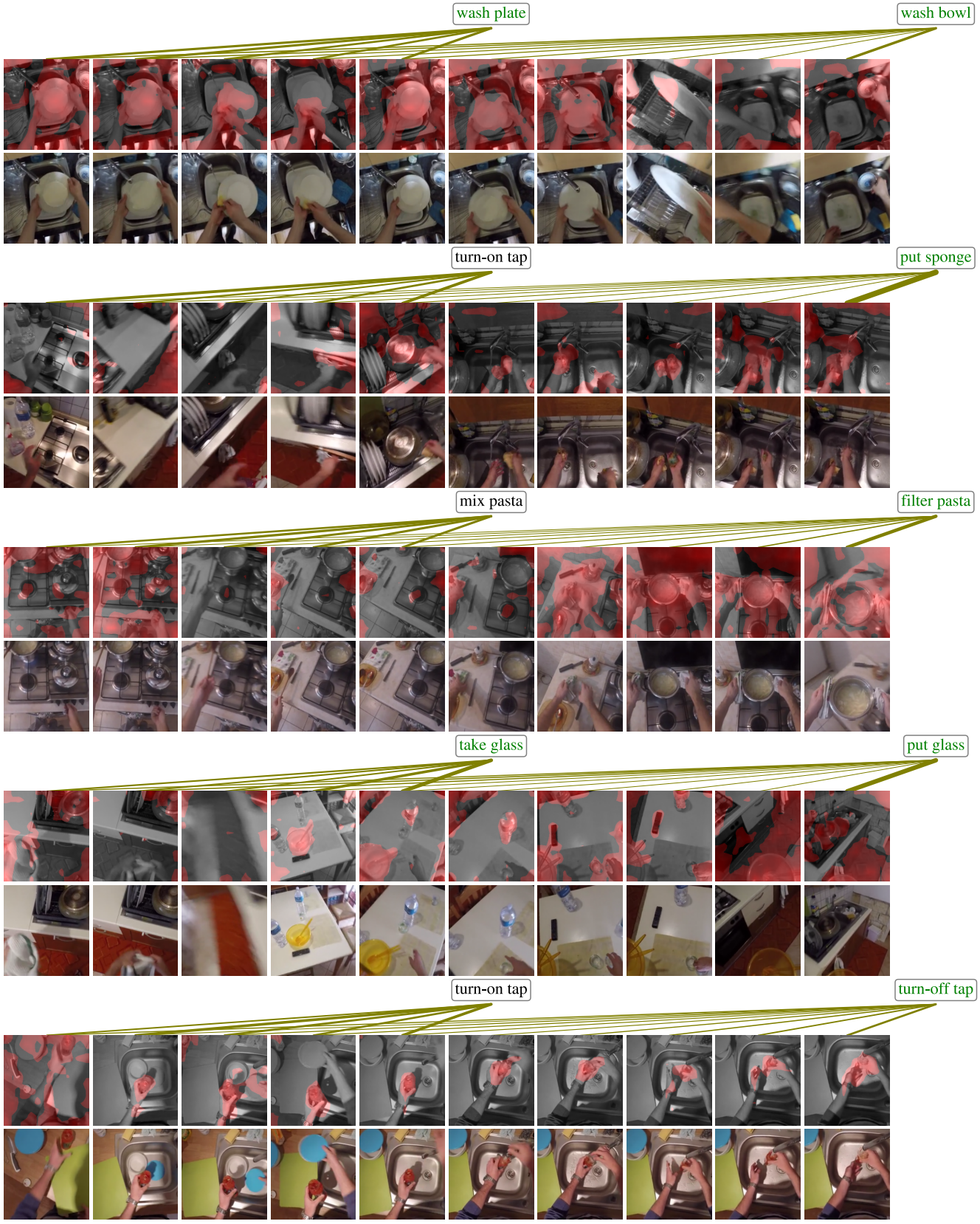

Figure 9: More Qualitative Results. (Continued) Here we also see some failure cases (the text in black—does not match the labeled ground truth). Note that the predictions in those failure cases are still reasonable. For instance in the second example the model predicts 'turn-on tap', while the groundtruth on that frame is 'wash cloth'. As we can see in the frame that the water is running, hence the 'turn-on tap' does happen before the eventual labeled action of 'wash cloth', albeit slightly sooner than when the model predicts.