

Multi-Granularity Hand Action Detection

Supplementary Material

Anonymous Author(s)

A APPENDIX

A.1 FHA-Kitchens Dataset

A.1.1 Data Annotation.

We recruited 10 voluntary annotators to annotate hand actions for each frame with high quality. Their responsibility was to annotate bounding boxes and multi-granularity action categories for each hand interaction region. To enhance annotation efficiency, we implemented a parallel annotation approach. We utilized the LabelBee tool for annotating bounding boxes and coarse-grained categories, while fine-grained action triplets were annotated on the Amazon Mechanical Turk platform. To ensure annotation quality, we conducted three rounds of cross-checking and corrections. In the main paper, we described the annotation details for bounding boxes and categories. In addition to this, we also provide segment annotations for the objects.

Bounding Box Annotation. During annotation, we may encounter overlapping bounding boxes, *i.e.*, the same interacting object will satisfy two annotation definitions, for example, the *utility knife* in Figure 1, which is both the object directly touched by the right hand in the R-O interaction region and the active force provider in the O-O interaction region. In this case, we annotated all the labels because the same object participates in different interaction actions and has different roles (The corresponding visualizations are shown in Figure 1 and the annotation details are listed in Figure 2). Finally, we annotated a total of 198,839 bounding boxes over 9 types, including 49,746 hand boxes, 66,402 interaction region boxes, and 82,691 interaction object boxes. Compared to existing datasets [5], we added an average of 5 additional annotation types per frame.

Object Segment Annotation. To enrich our FHA-Kitchens, we utilized the state-of-the-art SAM model [9] to annotate object masks in all video frames, which can be used for action segmentation relevant tasks.

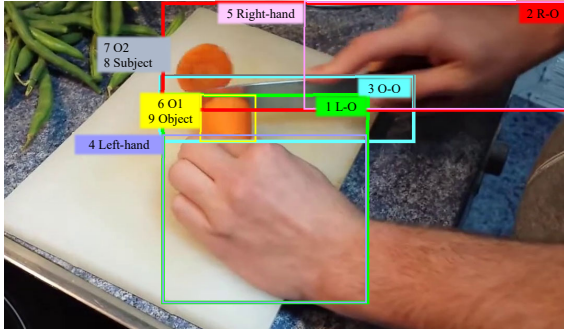


Figure 1: Visualization of bounding box annotations for the example of “fry vegetables”.

Bounding Box Annotation			Action Triplets Annotation
id	definition	b-box label	action triplet label
1	left hand-object interaction region	L-O	(hand_left, press-on, carrot_end)
2	right hand-object interaction region	R-O	(hand_right, hold-in, utility-knife_handle)
3	object-object interaction region	O-O	(utility-knife_body, cut-slice, carrot_head)
4	left hand	Left-hand	-
5	right hand	Right-hand	-
6	object touched by left hand in L-O	O1	-
7	object touched by right hand in R-O	O2	-
8	active force provider in O-O	Subject	-
9	passive force receiver in O-O	Object	-

Figure 2: Descriptive list of action triplets and bounding box annotations.

A.1.2 More statistics of the FHA-Kitchens Dataset.

In this part, we re-arrange some figures in the paper to make them more readable and provide more statistics of the FHA-Kitchens dataset. Our annotation primarily focuses on hand interaction regions, interaction objects, and their corresponding interaction actions, resulting in a diverse array of verbs, nouns, and bounding boxes.

Verbs. The annotated dataset comprises 130 action verbs that have been grouped into 43 parent verb categories (Figure 3 and Figure 4). The three most prevalent parent verb categories, based on the count of sub-action verbs, are *Cut*, *Hold*, and *Take*, representing the most frequently occurring hand actions in human interactions. Figure 4 visually depicts the distribution of all verb categories within FHA-Kitchens, ensuring the presence of at least one instance for each verb category. Specifically, the mapping between action verb IDs and their corresponding category names can be seen in Table 11.

Nouns. In our annotation process, we identified a total of 384 interaction object noun categories that are associated with actions, categorized into 17 super-categories. Figure 13 shows the distribution of noun categories based on their affiliations with super-categories. Notably, the super-category “vegetables & plants” exhibits the highest number of sub-categories, followed by “kitchenware”, which aligns with typical kitchen scenes. Specifically, the mapping between interaction object noun IDs and their corresponding category names can be seen in Table 12, Table 13, and Table 14.

Bounding Boxes. We performed a comprehensive statistical analysis on the bounding boxes of the three hand interaction regions and the corresponding interacting objects. Specifically, we focused on two aspects: the box area and the aspect ratio. Detailed results can be found in Figure 5 and Figure 6. Figure 5 shows the considerable range of sizes covered by our bounding boxes, with many interaction objects exhibiting small and challenging sizes for accurate detection. Moreover, in Figure 6, the aspect ratios of the bounding boxes exhibit notable variation. The aspect ratios of the

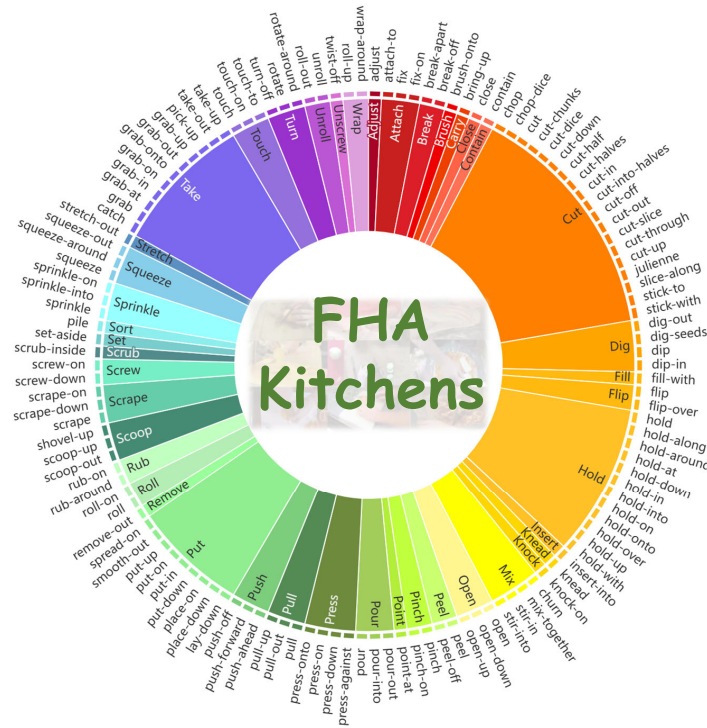


Figure 3: An overview of the action verbs and their parent action categories in FHA-Kitchens.

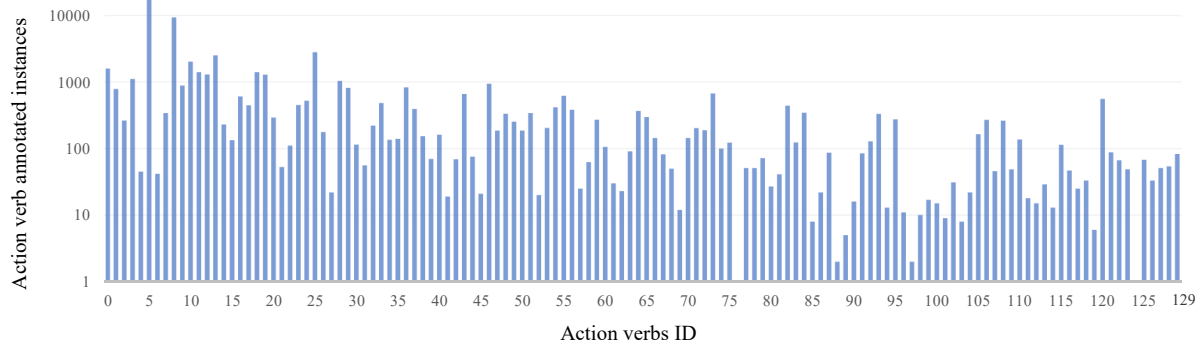


Figure 4: The distribution of instances per action verb category (the outer ring of the circle in Figure 3) in the FHA-Kitchens dataset.

three regions tend to concentrate within the range of $[0.5, 2]$, which can be attributed to the typical composition of interaction regions involving two interacting objects. Consequently, the bounding box encompasses the combined region of both objects. For instance, the R-O interaction region frequently involves the interaction between the “*right hand*” and “*utility knife*”. In such cases, the aspect ratio of the bounding box is observed to be 2:1, as depicted in Figure 1. These findings highlight the significant challenges of the detection task in our dataset.

Long-tail Property. The distribution of instances per action triplet category in FHA-Kitchens, as depicted in Figure 7, depicts a

long-tail property. This distribution reflects the frequency of hand interactions in real-world kitchen scenes, taking into account the varying commonness or rarity of specific hand actions. For instance, the action triplet “<*hand_right*, *hold-in*, *utility-knife_handles*>” consists of 9,887 instances, which is nine times more prevalent than the “<*hand_left*, *hold-in*, *utility-knife_handle*>” triplet. This long-tail characteristic of the distribution renders FHA-Kitchens a challenging benchmark for hand action recognition, making it suitable for investigating few-shot learning and out-of-distribution generalization in action recognition as well.

A.1.3 More quantitative and qualitative results.

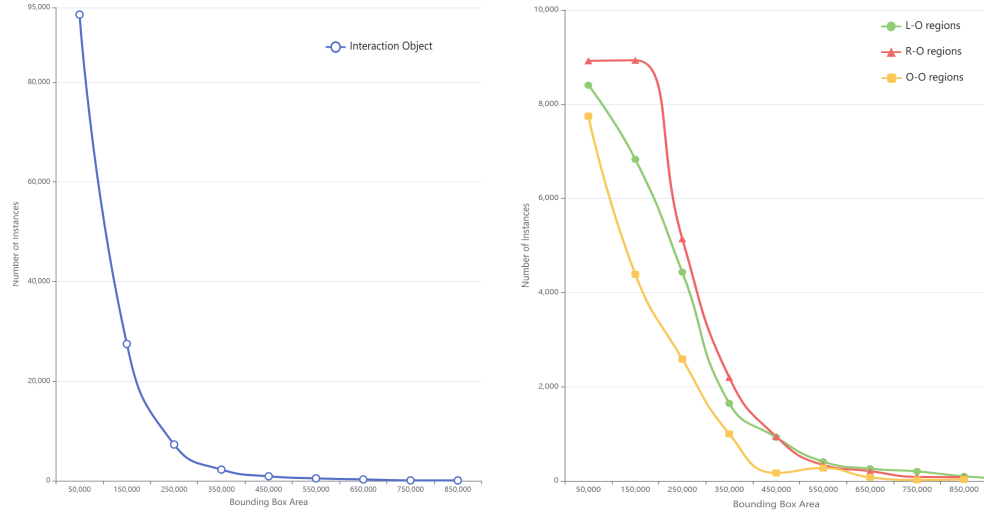


Figure 5: The distributions of bounding box areas of interaction objects (left) and interaction regions (right) in the FHA-Kitchens dataset.

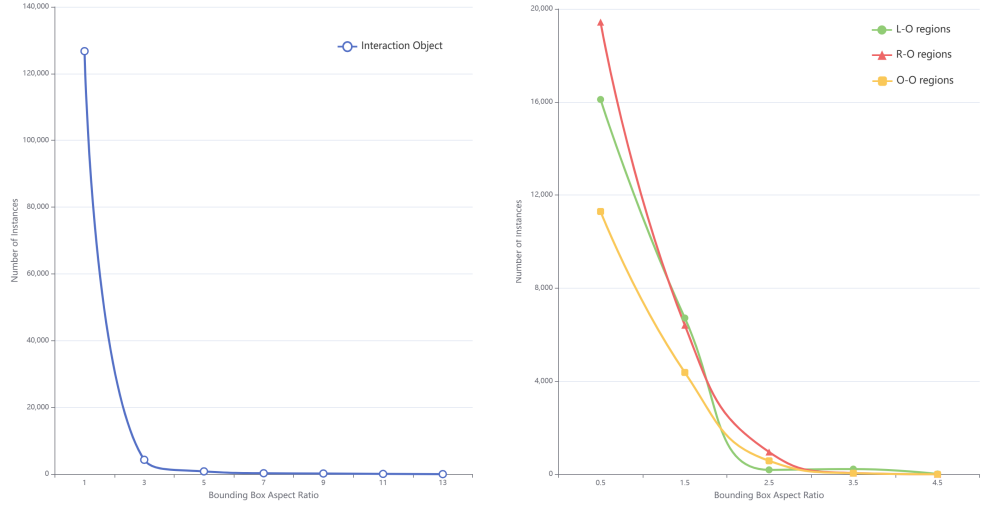


Figure 6: The distributions of bounding box aspect ratios of interaction objects (left) and interaction regions (right) in the FHA-Kitchens dataset.

1) Quantitative Results

SL-AR Track: Supervised Learning for Fine-grained Hand Action Recognition

Settings. The SL-AR track primarily evaluates the performance of different action recognition models on fine-grained hand actions. We adopted the representative TSN [19] and Slowfast [6] with the ResNet50 and ResNet101 backbones, VideoSwin [12] with the Swin-B backbone, VideoMAE V2 [18] with the three different size backbones, and Hiera [15] with the Hiera-B backbone. We trained these models on the FHA-Kitchens dataset using two settings: (1) **Pre-training on Kinetics 400 [3] and hybrid dataset**, where

we initialized the backbone with Kinetics 400 or Hybrid dataset pre-trained weights and fine-tuned the entire model on the FHA-Kitchens training set; and (2) **Training from scratch on FHA-Kitchens**, where we randomly initialized the model weights and directly train them on FHA-Kitchens. For different models, we used the recommended optimization strategy and batch size, and the maximum training period was set to 210 epochs.

Results on the SL-AR Track. Table 1 presents the performance of different action recognition methods on the Kinetics 400 [3] dataset and the proposed FHA-Kitchens dataset, with and without pre-trained models. From the experimental results, it can be

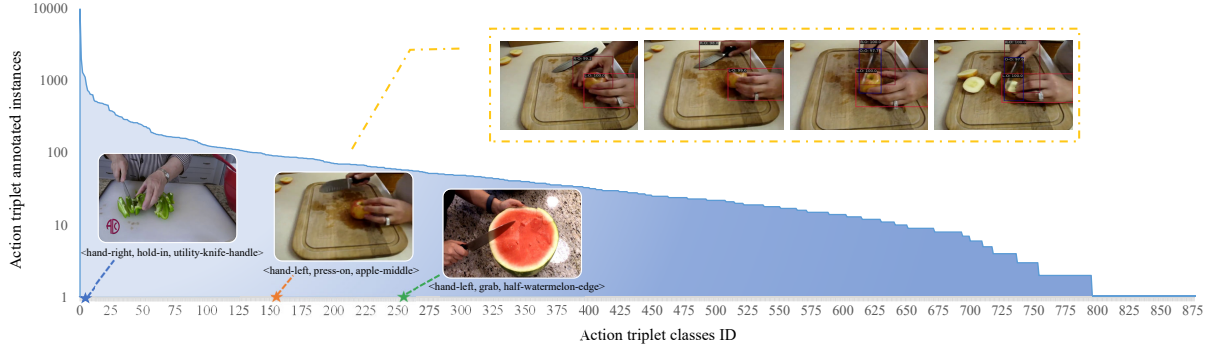


Figure 7: The distribution of instances per action triplet category in the FHA-Kitchens dataset.

Table 1: Classification results (Top-1 and Top-5 accuracy) of fine-grained hand actions using different methods on the validation set of the SL-AR track. w/ Pre-train : using pre-trained weights. w/o Pre-train: Training from scratch (the Kinetics 400 [3] dataset results from mmaction2 [4], VideoMAE V2 [18], and Hiera [15]).

Dataset	Method	Backbone	Pre-train Data	w/ Pre-train		w/o Pre-train	
				Top-1	Top-5	Top-1	Top-5
Kinetics 400	TSN [19]	ResNet50	ImageNet	72.83	90.65	-	-
		ResNet101	ImageNet	75.89	92.07	-	-
	SlowFast [6]	ResNet50	-	-	-	76.65	92.86
		ResNet101	-	-	-	78.65	93.88
	VideoSwin [12]	Swin-B	ImageNet	80.57	94.49	-	-
	VideoMAE V2 [18]	ViT-B	UnlabeledHybrid	81.50	-	-	-
		ViT-L	UnlabeledHybrid	85.40	-	-	-
		ViT-H	UnlabeledHybrid	86.90	-	-	-
	Hiera [15]	Hiera-B	Kinetics 400	84.00	-	-	-
Dataset	Method	Backbone	Pre-train Data	w/ Pre-train		w/o Pre-train	
				Top-1	Top-5	Top-1	Top-5
FHA-Kitchens	TSN [19]	ResNet50	Kinetics 400	30.37	74.26	29.11	73.84
		ResNet101	Kinetics 400	30.80	73.42	30.38	74.26
	SlowFast [6]	ResNet50	Kinetics 400	33.33	70.46	27.85	68.35
		ResNet101	Kinetics 400	36.71	67.93	31.22	69.62
	VideoSwin [12]	Swin-B	Kinetics 400	37.13	70.89	34.18	66.67
	VideoMAE V2 [18]	ViT-B	UnlabeledHybrid	21.67	57.08	-	-
		ViT-L	UnlabeledHybrid	32.92	68.75	-	-
		ViT-H	UnlabeledHybrid	34.58	68.33	-	-
	Hiera [15]	Hiera-B	Kinetics 400	27.00	69.20	-	-

observed that the performance trends of all action recognition methods on FHA-Kitchens are similar to their performance on Kinetics 400 [3], while the models perform much better on the coarse-grained actions of Kinetics 400. For the best-performing VideoSwin [12] model, the top-1 accuracy on Kinetics 400 surpasses the top-1 accuracy on FHA-Kitchens by 43.44%. And those methods with even large models cannot achieve satisfactory performance. This is clear evidence that validates the challenging nature of the fine-grained hand action recognition on FHA-Kitchens. Besides, the utilization of pre-trained weights has proven to be beneficial, resulting in improved accuracy compared to training models from

scratch. This finding suggests that despite the existence of a domain gap between coarse-grained and fine-grained actions, pre-training remains an effective strategy for addressing the challenges inherent in FHA-Kitchens, which have a larger number of action categories and relatively limited training data.

In addition, we further supplemented the hand pose information and conducted experiments using the skeleton-based STGCN [20] method. We used STGCN pre-trained on NTU60 [16] and NTU120 [11] and fine-tuned the models on the SL-AR track using different features for fine-grained hand actions, the results (Top-1 and Top-5 accuracy) can be seen in Table 2.

Table 2: Classification results (Top-1 and Top-5 accuracy) of fine-grained hand actions using different features and the skeleton-based STGCN [20] method (pre-trained on NTU60 [16] and NTU120 [11]) on the validation set of the SL-AR track.

Pre-train Data	Feature	Top-1	Top-5	Pre-train Data	Feature	Top-1	Top-5
NTU60	joint-2d	22.78	47.68	NTU120	joint-2d	20.68	48.10
	joint-3d	22.36	52.32		joint-3d	21.10	47.68
	joint-motion-2d	8.02	19.83		joint-motion-2d	9.28	20.25
	joint-motion-3d	10.97	23.63		joint-motion-3d	11.81	26.16
	bone-2d	22.36	49.79		bone-2d	24.05	57.81
	bone-3d	24.05	52.32		bone-3d	24.05	51.05
	bone-motion-2d	10.55	23.21		bone-motion-2d	9.28	23.21
	bone-motion-3d	13.50	26.16		bone-motion-3d	12.24	27.00

According to the experimental results, it can be observed that 3D pose features outperform 2D pose features and bone features achieve better results than joint features. Nevertheless, the overall results did not surpass the efficacy of hand-object interaction-based approaches, highlighting that relying only on hand pose information is insufficient for accomplishing fine-grained action recognition tasks. Because the generation of hand actions involves interacting objects, achieving a fine-grained hand action recognition task is required to consider the information of the objects interacting with the hand, which is different from a whole-body action recognition task (e.g., AVA, FineGym dataset).

DG Track: Intra- and Inter-class Domain Generalization for Interaction Region Detection

• Intra-class Domain Generalization

Settings. We conducted intra-class DG experiments using the three most prevalent parent action categories, *i.e.*, *Cut*, *Hold*, and *Take*. For each parent action category, we selected the most prevalent sub-categories and adopted the cross-validation protocol, *i.e.*, randomly choosing one sub-category as the test set while using all other sub-categories for training. Following the SL-AD track, we selected the Faster RCNN [14] model with the ResNet50 backbone as the default model, which is pre-trained on the MS COCO [10] object detection dataset.

Results on the Intra-class DG Track. The results on *Cut*, *Hold*, and *Take* are summarized in Table 3, 4, and 5. In the “*Cut*” parent category, the performance of all four detection models remains stable for the sub-categories seen during training but deteriorates for unseen sub-categories, as evidenced by the diagonal scores, which exhibit a minimum drop of 15 mAP. The findings in the results of the other two parent categories align with the observations in the “*Cut*” parent category. This finding suggests that there is still potential for enhancing the models’ generalization abilities, *e.g.*, by exploring the domain generalization or unsupervised domain adaptation techniques.

• Inter-class Domain Generalization

Settings. We chose the three most prevalent parent action categories *Cut*, *Hold*, and *Take*, and adopted the cross-validation protocol, *i.e.*, randomly choosing one parent category for training and using the other parent categories for testing. Other settings follow those in the intra-class DG track.

Results on the Inter-class DG Track. The results are listed in Table 6. Similar to the results in the intra-class DG track, the

Table 3: Intra-class DG test results of Faster RCNN [14] with the ResNet50 backbone on the “Cut” Setting. $\Delta_i = \frac{1}{3} \sum_{j,j \neq i} ji - ii$, $\Delta_i^* = \frac{1}{3} \sum_{j,j \neq i} ij - ii$, $i = 0, 1, 2, 3$.

Train	Test (mAP)				Δ^*
	cut-slice	cut-off	cut-down	cut-dice	
w/o cut-slice	33.30	65.00	56.00	60.90	27.33
w/o cut-off	57.10	48.00	54.80	62.80	10.23
w/o cut-down	57.30	64.40	41.30	63.50	20.43
w/o cut-dice	57.50	64.90	58.70	41.10	19.27
Δ	24.00	16.77	15.20	21.30	

Table 4: Intra-class DG test results of Faster RCNN [14] with the ResNet50 backbone on the “Hold” Setting. $\Delta_i = \frac{1}{2} \sum_{j,j \neq i} ji - ii$, $\Delta_i^* = \frac{1}{2} \sum_{j,j \neq i} ij - ii$, $i = 0, 1, 2$.

Train	Test (mAP)			Δ^*
	hold-up	hold-in	hold-around	
w/o hold-up	44.00	53.50	71.70	18.60
w/o hold-in	44.30	7.30	69.10	49.40
w/o hold-around	52.30	52.80	47.80	4.75
Δ	4.30	45.85	22.60	

Table 5: Intra-class DG test results of Faster RCNN [14] with the ResNet50 backbone on the “Take” Setting. $\Delta_i = \frac{1}{2} \sum_{j,j \neq i} ji - ii$, $\Delta_i^* = \frac{1}{2} \sum_{j,j \neq i} ij - ii$, $i = 0, 1, 2$.

Train	Test (mAP)			Δ^*
	pick-up	grab	catch	
w/o pick-up	0.40	47.10	46.60	46.45
w/o grab	19.00	4.50	39.00	24.50
w/o catch	19.10	46.00	15.60	16.95
Δ	18.65	42.05	27.20	

detection models perform well on the seen categories while deteriorating on the unseen categories. Nevertheless, it is interesting

to find that the performance gap ($\Delta_0 = 6.05$ and $\Delta_0^* = 8.05$) between *Cut* and others are smaller than those in the intra-class DG track, implying that there is likely a large intra-class variance, and the detection model is prone to overfitting the seen categories, particularly when the volume of training data is smaller (there are 7,463 training frames in *Hold* while only 1,680 in *Take*).

Table 6: Inter-class DG test results. $\Delta_i = ii - \frac{1}{2} \sum_{j,j \neq i} ji$, $\Delta_i^* = ii - \frac{1}{2} \sum_{j,j \neq i} ij$, $i = 0, 1, 2$.

Train	Test (mAP)			
	Cut	Hold	Take	Δ^*
Cut	37.40	29.50	29.20	8.05
Hold	48.70	52.30	41.80	7.05
Take	14.00	13.20	41.20	27.60
Δ	6.05	30.95	5.70	

2) Qualitative Results

The visual results of the SL-AD and SL-AR track experiments are presented in Figure 10, Figure 11, and Figure 12. We showcased the visualization results of interaction region detection, interacting object detection, and action recognition, focusing on hand interaction scenarios of varying complexity. In the interaction region detection results, we provide coarse-grained action categories corresponding to the sub-interaction regions, *i.e.*, $\langle L-O, R-O, O-O \rangle$. In the recognition results, we provide fine-grained action verbs corresponding to the three hand sub-interaction regions, denoted as $\langle L-O \text{ action verb}, R-O \text{ action verb}, O-O \text{ action verb} \rangle$. Figure 10 shows some challenging cases of hand interactions, providing compelling evidence of the good prediction performance of detection and recognition models, *i.e.*, the Faster-RCNN [14] with a ResNet50 backbone for detection and a pre-trained TSN [19] model with a ResNet50 backbone for action recognition. Moreover, Figure 11 and Figure 12 also demonstrate accurate detection and recognition results for some common interaction cases.

A.2 MG-HAD: Multi-Granularity Hand Action Detection

A.2.1 Multi-dimensional Action Queries.

To enhance the model's understanding of multi-dimensional action information from global and local perspectives, in our design of multi-dimensional action queries, we introduce an action dimensional hyper-parameter w_d ($d \in \{s, a, o\}$), to control the proportion of local information (sub-categories) fused with global information (triplet categories). In the three action dimensions $\langle s, a, o \rangle$, the a dimension is the most crucial for our task. Therefore, we use w_a as the key weight to dynamically adjust the weight proportions of the three action dimensions, with a total sum of 1. To determine the optimal weight distribution, we conducted a total of 10 comparative experiments under different backbones, with detailed results provided in Table 7. Based on the experimental results, the action dimensional weight setting we selected is as follows:

$$w_d = \begin{cases} 0.6 & \text{if } d = a \\ 0.2 & \text{if } d = s \text{ or } d = o \end{cases}, \quad (1)$$

Table 7: Comparative experiments of the action dimensional weight, *i.e.*, hyper-parameter w_d ($d \in \{s, a, o\}$), under different ratio settings. $w_s = w_o = (1 - w_a)/2$, $w_s + w_a + w_o = 1$, **M-G: Mixed-Grained.**

Method	Backbone	Train Data	w_a	M-G mAP(%)
Ours-4scale	ResNet-50		0.5	54.9
			0.6	57.0
			0.7	55.6
			0.8	56.0
			0.9	54.7
Ours-5scale	Swin-L	Multi-Granularity		
			0.5	57.2
			0.6	59.4
			0.7	58.2
			0.8	57.4
			0.9	57.1

where $w_s = w_o = (1 - w_a)/2$, $w_s + w_a + w_o = 1$.

A.2.2 Coarse-Fine Contrastive DeNoising (C-F CDN).

To enable the model to handle multi-granularity hand action labels and understand the differences between different granularity labels, we propose the C-F CDN module. When designing the noise label generation strategy, we replaced the “random” generation of noise with “specified”, reducing randomness by specifying noise positions and categories. This ensures noise is added to different granularity labels, generating CDN queries encompassing various granularity. To determine the optimal setting, we considered the noise distribution of different granularity categories in real-world scenarios and ensured contrastive learning between coarse- and fine-grained information. We conducted three sets of comparative experiments (see Table 8). The final selected setting, as shown in Eq. (2), exhibits the most significant improvement. Hence, subsequent experiments were conducted using this setting for further investigation. Our method's success lies in its ability to suppress confusion at the category level and select appropriate granularity to predict hand action categories, thus enhancing its ability to predict multi-granularity information.

$$\text{noise label} = \begin{cases} \text{fine-grained} & i \in [0, 3) \\ \text{mixed-grained} & i \in [3, C) \end{cases}, \quad (2)$$

where i indexes the multi-granularity action category for a specific instance (*i.e.*, 0~2 denote coarse-grained categories while 3~C-1 denote fine-grained categories), and C is the number of categories. “fine-grained” and “mixed-grained” denote that the noise label is chosen randomly from the fine-grained categories and the combination of the coarse-grained and fine-grained categories, respectively.

A.2.3 Action Detection Results of MG-HAD.

Visualization detection results. Based on the 5scale-Swin-L backbone, qualitative comparison results with the baseline [21] on the FHA-Kitchens dataset are shown in Figure 8. Our model accurately detects three hand sub-interaction regions (*i.e.*, “Left hand-Object interaction region (L-O)”, “Right hand-Object interaction region (R-O)”, and “Object-Object interaction region (O-O)”).

Table 8: Three sets comparative experiments of the “noise label generation strategy” in the Coarse-Fine Contrastive DeNoising (C-F CDN). M-G: Mixed-Grained, C-G: Coarse-Grained, F-G: Fine-Grained.

Method	Backbone	Train Data	noisy label location	noisy label class	FHA-Kitchens val mAP(%)		
					M-G label	C-G sub-label	F-G sub-label
<i>Baseline</i> [21]	ResNet-50	multi-granularity	random	random	54.7	76.3	54.5
<i>Ours-4scale</i>			a_1 : fine-grained a_2 : coarse-grained	$a_1 \rightarrow$ mixed-grained $a_2 \rightarrow$ fine-grained	56.6	75.9	56.4
				$a_1 \rightarrow$ coarse-grained $a_2 \rightarrow$ fine-grained	56.1	75.5	55.8
				$a_1 \rightarrow$ coarse-grained $a_2 \rightarrow$ mixed-grained	55.8	74.5	55.6

and provides multi-granularity hand action categories (*i.e.*, “Coarse-Grained” and “Fine-Grained”). Compared to the DINO, we demonstrate superior performance across multiple dimensions of fine-grained categories, illustrating the effectiveness of our designed multi-dimensional action queries. Additionally, we present our model’s multi-granularity hand action detection results in more kitchen scenarios, as shown in Figure 9. We randomly selected four different kitchen scenarios, *i.e.*, “*fry vegetables*”, “*sandwich*”, “*salad*”, and “*fruit*”, showcasing complex hand actions. Our model offers accurate bounding boxes and multi-granularity hand action information for three hand sub-interaction regions.

Basic Hyper-parameters. For the basic hyper-parameters, consistent with DINO [21], we utilized a 6-layer Transformer encoder and a 6-layer Transformer decoder with a hidden feature dimension of 256. We set the initial learning rate (lr) to 1×10^{-4} and employed a MultiStep lr scheduler, dropping lr by multiplying 0.1 at the 11-th and 20-th epochs for ResNet50, corresponding to the 12 and 24 epoch settings. We employed the AdamW [8, 13] optimizer with a weight decay of 1×10^{-4} and trained our model on NVIDIA GeForce RTX 3090 GPUs. The ResNet50 backbone was trained with a batch size of 2 per GPU, while the SwinL backbone had a batch size of 1. We initialized 900 decoder queries, maintaining the same computational cost as DINO. We provide detailed hyper-parameters in Table 10 for reproducibility.

A.3 Datasheets for Datasets

A.3.1 Motivation.

1. For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

A1: FHA-Kitchens is created to facilitate research in the field of complex multi-granularity hand action. It is important to study several challenging questions in the context of more training data from diverse multi-granularity hand actions, such as: (1) How do different representative action recognition models perform on fine-grained hand action tasks? (2) How do state-of-the-art detection models perform on the refined hand interaction regions with multi-granularity hand action categories? (3) How about the impact of pre-training, *e.g.*, on the whole-body actions dataset [3], in the context of the large-scale dataset with diverse multi-granularity hand

actions? and (4) How do the intra-class and inter-class generalization capabilities of models trained with specific fine-grained hand actions or parent hand actions perform? However, existing action datasets primarily focus on whole-body actions or coarse-grained action categories, lacking finer-grained hand-action localization and category information. Therefore, it is impossible to study these questions using existing datasets. In contrast, FHA-Kitchens primarily focuses on hand actions and refines hand interaction regions into three sub-interaction regions. We annotated coarse- and fine-grained actions for each sub-interaction region. Coarse-grained categories, denoted by the generic terms “L-O”, “R-O”, and “O-O”, represent the coarse actions within the sub-interaction regions. Fine-grained action category in a triplet format: $\langle \text{subject}, \text{action verb}, \text{object} \rangle$. Overall, we meticulously annotated 880 hand action categories (coarse- and fine-grained) for approximately 220k bounding boxes, with each category corresponding to a sub-interaction region’s localization box. Fine-grained categories per frame have nine dimensions, resulting in 877 action triplets, significantly enhancing the granularity of actions and providing valuable resources for researchers to study these questions effectively.

FHA-Kitchens aims to provide a better, more comprehensive, and finer-grained benchmark for hand action. However, existing hand-action datasets exhibit limitations including insufficient representation of hand-action granularity, lack of annotation of hand-action interaction regions, and neglect of the relationships between interacting objects. With its diverse and finer-grained hand action information, the FHA-Kitchens dataset enables a better evaluation performance for hand action tasks.

2. Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

A2: Our dataset is created by the authors as well as some volunteer undergraduate students.

3. Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

A3: This information will be made public once the paper is accepted after peer review.

A.3.2 Composition.

1. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people

and interactions between them; nodes and edges)? Please provide a description.

A1: FHA-Kitchens consists of video clips, each video clip consists of consecutive video frames, including 880 coarse- and fine-grained hand action categories. For each frame, we provide bounding boxes for three hand sub-interaction regions (*i.e.*, left hand-object (L-O), right hand-object (R-O), and object-object (O-O) interaction regions) and the interaction objects. Each sub-interaction region action was annotated using coarse- and fine-grained action categories. Coarse-grained categories, denoted by the generic terms “L-O”, “R-O”, and “O-O”, fine-grained action category in a triplet format: *<subject, action verb, object>*. Additionally, we provide segmentation masks related to hands and interaction objects.

2. How many instances are there in total (of each type, if appropriate)?

A2: The FHA-Kitchens contains 30,047 frames from 2,377 video clips, with each frame annotated for three hand sub-interaction regions, resulting in a total of 877 fine-grained action triplets and 3 coarse-grained action categories. Among them, there are 597 frames where no hand interaction action occurs, represented as L-O_triplet:<none>, R-O_triplet:<none>, O-O_triplet:<none>.

3. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

A3: FHA-Kitchens is a real-world sample of human hands part in the kitchen scenes, including information about their hand actions. The data is sourced from an existing large-scale whole-body action dataset [17], from which we selected videos featuring hand interaction actions. We extracted a total of 2,377 video clips, amounting to 84.22 minutes of footage, encompassing 8 distinct types of dishes. Due to the diversity of real-world human hand actions, it’s impossible to cover all types of actions. The FHA-Kitchens dataset focuses primarily on multi-granularity hand action tasks. To address the granularity issue, we improved the hand action information in the existing dataset. Compared to the data’s original annotations in Kinetics-700_2020 [17], our dataset expanded the action labels by 7 dimensions, increased the number of action categories by 52 times, and introduced 122 new action verbs. We provide a finer-grained set of hand-action instances than ever before, facilitating further research in hand-action.

4. What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.

A4: Each video frame consists of at most 9 types of bonding boxes (*i.e.*, three hand sub-interaction regions and interaction objects within interaction region) and sub-interaction region corresponding coarse- and fine-grained descriptions (*i.e.*, L-O, R-O, O-O, and *<subject, action verb, object>*). Additionally, we took into account the “active-passive” relationships between object pairs and the specific contact areas involved in the interaction actions. Consequently, our annotation process encompassed a total of nine dimensions, resulting in a total of 877 fine-grained hand action triplets and 3

coarse-grained hand action categories. The annotated visualizations are shown in Figure 1 and corresponding details are listed in Figure 2.

5. Is there a label or target associated with each instance? If so, please provide a description.

A5: Yes. Due to our parallel annotation process, we generated annotation files in different styles. However, we consolidated all the bounding box and triplet annotation information into a single CSV file. In the merged CSV file, each instance is annotated with labels following the style of the Kinetics [1–3, 17] and AVA [7] datasets, which include video_name, video_id, clip_id, clip_name, frame_name, timestamp, L-O_triplet, L-O_action_verb_id, L-O_action_verb_class, L-O_action_bbox, left_hand_bbox, O1_class, O1_bbox, R-O_triplet, R-O_action_verb_id, R-O_action_verb_class, R-O_action_bbox, right_hand_bbox, O2_class, O2_bbox, O-O_triplet, O-O_action_verb_id, O-O_action_verb_class, O-O_action_bbox, subject_class, subject_bbox, object_class, object_bbox, action_verb_triplet, action_verb_triplet_id.

6. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information but might include, e.g., redacted text.

A6: Yes. Some instances may not have all 9 types of bonding boxes and their corresponding coarse-fine-grained action categories and segmentation annotation because of interaction action scenes, severe occlusion, truncation, blur, or small scale. We just annotated “None” in our annotation file to represent this situation.

7. Are relationships between individual instances made explicit (e.g., users’ movie ratings, and social network links)? If so, please describe how these relationships are made explicit.

A7: Yes. We provide different styles of annotation files, in COCO-style, the annotations are connected by image id and category id, you can easily access them by COCO APIs. In CSV style, one line represents the annotations of one frame and can be processed by the pandas library easily.

8. Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

A8: Yes. We randomly split the dataset into the disjoint train, validation, and test sets following the ratio of 7:1:2.

9. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

A9: Although we conducted three rounds of cross-checking and corrections, there may still be some errors in the annotations, *e.g.*, inappropriate bounding box annotations, or small drifts of the bounding box locations, incorrectly written verbs or nouns, insufficient granularity in verb or noun descriptions, inappropriate formatting of triplets, etc. However, we have made every effort to minimize such occurrences.

To analyze the quality of annotations, we randomly selected 500 frames and conducted manual evaluations for correctness. The results are reported in Table 9. These error rates are comparable to recently published datasets [5].

10. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are

Table 9: Error rate in FHA-Kitchens. I-O: Interaction Objects, I-R: Interaction Regions.

	Frames	I-O Boxes	I-R Boxes	Verb	Noun
Total Number	500	3,006	1,503	1,503	2,006
Error Rate (%)	-	4.9	2.5	2.2	5.3

there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.g., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

A10: Our dataset was derived from a large-scale publicly available dataset, namely Kinetics-700_2020 [17], which is publicly available for download from their website. The Kinetics dataset follows the Creative Commons Attribution 4.0 International License. We would like to express our gratitude to the authors for their significant contributions to the research community.

11. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

A11: No.

12. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

A12: No.

A.3.3 Collection Process.

1. How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

A1: Our data was obtained from the existing large-scale publicly available dataset, namely Kinetics-700_2020 [17], which is publicly available for download from their website, and then further cleaned, frame segmented, and reorganized to obtain 2377 video clips.

2. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

A2: The data in FHA-Kitchens come from dataset publicly available datasets described above, which can be directly downloaded from their websites.

3. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

A3: Currently, we focus exclusively on hand interaction actions in kitchen scenes, thus primarily extracting data that includes hand interaction actions in kitchen scenes.

4. Who was involved in the data collection process (e.g., students, crowdworkers, contractors), and how were they compensated (e.g., how much were crowdworkers paid)?

A4: The authors collected this dataset. The annotation compensation is based on the prevailing market rates.

5. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., the recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

A5: It took about 1 week to collect the data and about 6 weeks to complete organization and annotation, as each participant labeled the bonding boxes and action triplets about four hours per work-day. And the segmentation masks are generated by the Segment-Anything Model [9] guided by the bonding boxes, and corrected by human annotators for about one week.

A.3.4 Preprocessing/cleaning/labeling.

1. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

A1: Yes. Since we focus on hand actions, we performed filtering and processing operations on the original videos, including the following three steps. (1) First, we observed that kitchen scenes often featured hand actions, with video content prominently showcasing human hand parts. Therefore, we sought out and extracted relevant videos that were set against a kitchen backdrop. (2) Then, to ensure the quality of the dataset, we selectively chose videos with higher resolutions. Specifically, 87% of the videos were recorded at 1,280 × 720 resolution, while another 13% had a shorter side of 480. Additionally, 67% of the videos were captured at 30 frames per second (fps), and another 33% were recorded at 24~25 fps. (3) Subsequently, we imposed a duration constraint on the videos, ranging from 30 seconds to 5 minutes, to exclude excessively long-duration videos. This constraint aimed to maintain a balanced distribution within the sample space. Finally, we collected a total of 2,377 video clips, amounting to 84.22 minutes of footage, encompassing 8 distinct types of dishes.

The collected video data was reorganized and cleaned to align with our annotation criteria. First, we split the collected video data into individual frames, as our annotated units are frames. Subsequently, we conducted further cleaning of the frames by excluding those that did not depict hands or exhibited meaningless hand actions. This cleaning process took into consideration factors such as occlusion, frame quality (i.e., without significant blur, subtitles, and logos), meaningful hand actions, and frame continuity. As a result, we obtained a total of 30,047 high-quality candidate video frames containing diverse hand actions for our FHA-Kitchens dataset. Compared to the initial collection, 113,436 frames were discarded during the cleaning process.

We recruited 10 voluntary annotators, whose responsibility was to annotate bounding boxes and multi-granularity action categories for each hand interaction region. To enhance annotation efficiency, we implemented a parallel annotation pipeline. The annotation of fine-grained action triplets was carried out on the Amazon Mechanical Turk platform, while the bounding box and coarse-grained actions annotation was facilitated using the LabelBee tool. To ensure the annotation quality, three rounds of cross-checking and corrections were conducted.

2. Was the “raw” data saved in addition to the preprocessed/cleaned/ labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.

A2: No.

3. Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

A3: The annotation of fine-grained action triplets was carried out on the Amazon Mechanical Turk platform, while the bounding box and coarse-grained actions annotation was facilitated using the LabelBee tool.

A.3.5 Uses.

1. Has the dataset been used for any tasks already? If so, please provide a description.

A1: No.

2. Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

A2: N/A.

3. What (other) tasks could the dataset be used for?

A3: FHA-Kitchens can be used for the research of fine-grained hand action recognition, multi-granularity hand action detection, and interaction object detection. Besides, it can also be used for specific machine learning topics such as domain generalization and action segmentation. Please see the Discussion part of the main paper and supplementary material.

4. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

A4: No.

5. Are there tasks for which the dataset should not be used? If so, please provide a description.

A5: No.

A.3.6 Distribution.

1. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

A1: Yes. The dataset will be made publicly available to the research community.

2. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

A2: It will be publicly available on the dataset project website at GitHub.

3. When will the dataset be distributed?

A3: The dataset will be distributed once the paper is accepted after peer review.

4. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

A4: It will be distributed under the MIT license.

5. Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

A5: No.

6. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

A6: No.

A.3.7 Maintenance.

1. Who will be supporting/hosting/ maintaining the dataset?

A1: The authors.

2. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

A2: They can be contacted via email available on our dataset project website.

3. Is there an erratum? If so, please provide a link or other access point.

A3: No.

4. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

A4: No. We have carefully three rounds of cross-checking the annotations to reduce the labeling errors. There may be very few labeling errors, which can be treated as noise.

5. Will older versions of the dataset continue to be supported/ hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

A5: N/A.

6. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

A6: N/A.

REFERENCES

- [1] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. 2018. A short note about kinetics-600. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- [2] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2019. A short note on the kinetics-700 human action dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- [3] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.
- [4] MMAAction2 Contributors. 2020. OpenMMLab’s Next Generation Video Understanding Toolbox and Benchmark. <https://github.com/open-mmlab/mmaaction2>.
- [5] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision*. 720–736.
- [6] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slow-fast networks for video recognition. In *Proceedings of the IEEE International Conference on Computer Vision*. 6202–6211.
- [7] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. 2018. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (challenge)*. 6047–6056.
- [8] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.
- [9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643* (2023).
- [10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V* 13. Springer, 740–755.
- [11] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. 2019. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 10 (2019), 2684–2701.
- [12] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3202–3211.
- [13] Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations*.
- [14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems* 28 (2015).
- [15] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. 2023. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *Proceedings of the International Conference on Machine Learning*. PMLR, 29441–29454.
- [16] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1010–1019.
- [17] Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. 2020. A short note on the kinetics-700-2020 human action dataset. *arXiv preprint arXiv:2010.10864* (2020).
- [18] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. 2023. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 14549–14560.
- [19] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *Proceedings of the European Conference on Computer Vision*. Springer, 20–36.
- [20] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the Association for the Advancement of Artificial Intelligence Conference*, Vol. 32.
- [21] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. 2023. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *Proceedings of the International Conference on Learning Representations*.

Table 10: Hyper-parameters used in MG-HAD.

Item	Value
lr	0.0001
lr_backbone	1e-05
weight_decay	0.0001
clip_max_norm	0.1
pe_temperature	20
enc_layers	6
dec_layers	6
dim_feedforward	2048
hidden_dim	256
dropout	0.0
nheads	8
num_queries	900
enc_n_points	4
dec_n_points	4
transformer_activation	“relu”
set_cost_class	2.0
set_cost_bbox	5.0
set_cost_giou	2.0
cls_loss_coef	1.0
bbbox_loss_coef	5.0
giou_loss_coef	2.0
focal_alpha	0.25
dn_box_noise_scale (γ_1)	1.0
dn_label_noise_scale (γ_2)	0.5
subject_weight (w_s)	0.2
action_weight (w_a)	0.6
object_weight (w_o)	0.2

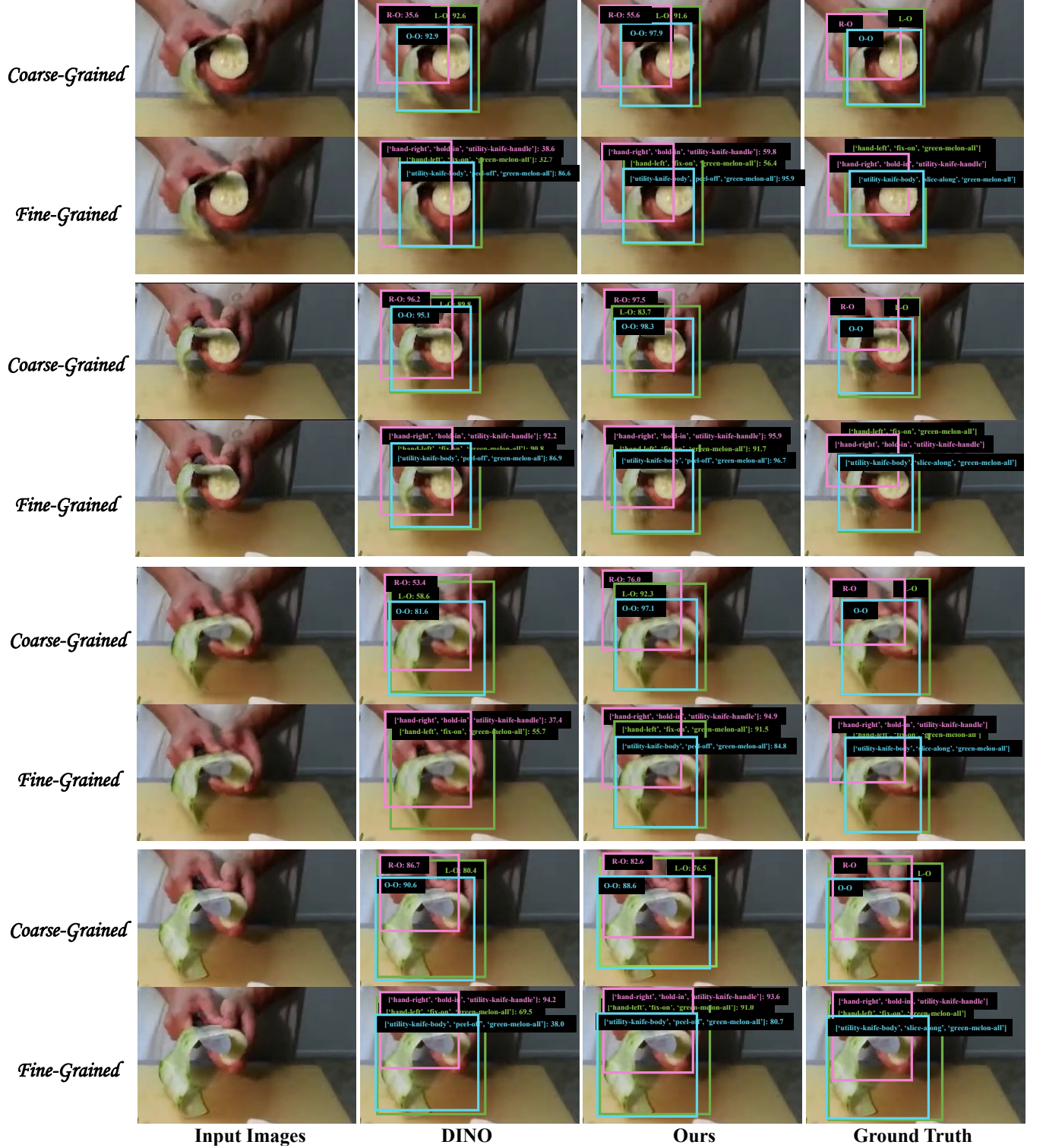


Figure 8: Qualitative comparison on the FHA-Kitchens dataset. Our model accurately detects three hand sub-interaction regions and provides multi-granularity hand action categories. Compared to the baseline [21], our model performs better across multi-dimensional fine-grained categories, demonstrating the effectiveness of our designed multi-dimensional action queries.

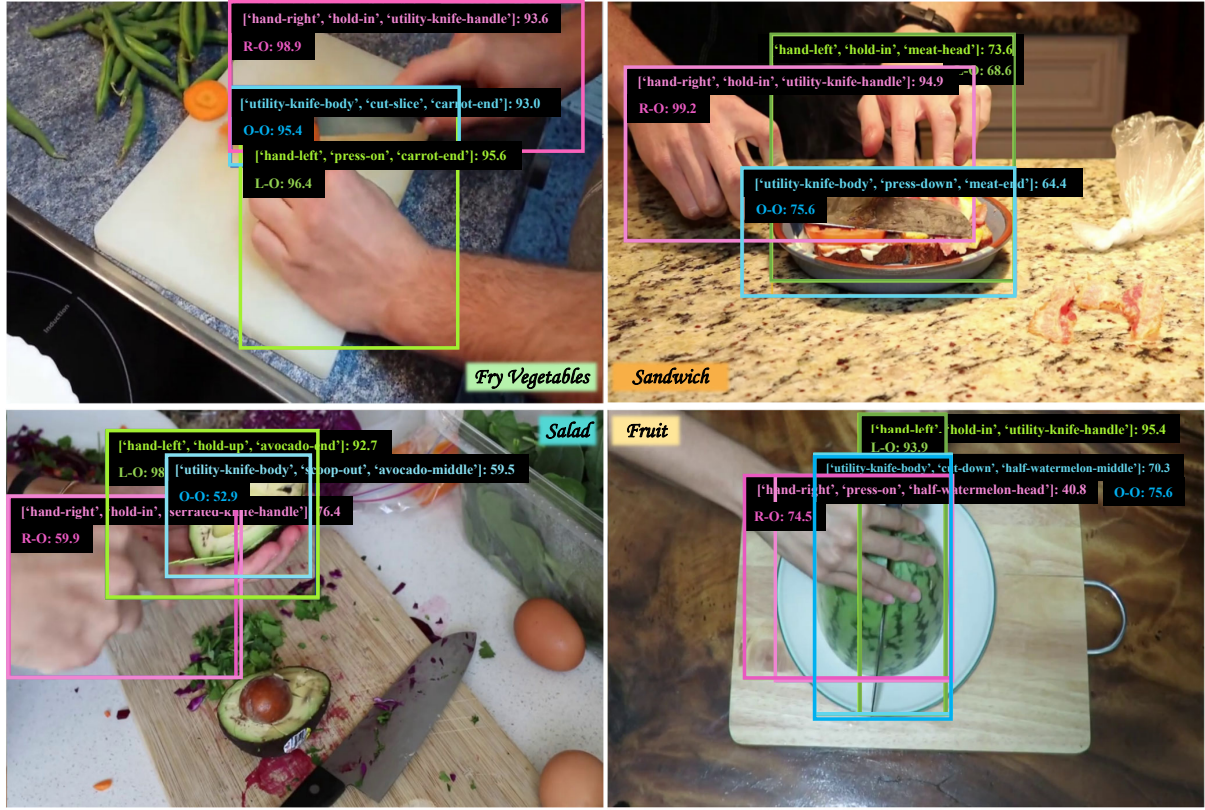


Figure 9: Visual detection results of our method in four different kitchens scenarios containing complex hand actions, i.e., “fry vegetables”, “sandwich”, “salad”, and “fruit”. Our model offers accurate bounding boxes and multi-granularity hand action information for three hand sub-interaction regions.

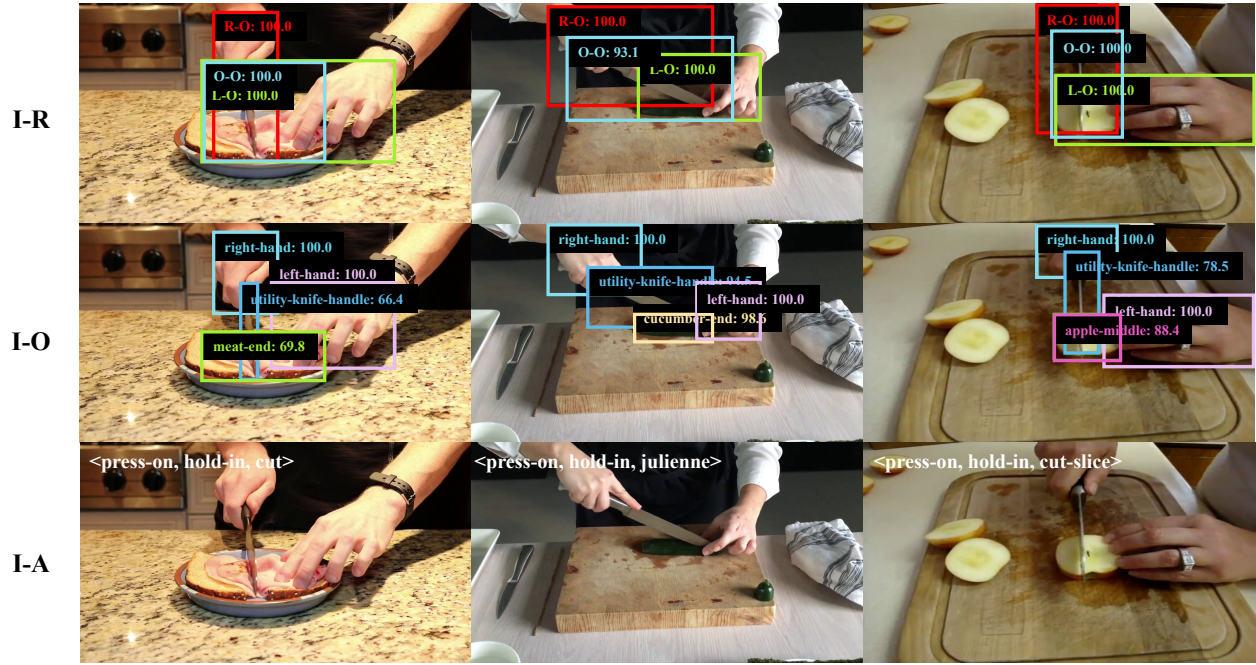


Figure 10: Visual results of Faster-RCNN [14] and TSN [19] methods in the SL-AD and SL-AR track experiments on our FHA-Kitchens dataset, showcasing interaction scenes with *three* hand sub-interaction regions, i.e., “Left hand-Object interaction region (L-O)”, “Right hand-Object interaction region (R-O)”, and “Object-Object interaction region (O-O)”. I-R: Interaction Region, I-O: Interaction Object, I-A: Interaction Region Action Verb.

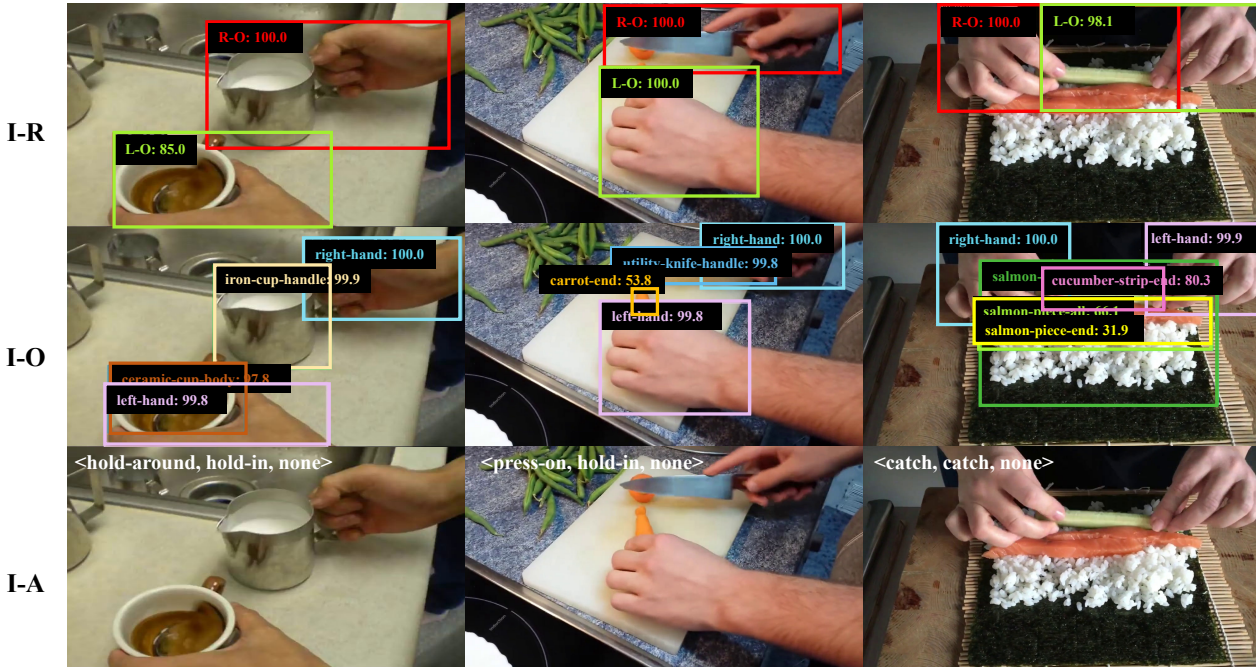


Figure 11: Visual results of Faster-RCNN [14] and TSN [19] methods in the SL-AD and SL-AR track experiments on our FHA-Kitchens dataset, showcasing interaction scenes with *two* hand sub-interaction regions, i.e., “Left hand-Object interaction region (L-O)” and “Right hand-Object interaction region (R-O)”. I-R: Interaction Region, I-O: Interaction Object, I-A: Interaction Region Action Verb.

I-R

I-O

I-A

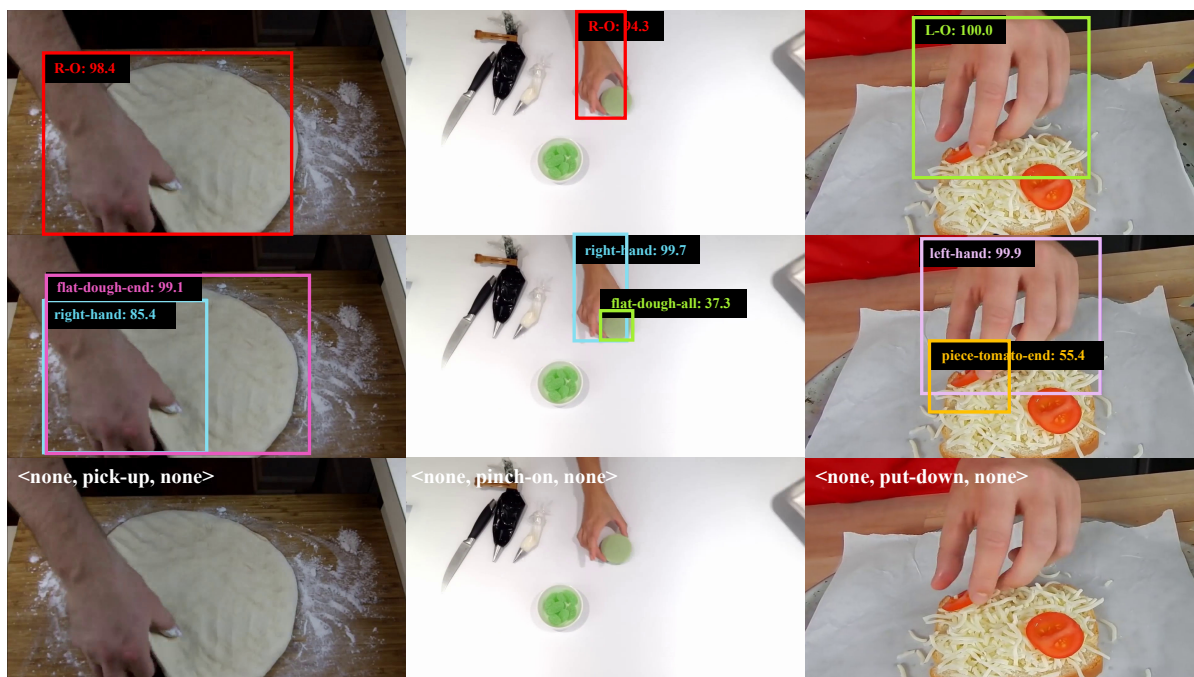


Figure 12: Visual results of Faster-RCNN [14] and TSN [19] methods in the SL-AD and SL-AR track experiments on our FHA-Kitchens dataset, showcasing interaction scenes with a *single* hand sub-interaction region, i.e., “Left hand-Object interaction region (L-O)” or “Right hand-Object interaction region (R-O)”. I-R: Interaction Region, I-O: Interaction Object, I-A: Interaction Region Action Verb.

Table 11: Vocabulary of fine-grained hand action verbs.

ID	Verb	#Instance	ID	Verb	#Instance
0	hold-around	1,593	65	knead	298
1	hold-at	788	66	contain	144
2	fill-with	265	67	roll-on	82
3	pinch-on	1,115	68	stick-to	50
4	rub-around	45	69	touch-to	12
5	hold-in	20,520	70	smooth-out	144
6	touch-on	42	71	sprinkle-on	203
7	hold-with	341	72	squeeze-around	189
8	press-on	9,369	73	press-down	675
9	cut-out	889	74	cut-up	100
10	fix-on	2,037	75	shovel-up	123
11	peel-off	1,413	76	grab-out	1
12	slice-along	1,306	77	close	51
13	grab	2,531	78	rotate	51
14	cut-half	230	79	open	72
15	take-up	134	80	open-down	27
16	pinch	609	81	hold-down	41
17	catch	446	82	cut-dice	443
18	put-down	1,406	83	dig-seeds	124
19	roll-up	1,296	84	chop	346
20	fix	293	85	push-forward	8
21	scrub-inside	53	86	cut-halves	22
22	lay-down	111	87	peel	87
23	hold-onto	453	88	push-ahead	2
24	pick-up	526	89	screw-on	5
25	cut-slice	2,808	90	sprinkle-into	16
26	take-out	178	91	scoop-up	85
27	turn-off	22	92	hold-along	129
28	cut-down	1,040	93	scrape-on	331
29	cut-off	820	94	stick-with	13
30	grab-up	115	95	cut-in	275
31	put-up	56	96	rub-on	11
32	break-apart	221	97	put-on	2
33	touch	483	98	push-off	10
34	cut-into-halves	136	99	place-on	17
35	bring-up	140	100	cut	15
36	pour-out	832	101	dip-in	9
37	pour-into	395	102	stretch-out	31
38	pour	154	103	flip	8
39	scrape	70	104	set-aside	22
40	rotate-around	162	105	julienne	165
41	screw-down	19	106	unroll	270
42	remove-out	69	107	adjust	46
43	hold-up	664	108	place-down	262
44	scoop-out	76	109	pile	49
45	open-up	21	110	pull	137
46	hold	946	111	attach-to	18
47	hold-on	187	112	grab-in	15
48	squeeze	334	113	knock-on	29
49	squeeze-out	254	114	press-against	13
50	mix-together	187	115	stir-in	114
51	spread-on	342	116	pull-up	47
52	twist-off	20	117	point-at	25
53	wrap-around	204	118	pull-out	33
54	break-off	417	119	scrape-down	6
55	grab-at	621	120	grab-onto	558
56	grab-on	384	121	hold-into	88
57	cut-through	25	122	hold-over	67
58	chop-dice	63	123	stir-into	49
58	chop-dice	63	124	press-onto	1
60	insert-into	106	125	roll	68
61	put-in	30	126	roll-out	33
62	dig-out	23	127	dip	51
63	cut-chunks	91	128	brush-onto	54
64	churn	369	129	flip-over	83

Table 12: Vocabulary of fine-grained interaction object nouns.

Super category	ID	Noun	#Instance	Super category	ID	Noun	#Instance
Vegetables&Plants	0	basil-end	201	Fruits	65	apple-all	22
	1	beet-end	143		66	apple-end	34
	2	beet-head	136		67	apple-head	253
	3	beet-middle	36		68	apple-middle	301
	4	bell-pepper-all	2		69	avocado-end	156
	5	bell-pepper-end	210		70	avocado-head	12
	6	bell-pepper-head	67		71	avocado-left	34
	7	bell-pepper-middle	139		72	avocado-middle	174
	8	broccoli-head	38		73	avocado-right	34
	9	carrot-end	1,625		74	block-watermelon-edge	5
	10	carrot-head	205		75	green-melon-all	4,074
	11	carrot-middle	638		76	green-melon-end	68
	12	chopped-vegetables-surface	38		77	green-melon-middle	68
	13	courgette-end	921		78	half-apple-head	139
	14	courgette-middle	39		79	half-apple-middle	47
	15	cucumber-end	288		80	half-pineapple-head	6
	16	cucumber-middle	165		81	half-pineapple-middle	99
	17	cucumber-strip-all	14		82	half-tomato-end	1
	18	cucumber-strip-end	72		83	half-tomato-middle	1
	19	cucumber-strip-middle	22		84	half-watermelon-edge	79
	20	garlic-middle	240		85	half-watermelon-end	17
	21	garlic-end	164		86	half-watermelon-head	301
	22	garlic-head	46		87	half-watermelon-middle	232
	23	ginger-end	248		88	lemon-end	156
	24	ginger-head	169		89	lemon-middle	108
	25	ginger-middle	70		90	melon-skin-all	119
	26	green-beans-end	937		91	melon-skin-end	576
	27	green-pepper-dice	1		92	melon-pulp-all	28
	28	green-pepper-end	710		93	melon-pulp-end	163
	29	green-pepper-head	142		94	melon-pulp-middle	49
	30	green-pepper-middle	505		95	melon-slice-end	200
	31	half-bell-pepper-end	116		96	orange-all	22
	32	half-bell-pepper-middle	110		97	orange-end	24
	33	half-onion-all	23		98	orange-head	209
	34	half-onion-head	11		99	orange-middle	547
	35	half-onion-middle	11		100	peelless-orange-middle	93
	36	mushroom-middle	15		101	piece-orange-edge	276
	37	nori-all	506		102	pineapple-all	26
	38	nori-end	262		103	pineapple-end	476
	39	onion-end	78		104	pineapple-head	959
	40	onion-head	28		105	pineapple-middle	1,346
	41	onion-middle	49		106	slice-pineapple-end	25
	42	pepper-seeds-all	4		107	slice-pineapple-middle	63
	43	piece-onion-middle	38		108	watermelon-edge	29
	44	piece-tomato-end	41		109	watermelon-end	966
	45	purple-cabbage-end	77		110	watermelon-head	155
	46	purple-cabbage-head	70		111	watermelon-middle	631
	47	purple-cabbage-middle	23	Dairy&Eggs	112	boiled-egg-end	226
	48	red-pepper-all	21		113	boiled-egg-head	22
	49	red-pepper-head	25		114	boiled-egg-middle	14
	50	red-pepper-middle	18		115	boiled-egg-shell	88
	51	small-tomato-head	9		116	egg-all	248
	52	small-tomato-middle	27		117	egg-head	1
	53	spinach-end	4		118	egg-middle	26
	54	spinach-head	35		119	egg-liquid-all	10
	55	spinach-middle	30		120	egg-shell-all	86
	56	spring-garlic-all	18		121	egg-shell-edge	34
	57	spring-garlic-end	53		122	milk-all	594
	58	spring-garlic-head	23		123	yolk-all	112
	59	spring-garlic-middle	52	Meat&Fish	124	chicken-dice	8
	60	sun-flower-seeds	104		125	raw-chicken-dice	69
	61	tomato-cube	22		126	crab-shred	328
	62	tomato-end	280		127	meat-end	515
	63	tomato-middle	17		128	meat-head	573
	64	tomato-sliced-middle	109		129	meat-middle	142

Table 13: Vocabulary of fine-grained interaction object nouns.

Super category	ID	Noun	#Instance	Super category	ID	Noun	#Instance
Meat&Fish	130	meat-piece-end	442	Containers	194	bottle-body	52
	131	meat-slice-all	1		195	box-lid-bottom	21
	132	meat-slice-end	202		196	bottle-cap	8
	133	piece-pepperoni-all	87		197	bottle-cap-all	106
	134	piece-pepperoni-end	158		198	bottle-cap-bottom	19
	135	salmon-piece-all	36		199	can-cover-edg	2
	136	salmon-piece-end	399		200	can-cover-edge	106
	137	salmon-piece-middle	14		201	ceramic-cup-all	68
	138	salmon-slice-end	44		202	ceramic-cup-body	671
Spices&Sauces	139	butter-all	45		203	ceramic-cup-handle	12
	140	crumbles-cheese-all	116		204	ceramic-lid-all	25
	141	cheese-all	27		205	ceramic-lid-edge	85
	142	green-butter-all	85		206	ceramic-teapot-handle	132
	143	mozzarella-all	84		207	ceramic-teacup-body	69
	144	mozzarella-end	12		208	ceramic-teacup-edge	138
	145	pizza-sauce-end	193		209	cup-edge	79
	146	powder-all	29		210	glass-bottle-edge	40
	147	sauce-all	397		211	glass-bottle-top	59
	148	slice-cheese-end	44		212	glass-cup-body	1
	149	tomato-sauce-all	151		213	glass-cup-edge	217
Liquids	150	tomato-sauce-edge	15		214	glass-cup-handle	51
	151	sauce-mixed	70		215	glass-goblet-stem	333
	152	can-opener	108		216	glastic-bottle-edge	4
	153	green-mixture-all	248		217	glastic-bottle-top	4
	154	jam-all	23		218	grass-bottle-top	30
Baked&Baking	155	oil-all	60		219	iron-basin-body	29
	156	olive-oil-all	51		220	iron-basin-edge	145
	157	baking-paper-edge	99		221	iron-basin-middle	115
	158	baking-paper-top	25		222	iron-cup-body	1,005
	159	baking-plate-edge	54		223	iron-cup-handle	325
Cooked Food	160	piece-pizza-end	125		224	iron-dipper-handle	14
	161	pizza-all	63		225	plastic-basin-edge	33
	162	pizza-end	27		226	plastic-bottle-bottom	21
	163	pizza-middle	29		227	plastic-bottle-edge	352
	164	sandwich-edge	32		228	plastic-bottle-top	78
	165	sandwich-end	297		229	plastic-cup-body	19
	166	sandwich-head	141		230	sauce-container-end	4
	167	sandwich-middle	123		231	small-cup-edge	195
	168	sandwich-side	85		232	small-plastic-bottle-edge	154
	169	sandwich-top	27		233	small-plastic-bottle-end	120
	170	sandwich-all	23		234	small-plastic-bottle-top	167
	171	sushi-roll-all	151		235	teapot-lid-edge	126
	172	sushi-roll-end	1,659		236	teapot-lid-handle	267
	173	sushi-roll-head	233		237	wine-bottle-bottom	75
Packaging	174	sushi-roll-middle	622		238	yogurt-box-bottom	63
	175	bamboo-mat-edge	284	Cutlery	239	yogurt-box-edge	62
	176	bamboo-mat-end	528		240	yogurt-box-handle	2
	177	bamboo-mat-head	8		241	yogurt-box-top	14
	178	bamboo-mat-middle	244		242	sauce-cup-all	9
	179	mozzarella-bag-end	71		243	bowl-bottom	69
	180	mozzarella-bag-middle	24		244	bowl-edge	198
	181	onion-bag-end	86		245	glass-bowl-all	23
	182	onion-bag-middle	30		246	glass-bowl-body	23
	183	pepperoni-bag-end	60		247	glass-bowl-bottom	91
	184	pepperoni-bag-middle	16		248	glass-bowl-edge	415
	185	piping-bag-all	251		249	glass-bowl-handle	8
	186	pizza-box-edge	140		250	grass-bowl-edge	13
	187	tea-leaves-bag-body	38		251	green-bowl-edge	29
	188	tea-leaves-bag-bottom	42		252	small-bowl-edge	16
	189	tea-leaves-bag-top	38		253	steel-bowl-edge	153
Containers	190	black-bottle-top	42		254	steel-bowl-top	13
	191	bottle-all	1		255	metal-bowl-edge	470
	192	bottle-edge	19		256	plastic-bowl-all	46
	193	bottle-top	30				

Table 14: Vocabulary of fine-grained interaction object nouns.

Super category	ID	Noun	#Instance	Super category	ID	Noun	#Instance
Cutlery	257	plastic-bowl-body	10	Kitchenware	321	shovel-body	49
	258	plastic-bowl-edge	31		322	shovel-handle	49
	259	porcelain-bowl-edge	39		323	sieve-spoon-body	13
	260	porcelain-bowl-middle	2		324	sieve-spoon-handle	12
	261	porcelain-bowl-top	17		325	small-iron-pot-handle	90
	262	fork-handle	9		326	small-knife-body	663
	263	iron-spoon-body	353		327	small-knife-handle	706
	264	iron-spoon-handle	395		328	tea-strainer-body	108
	265	plastic-scoop-body	48		329	tea-strainer-edge	4
	266	plastic-scoop-handle	94		330	tea-strainer-handle	224
	267	plastic-spoon-body	22		331	turnplate-corner	15
	268	plastic-spoon-handle	22		332	turnplate-edge	9
	269	spoon-body	692		333	utility-knife-body	10,419
	270	spoon-handle	1,011		334	utility-knife-handle	11,157
	271	tablespoon-body	327		335	carrot-peeler-body	255
	272	tablespoon-handle	332		336	carrot-peeler-handle	518
	273	teaspoon-body	51	Appliances	337	grater-body	70
	274	teaspoon-handle	191		338	grater-handle	70
	275	wooden-spoon-body	116		339	oven-door-handle	22
Kitchenware	276	wooden-spoon-handle	152	Rice&Flour	340	ball-dough-end	27
	277	wooden-spatula-body	25		341	ball-dough-head	24
	278	wooden-spatula-handle	28		342	ball-dough-middle	173
	279	table-knife-handle	34		343	dough-all	413
	280	tableware-handle	11		344	dough-flour-all	117
	281	beater-end	25		345	dough-flour-middle	35
	282	plate-edge	335		346	flat-dough-all	42
	283	plate-end	20		347	flat-dough-edge	206
	284	brush-body	105		348	flat-dough-end	963
	285	brush-handle	195		349	flat-dough-middle	13
	286	can-opener-edge	83		350	flour-all	35
	287	can-opener-end	25		351	green-dough-all	123
	288	ceramic-plate-all	91		352	little-dough-al	1
	289	ceramic-plate-edge	208		353	little-dough-all	386
	290	chopping-board-edge	2		354	oval-dough-end	170
	291	cooking-spoon-body	77		355	rice-all	585
	292	cooking-spoon-handle	382		356	slice-bread-end	92
	293	food-mixer-handle	372		357	bread-end	510
	294	food-mixer	369		358	bread-head	199
	295	pizza-cutter-body	29		359	bread-middle	258
	296	pizza-cutter-handle	17	Dessert	360	chocolate-cake-edge	242
	297	food-plate-center	75		361	chocolate-cake-top	111
	298	food-plate-edge	36		362	chocolate-all	21
	299	handle	108		363	candies-all	159
	300	iron-plate-edge	76		364	candy-all	4
	301	kitchen-knife-body	552		365	candy-top	51
	302	kitchen-knife-handle	1,215		366	chocolate-bar-middle	87
	303	knife-body	26		367	chocolate-chips-all	49
	304	knife-handle	27		368	chocolate-cream-all	189
	305	jam-knife-body	50		369	chocolate-cream-top	261
	306	jam-knife-handle	53		370	chocolate-donut-all	35
	307	metal-plate-edge	47		371	cookie-all	41
	308	metal-spatula-body	71		372	cookie-end	57
	309	metal-spatula-handle	123		373	cookie-head	1
	310	pizza-spatula-body	70		374	cookies-all	49
	311	pizza-spatula-handle	99		375	cookie-top	9
	312	pizza-tray-edge	122		376	cream-all	139
Kitchenware	313	plastic-spatula-body	328	Drink	377	tea-all	14
	314	plastic-spatula-handle	531		378	tea-leaves-all	219
	315	rolling-pin-body	198		379	whisk-head-top	9
	316	rolling-pin-handle	174	Uncategorised	380	hand-left	23,305
	317	rolling-pin-middle	82		381	hand-right	26,441
	318	rolling-pin-miidle	5		382	towel-all	82
	319	serrated-knife-body	38		383	towel-edge	38
	320	serrated-knife-handle	43				



Figure 13: The distribution of instances per object noun category from 17 super-categories in the FHA-Kitchens dataset.