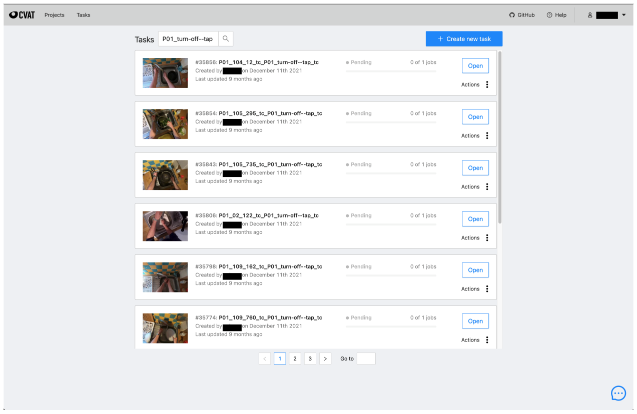
Supplementary Material

S1. Affordance annotation examples

We show some examples of the affordance annotation in Figure 5 where videos annotated with the same affordance label (*goal-irrelevant action, grasp type*) are listed in the same row. As we can see from the first and second rows, the annotation of goal-irrelevant actions such as "pull" is less ambiguous since we can easily determine it by the object's property (pullable) and the verb (open) performed in the video. A part of the verb / goal-irrelevant action mapping is shown in Table 1. However, the annotation of hand grasp types is more difficult due to the variation of the hand's appearance. During the annotation, we label the hand grasp type considering both the hand's appearance and the object's property to reduce ambiguity. For example, although the hands' appearance in the second and third example of the third row are different, they are annotated with the same label. After assigning the affordance labels and manually checking part of the automatically assigned labels, we get an accuracy of 88.32% and 98.96% on hand grasp types and goal-irrelevant actions. The complete annotation will be released once the paper is accepted.

S2. Data collection setup for affordance annotations

As mentioned in Section 3.2, we manually annotate affordance labels for videos of each action-participant pair. We randomly sample 5 video clips from each pair, then deploy the videos to the computer vision annotation tool (CVAT) [2] for labeling. The CVAT menu interface is shown in Figure 1. We can easily tell whether there is a scene change inside the videos of each pair from the gallery picture and annotate at least one video clip for each scene. During annotation, the annotator first watches the video and then labels the goal-irrelevant action and the type of hand grip. The annotation interface is shown in Figure 2.



#35856: P01_104_12_tc_P01_turn-off--tap_tc
Created by [redacted] on December 11th 2021
Last updated 9 months ago
Pending 0 of 1 jobs

#35854: P01_105_295_tc_P01_turn-off--tap_tc
Created by [redacted] on December 11th 2021
Last updated 9 months ago
Pending 0 of 1 jobs

#35843: P01_105_735_tc_P01_turn-off--tap_tc
Created by [redacted] on December 11th 2021
Last updated 9 months ago
Pending 0 of 1 jobs

#35806: P01_02_122_tc_P01_turn-off--tap_tc
Created by [redacted] on December 11th 2021
Last updated 9 months ago
Pending 0 of 1 jobs

#35798: P01_109_162_tc_P01_turn-off--tap_tc
Created by [redacted] on December 11th 2021
Last updated 9 months ago
Pending 0 of 1 jobs

#35774: P01_109_760_tc_P01_turn-off--tap_tc
Created by [redacted] on December 11th 2021
Last updated 9 months ago
Pending 0 of 1 jobs

Figure 1. Task menu of the annotation tool (CVAT).

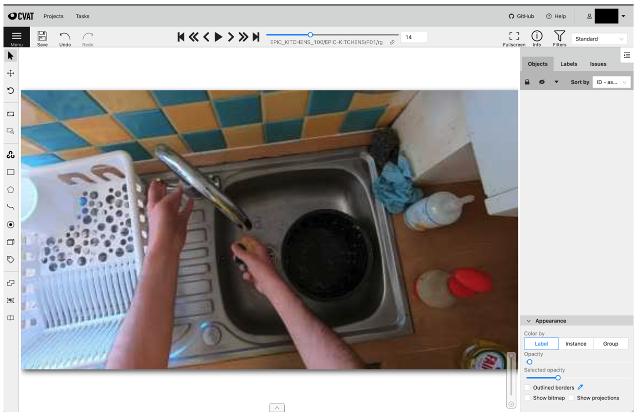


Figure 2. Annotation interface of the annotation tool (CVAT).

S3. Additional examples of interaction hotspot maps

Here we provide more results of the interaction hotspots prediction experiment in Sec 4.3, which are generated using action labels and affordance labels separately to compare their accuracy and granularity in Figure 6. The first group shows the hotspot maps generated by the "take" (red) action label (first row) and the hotspot maps generated by affordances related to "take": "pick-grasp1" (red), "pick-grasp3" (green), "pick-grasp4" (blue), "pick-grasp5" (cyan). The fine-grained affordance annotation helps the model distinguish diverse hand-object interactions on different object parts when performing the same action. In addition, the granularity of the affordance label helps the model better capture the possible interaction regions of the objects. For example, in the second row of the "put" action, the model captures more possible interaction regions on the board (second image from left) and the plate (third image) with affordance "place-grasp1" (green).

S4. Additional recognition results of verb / affordance / action

In this section, we compare the performance of verb/affordance/action recognition models trained on video clips within our affordance annotation. The distributions of verbs and actions are shown in Figure 4 and Figure 3. As shown in Table 2, although we introduce more affordance categories to represent various hand-object interactions, the performance does not drop much compared to the action recognition. The reason is that our affordance annotation focuses on the diverse hand-object interactions instead of the category of the objects, which benefits the model's recognition performance.

Verb	Goal-irrelevant action
close	push,push(rotate),press,rotate,place
open	pull,uplift,uncover,open1,rotate,pick,push
put-down	place
turn-off	press,rotate
turn-on	press,rotate
cut	move(press),scrape

Table 1. Part of verb / goal-irrelevant action mapping.

	Top1 Acc	Top5 Acc	mAP
Verb (21 classes)	0.6428	0.9414	0.5007
Affordance (60 classes)	0.5708	0.8771	0.4331
Action (91 classes)	0.4623	0.7244	0.3960

Table 2. Verb / affordance / action recognition results with Slowfast [1].

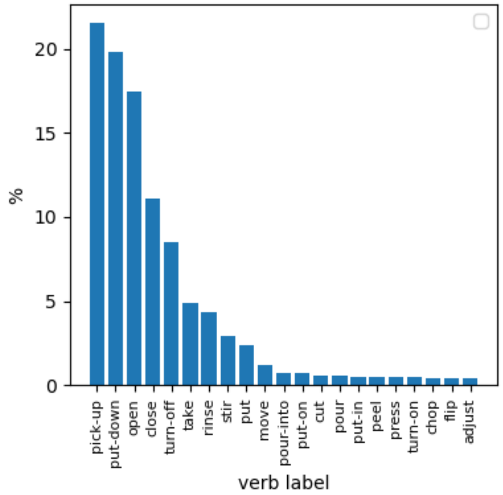


Figure 3. Distribution of verb classes.

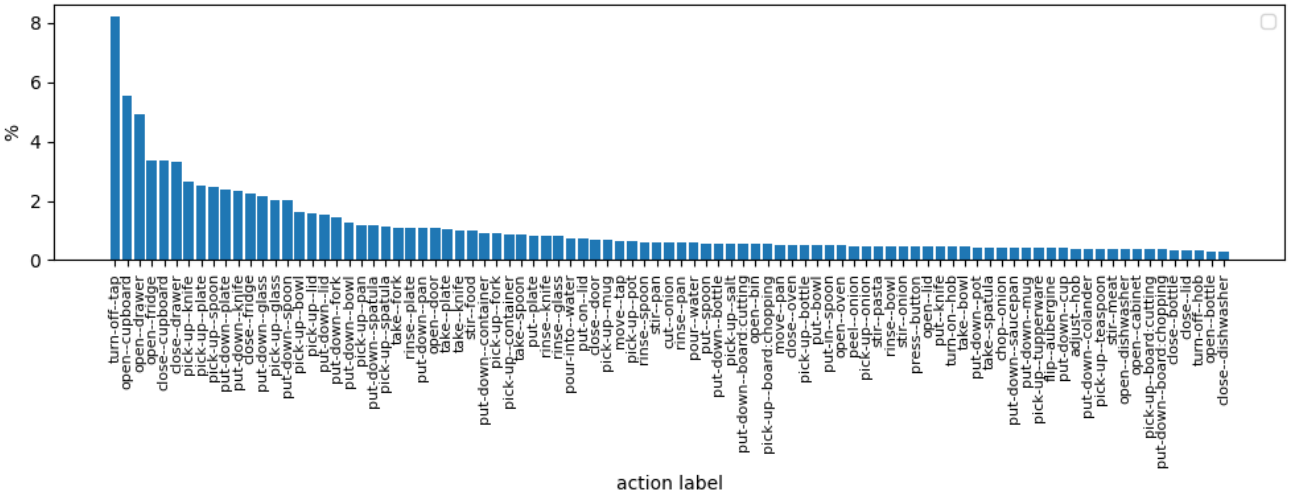


Figure 4. Distribution of action classes.

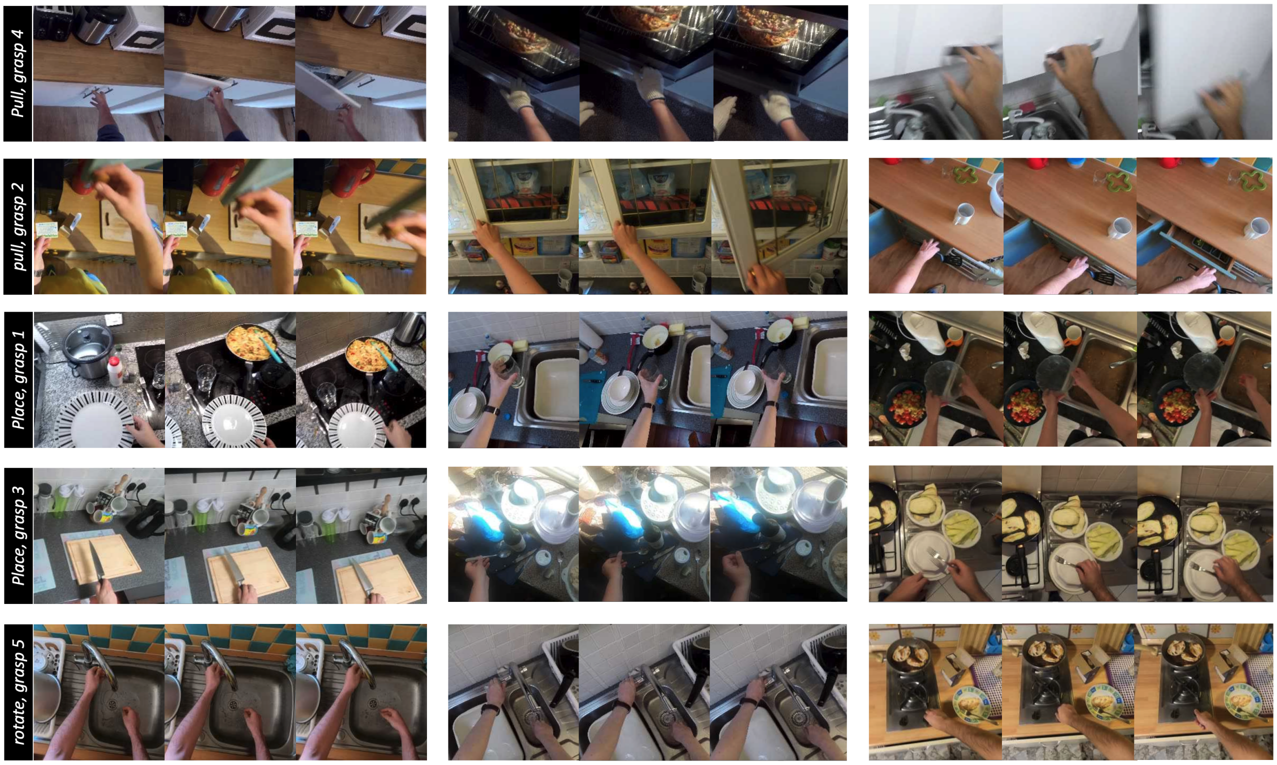


Figure 5. **Affordance annotation examples.** We show video clips contain the same affordance in each row.

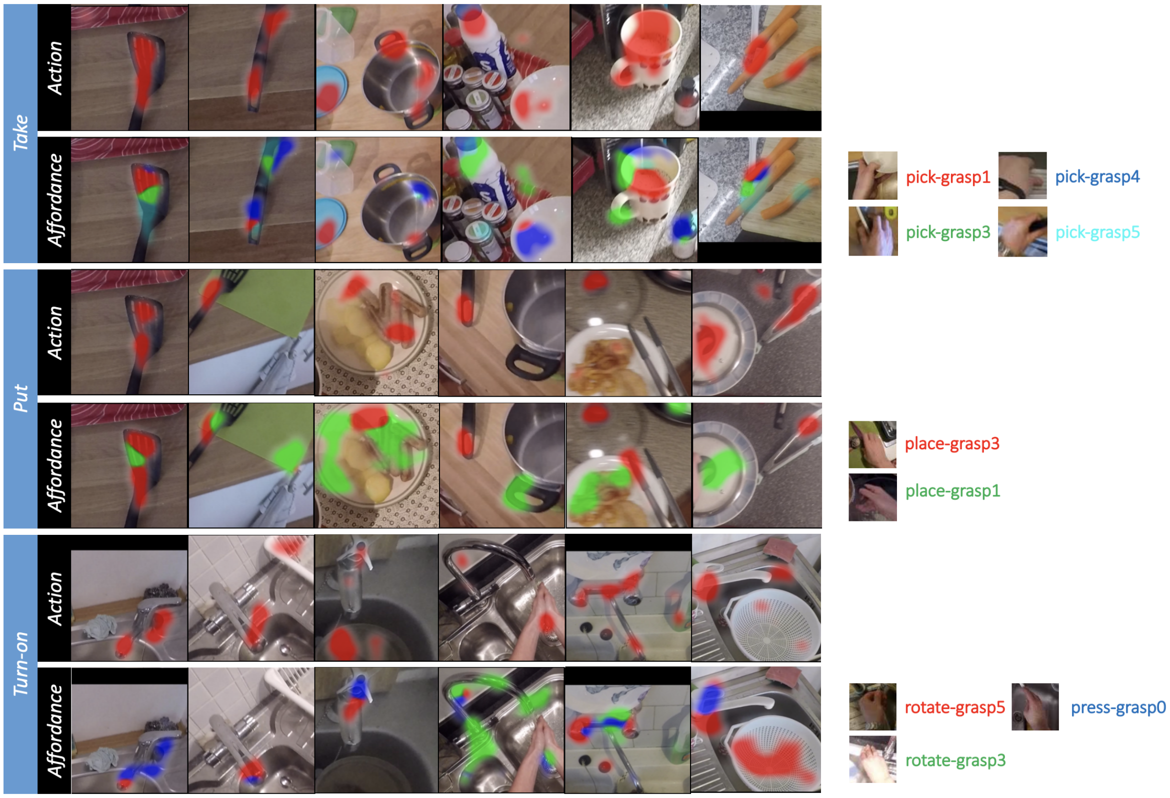


Figure 6. **Generated interaction hotspot maps on inactive object images.** These interaction hotspot maps show interaction regions of actions (take, put, turn-on) in the first row of each group and interaction regions of affordance related to each action (pick-grasp1, pick-grasp3, pick-grasp4, pick-grasp5, place-grasp3, place-grasp1, rotate-grasp5, rotate-grasp3, press-grasp0) in the second row.

References

[1] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019.

[2] Boris Sekachev, Nikita Manovich, Maxim Zhiltsov, Andrey Zhavoronkov, Dmitry Kalinin, Ben Hoff, TOsmanov, Dmitry Kruchinin, Artyom Zankevich, DmitriySidnev, Maksim Markelov, Johannes222, Mathis Chenuet, a andre, telenachos, Aleksandr Melnikov, Jijoong Kim, Liron Ilouz, Nikita Glazov, Priya4607, Rush Tehrani, Seungwon Jeong, Vladimir Skubriev, Sebastian Yonekura, vugia truong, zliang7, lizhming, and Tritin Truong. opencv/cvat: v1.1.0, Aug. 2020.