CDAD: A Common Daily Action Dataset with Collected Hard Negative Samples - Supplementary Material -

In this supplementary file, we provide:

1. Experiment settings of feature analysis on Charades that are omitted in the main paper, including the category mapping between the original action categories and the target ones.
2. Detailed illustration of our annotation interface.
3. Bounding box type division and statistics of different actions.
4. Video samples of multiple actions, positive/negative pairs and different actions with the same background.

1. Feature analysis on Charades

In order to tell the differences of models learned on CDAD and previous action datasets such as Kinetics, we employ the models pre-trained on these datasets and perform feature analysis on a third-party dataset, *i.e.*, Charades in our case. We re-organize the categories of Charades and build a four-class subset for feature analysis. We choose "eat" and "drink" as our target actions as these actions appear in both Kinetics and CDAD. We select their similar-looking actions such as "putting some food somewhere" for "negative eat" and "pouring something into a cup/glass/bottle" for "negative drink". In Table 1, we show the category mapping between original categories and the target ones.

ID	category name	Original category
0	Eat	c065 Eating a sandwich c156 Someone is eating something
1	Neg. eat	c062 Putting some food somewhere c063 Taking food from somewhere c064 Throwing food somewhere c066 Making a sandwich c068 Putting a sandwich somewhere c069 Taking a sandwich from somewhere
2	Drink	c106 Drinking from a cup/glass/bottle
3	Neg. drink	c108 Pouring something into a cup/glass/bottle c109 Putting a cup/glass/bottle somewhere c110 Taking a cup/glass/bottle from somewhere c111 Washing a cup/glass/bottle

Table 1. Class selected in Charades.

2. Annotation Interface

We show the user interface of our annotation tool in Figure 1. For each video, the annotation process contains the following steps: 1) select start and end frames of an action instance; 2) generate keyframes; 3) select action label; 4) annotate action bounding boxes for each keyframe; 5) save action instance annotation; 6) repeat the above steps for other action instances; 7) submit the annotation file for this video.

Figure 1. User interface of our annotation tool.

3. Bounding box annotation statistics

In this section, we show the action bounding box statistics of CDAD. We divide the 23 action classes into four action types, namely "half-body", "full-body", "human-object" and "multi-person" actions. We use five different types of bounding boxes for these actions. Details are showed in the Table 2.

4. Video samples

In this section, we show snapshots of video samples of different types of actions. The original videos of these samples and more video samples are enclosed in the supplementary materials.

Samples containing multiple actions. Figure 2 shows video samples containing multiple actions in our proposed CDAD dataset. We see that multiple actions can happen simultaneously (first row), or in sequence (second and third row). This demonstrates the diversity and complexity of video clips in our dataset.

Samples of positive and negative pairs. Figure 3 shows video samples of positive and negative pairs that are collected within the same group in our proposed CDAD dataset.

Samples of different actions with the same background. Figure 4 shows more video samples of different actions with the same background. We can see that the same person was doing different actions (rows 1-3 and rows 4-5, respectively) with the same background. This disentangles subject from background in the action.

Type	Action	BBox	BBox num.
Half-body	Raise hand	UB	11,881
	Hold head		9,089
	Hit		6,219
	Clap		9,963
	Point		11,753
Full-body	Lie	FB	11,347
	Jump		7,067
	Climb		11,990
	Nap on a desk		8,468
	Kick		11,020
	Fall		2,698
Human-object	Call/Answer a phone	UOB	8,193
	Smoke		6,999
	Use a phone		16,897
	Drink		13,526
	Eat		5,994
	Wash		9,943
	Write		11,252
	Throw	FOB	7,508
	Sweep		13,900
Multi-person	Hug	MB	4,185
	Shake hands		9,408
	Fight		2,851

Table 2. Action type, name and bounding box types of CDAD. UB represents upper-body bounding box, FB represents full-body bounding box, UOB represents upper-body-object-bounding box, FOB represents full-body and object bounding box, and MB represents multi-person bounding box.

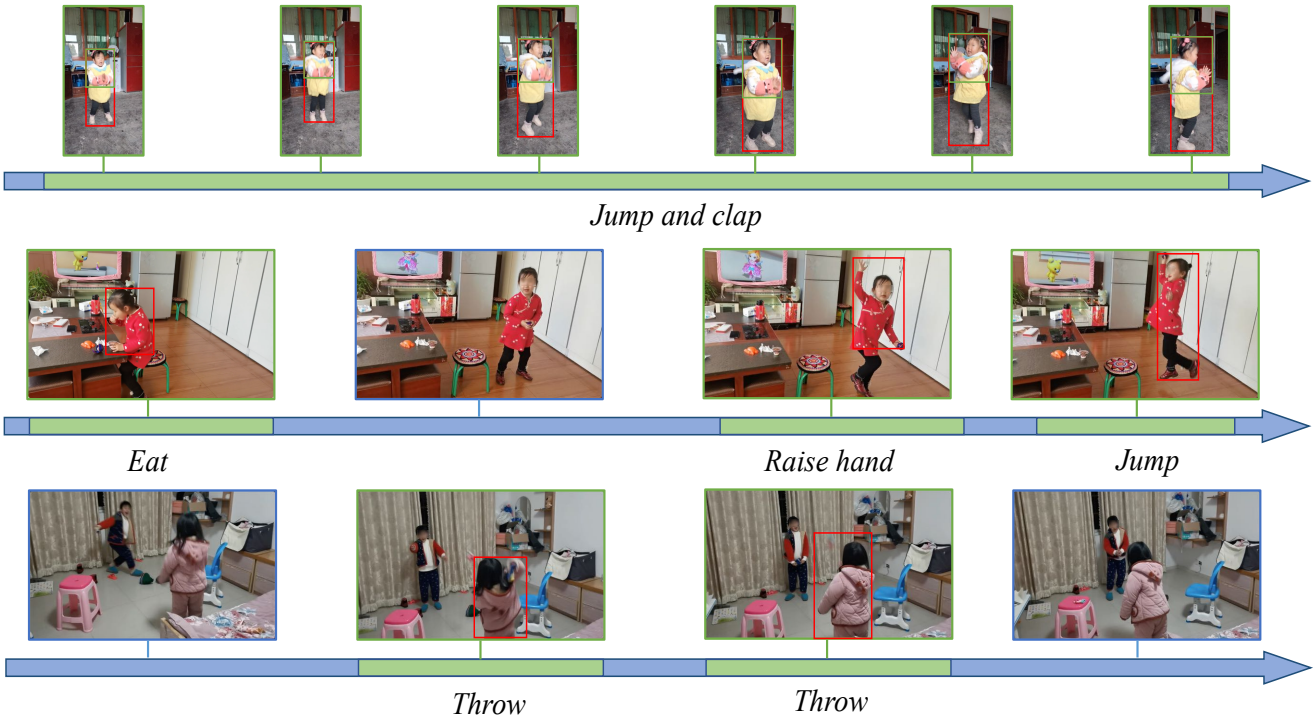

Figure 2. Sample video clips containing multiple actions.

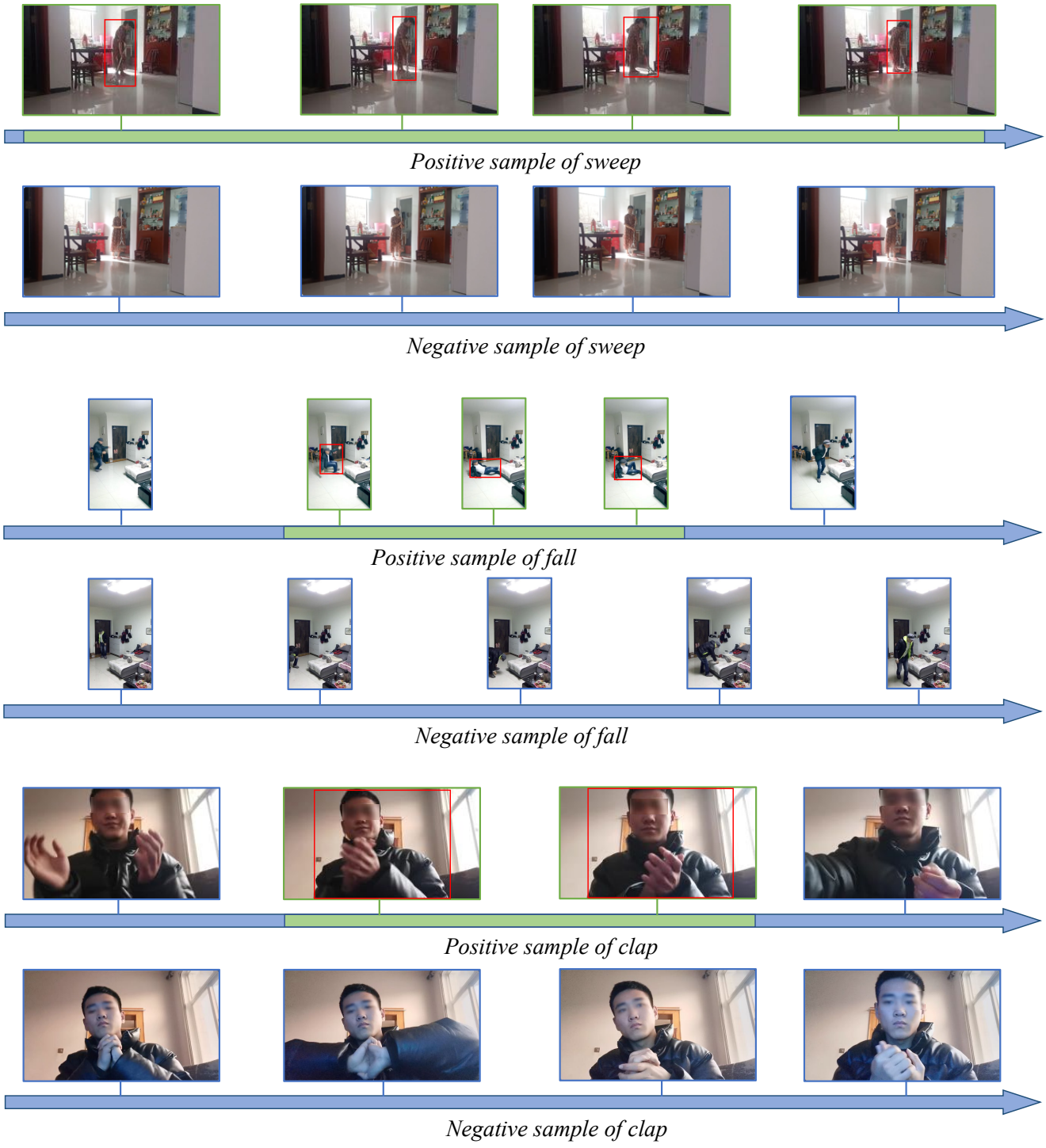

Figure 3. Sample annotated video clips of CDAD. Positive and negative samples are collected in pairs. We provide spatial-temporal annotations for all the positive video samples.

Figure 4. Sample annotated video clips of CDAD. We show different actions with the same person and background.