2HandedAfforder: Learning Precise Actionable Bimanual Affordances from Human Videos

Marvin Heidinger*[1], Snehal Jauhri*[1], Vignesh Prasad[1], Georgia Chalvatzaki[1,2]
* indicates equal contribution
[1]Computer Science Department, Technische Universität Darmstadt, Germany
[2]Hessian.AI, Darmstadt, Germany
{snehal.jauhri, vignesh.prasad, georgia.chalvatzaki}@tu-darmstadt.de

***Abstract*—When interacting with objects, humans effectively reason about which regions of objects are viable for an intended action, i.e., the affordance regions of the object. They can also account for subtle differences in object regions based on the task to be performed and whether one or two hands need to be used. However, current vision-based affordance prediction methods often reduce the problem to naive object part segmentation. In this work, we propose a framework for extracting affordance data from human activity video datasets. Our extracted 2HANDS dataset contains precise object affordance region segmentations and affordance class-labels as narrations of the activity performed. The data also accounts for bimanual actions, i.e., two hands co-ordinating and interacting with one or more objects. We present a VLM-based affordance prediction model, 2HandedAfforder, trained on the dataset and demonstrate superior performance over baselines in affordance region segmentation for various activities. Finally, we show that our predicted affordance regions are actionable, i.e., can be used by an agent performing a task, through demonstration in robotic manipulation scenarios.**

I. Introduction

When humans perceive objects, they understand different object regions and can predict which object region *affords* which activities [8], i.e., which object regions can be used for a task. We wish our machines to have this ability, referred to in literature as "affordance grounding". Affordance grounding has several downstream applications, including building planning agents, VR, and robotics. Affordance grounding is especially important for robotics since robots must reason about various actions that can be performed using different object regions which is a crucial step towards performing useful tasks in everyday, unstructured environments. For example, to pour into a bowl, the robot should know that it should hold the bottle in a region close to the center of mass of the bottle (Figure 1), i.e., a region that *affords* pouring. Predicting such affordance regions is challenging since it requires a fine-grained understanding of object regions and their semantic relationship to the task.

Recent advances in large-language and multimodal models have shown impressive visual reasoning capabilities using self-supervised objectives [38, 32, 7]. However, there is still a big gap in their ability to detect accurate object affordance regions in images [23]. Moreover, most existing state-of-the-art affordance detection methods [13, 35, 37, 45, 20] use labeled data [31, 35, 28, 15, 21] that lacks precision and is more akin to object part segmentation rather than *actionable* affordance-region prediction. When humans interact with objects, they are much more *precise* and use specific object regions important in the context of the task. An example is provided in Fig. 1. For the task of pouring into the bowl, part segmentation labels the entire bottom of the bottle with the affordance 'pour'. But, to pour correctly, humans leverage the appropriate region of the bottle. Moreover, the affordances are inherently bimanual, i.e., the affordance regions of the bowl and bottle are interconnected.

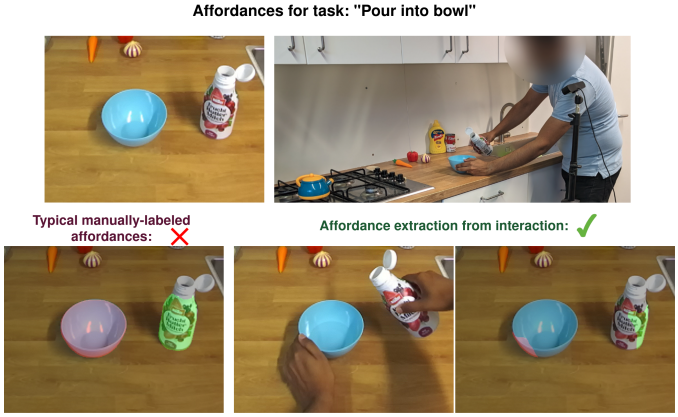

Fig. 1: **A motivating example:** When required to label affordances for a task 'Pour into bowl', the typically labeled affordances provided by annotators are not precise and reduce the problem to object part segmentation. Alternatively, our affordance extraction method uses the hand-object interaction sequence to obtain precise, actionable, bimanual object affordance regions.

We argue that affordances should not be labeled but automatically extracted from observations of humans performing tasks, such as human activity video datasets. We propose a method that uses hand-inpainting and mask completion to extract affordance regions occluded by human hands. This

Dataset	Image type & source	# Images	Affordance data					
			Annotation source	Annotation type	# Aff. classes	# Obj. classes	Class-labels	Bimanual
IIT-AFF [31]	Exocentric [41]	8.8K	Manually-labeled	Masks	9	10	Explicit	No
AGD20K [28]	Exo+Egocentric [25, 3]	23.8K	Manually-labeled	Heatmaps	36	50	Explicit	No
3DOI [35]	Exo+Egocentric [36, 5]	10K	Manually-labeled	Points	3	n.a.	Explicit	No
ACP [10]	Egocentric [5]	15K	Auto-labeled	Heatmaps	n.a.	n.a.	none	No
VRB [1]	Egocentric [5]	54K	Auto-labeled	Heatmaps	n.a.	n.a.	none	No
2HANDS	Egocentric [5]	**278K**	Auto-labeled	**Precise Masks**	**73**	**163**	**Narrations**	**Yes**

TABLE I: Comparison of our dataset 2HANDS against other affordance prediction datasets. For 2HANDS, we auto-label a large number of affordance region masks from human egocentric videos and use narration-based affordance class-labels. Our dataset also contains bimanual masks, with the goal of addressing the challenging problem of precise bimanual affordance prediction in images.

has several advantages. First, by using this procedure, we are able to obtain **bimanual** and **precise** affordances (Figure 1) rather than simply predicting object parts. Second, it makes affordance specification more natural since it is often easier for humans to *show* the object region to interact with, rather than label and segment it correctly in an image. Third, using human activity videos gives us diverse task-specific affordances, with the affordance class label naturally coming from the narration of what task is being done by the human. This makes our affordances **task-oriented** with natural language specification, unlike previous methods focused on predicting task-agnostic interaction hotspots [10, 1].

We extract a dataset, 2HANDS (2-Handed Affordance + Narration DataSet), consisting of a large number of unimanual and bimanual object affordance segmentation masks and task narrations as affordance class-labels. We propose a VLM-based affordance prediction model, 2HandedAfforder, that is trained on the 2HANDS dataset and predicts affordance masks in images based on an input text prompt. To evaluate the performance on this challenging problem, we also present a novel benchmark, ActAffordance, using annotations on images from two egocentric human activity datasets [5, 11].

Our contributions are:

- a method to extract precise affordance regions from human-object interaction videos.
- a dataset, 2HANDS, consisting of 278K images with extracted affordance masks, narration-based class labels, and unimanual/bimanual taxonomy labels.
- an affordance network, 2HandedAfforder, for predicting task-aware unimanual and bimanual affordance regions.
- a novel benchmark, ActAffordance, to evaluate affordance detection aligned with the "ground truth" considered by humans.
- the first comprehensive dataset and evaluation of task-specific bimanual object affordance regions in images.

II. Related Work

Fully supervised affordance detection. In fully supervised affordance detection datasets and methods such as by Qian and Fouhey [35], AffordanceLLM [37], the dataset is fixed and hand-annotated such as from IIT-AFF [31] and 3DOI [35]. The affordance classes in these datasets are explicit and annotators guess which affordance class may apply to object regions. Other methods, such as VLPart [45], use a general open vocabulary segmentation pipeline. LISA [20] performs open-vocabulary, prompt-based "reasoning segmentation". However, these methods do not consider actions and typically segment either the whole object [20] or object parts[45], and not precise affordance regions.

Weakly supervised affordance detection. Weakly supervised methods such as Cross-viewAG [28] and Locate [22] learn to predict affordances by observing exocentric images of humans interacting with objects based on the AGD20K dataset [28]. The model maps object parts across images, transferring the learned affordances to non-exocentric images where no hand-object interaction occurs. This is similar to saliency matching methods that use one-shot affordance transfer [46, 14]. However, these methods still require an initial smaller manually-labeled dataset with explicit affordance classes.

Auto-labeled affordance detection. Egocentric videos of humans performing tasks [5, 6, 11, 47, 12] are an attractive option for extracting affordance data since they include object interactions up close and in the camera field of view. Recently, Goyal et al. [10] and Bahl et al. [1] have shown that videos from datasets such as EPIC kitchens [5] and Ego4D [11] can be used to segment regions of interest in objects using weak supervision from hand and object bounding-boxes. However, these works focus on segmenting task-agnostic 'hotspot' interaction regions of objects. The region of interactions do not consider the task and whether one or two hands would be needed.

Our approach and goals. In this work, we propose a method to extract affordance masks leveraging recent video-based hand inpainting techniques [2]. Since our dataset contains precise segmentation masks, we can predict pixel-wise affordance segments in the image, as opposed to methods only trained with point-labels of affordance [35] or that only predict heatmaps [28, 1]. Moreover, we consider the especially challenging problem of bimanual affordance detection, for which the spatial context of the objects and their interconnection is also important. Although bimanual affordances have been considered in previous work [19, 9, 42, 33], to the best of our knowledge, ours is the first method to extract bimanual affordances from videos which we then use to train our model to predict task-specific affordance masks based on a text prompt.

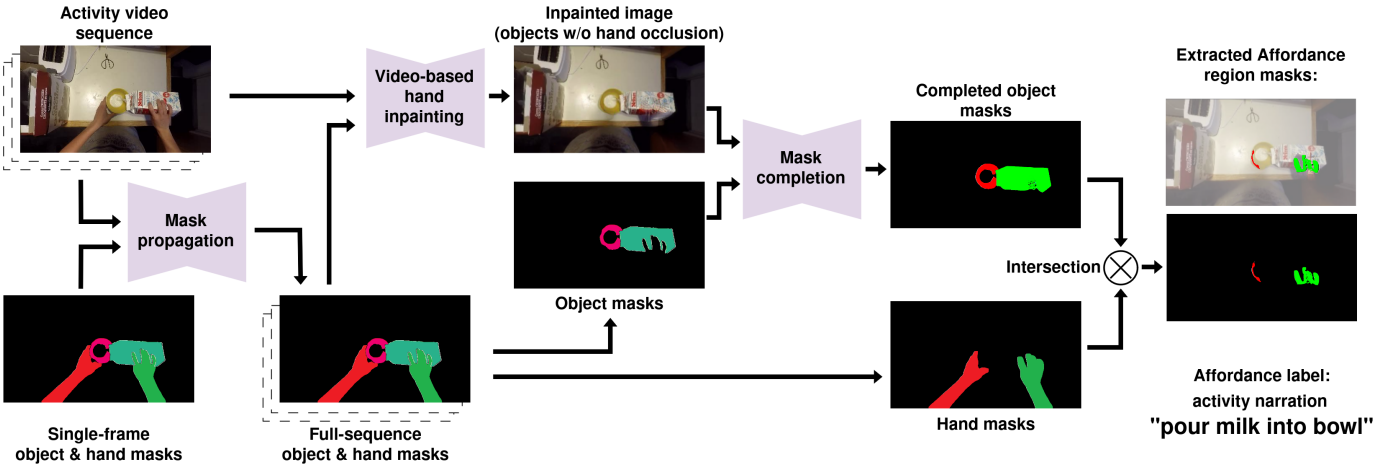

Fig. 2: Affordance extraction pipeline. Given a human activity video sequence and a single-frame object and hand masks, we first obtain dense, full-sequence object and hand masks using a video mask-propagation network [4]. We then inpaint out the hands in the RGB images using a video-based hand inpainting model [2]. This gives us an image with the objects reconstructed and un-occluded by the hands. With the inpainted image and the original object masks, we use [39] to "complete" the object masks by again propagating the object masks to the inpainted image. Finally, we can extract the affordance region masks for the given task as the intersection between the completed masks and the hand masks. We also label the affordance class using the narration of the task.

III. Extraction and Learning of Bimanual Affordances from Human Videos

In this section, we detail our affordance extraction approach used to generate our 2HANDS dataset from videos of humans performing everyday tasks (Sec. III-A). Then, we present our approach, "2HandedAfforder", for predicting meaningful task-oriented bimanual affordance regions in images in Sec. III-B.

A. Affordance Extraction from Human Videos

We use videos of humans performing tasks to extract precise affordance masks. This involves closely examining the contact regions between the hands and objects in the videos. Several recent methods [44, 34] have shown impressive performance in hand-object segmentation and reconstruction. However, the challenge in affordance region extraction is that the hand typically occludes the object region with which it interacts. Bahl et al. [1] circumvent this issue by only considering videos where objects are initially un-occluded before the interaction and only use the hand bounding-box to denote the interaction region. However, not only is this a limiting assumption, but also the bounding-boxes can only be used to detect interaction hotspots and not precise object affordance masks. Precise masks are more explicit and useful for downstream application, for example, for providing graspable regions of an object for robotic manipulation tasks. We propose a pipeline to extract affordances that leverages recent advances in hand inpainting [2] and object mask completion [43, 39], providing the first bimanual affordance region segmentation dataset. Moreover, we use the narration of the task being performed as the affordance text label, which helps obtain a diverse set of affordance classes for various objects. The full extraction pipeline is visualized in Figure 2.

We extract affordances from EPIC-KITCHENS [5], which contains ∼100 hours of egocentric videos of human activities in kitchens. We use the VISOR [6] annotations of the dataset, which contain some sparse hand-object mask segmentations and binary labels denoting whether the hand is in contact with the object. Note that we can also use other video datasets like Ego4D [11] along with hand segmentation methods [44] to extract hand-object masks. To obtain dense hand-object masks for entire video sequences, we use a video-based mask propagation network [4].

With the hand and object masks available over the entire video sequence, we obtain an un-occluded view of the objects by inpainting out the hands. We use a video-based hand inpainting model, VIDM [2], that uses 4 frames from the sequence as input to inpaint the missing regions. This makes reconstructing the target objects more likely if the objects have already been in another frame of the action sequence without occlusion. Inpainting provides us with an un-occluded view of the objects. We then precisely segment these un-occluded objects in the inpainted image using mask completion. For this, we use the segmentation masks from the original image and prompt SAM2 [39] to propagate these masks to the new inpainted image. We observe that this process gives us more precise object masks compared to directly using mask completion methods [43], detailed in the appendix (Sec. F).

To obtain the final affordance region where the hand interacted with the object, we can simply compute the intersection of the un-occluded object masks and the hand masks. The full pipeline is shown in Fig. 2. For bimanual affordances, it is also useful to classify the affordances into a bimanual taxonomy [19]. Thus, we distinguish between unimanual left, unimanual right, and bimanual actions. Additional details about the extraction procedure have been provided in the

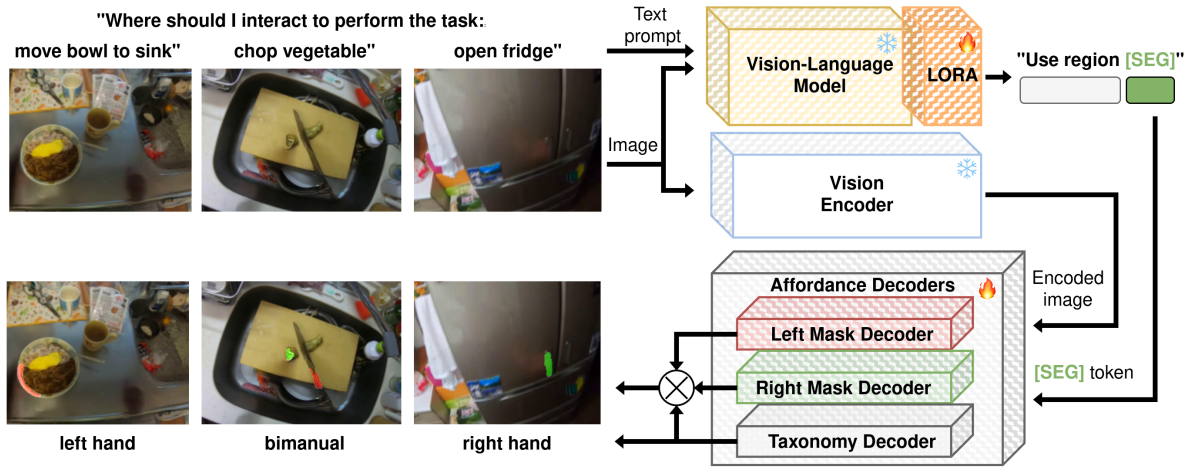

Fig. 3: Affordance prediction network. Given an input image and task, we use a question asking where the objects should be interacted for the desired task as a text prompt to a Vision-Language model (VLM). The VLM produces language tokens and a [SEG] token which is passed to the affordance decoders. We also use a SAM [18] vision-backbone to encode the image and pass it to the affordance decoders. The decoders predict the left hand and right hand affordance region masks as well as a taxonomy classification indicating whether the interaction is supposed to be performed with the left hand, right hand, or both hands. The vision encoder is frozen, while the VLM predictions are fine-tuned using LORA [16].

appendix.

With the above procedure, we obtain a dataset of 278K images with extracted affordance segmentation masks, narration-based class-labels, and bimanual taxonomy annotations. We call this dataset 2HANDS, i.e., the **2**-**H**anded **A**ffordance + **N**arration **D**ata**S**et.

B. Task-oriented Bimanual Affordance Prediction

Reasoning segmentation, i.e., text-prompt-based segmentation of full objects, is a difficult task. Segmentation of precise object affordance regions is even more challenging. The complexity is further increased when considering bimanual affordances with multiple objects. To address this challenge, we develop a model for general-purpose bimanual affordance prediction that can process both an input image and any task prompt (e.g., "pour tea from kettle"). We call this model "2HandedAfforder." We leverage recent developments in reasoning-based segmentation methods [24, 20] and train a VLM-based segmentation model to reason about the required task and predict the relevant affordance region in the input image. Since our 2HANDS dataset provides precise segmentation masks, we can predict pixel-wise affordance segments in the image, as opposed to other methods that are only trained with point labels of affordance [35] or that only predict heatmaps [28, 1].

Inspired by reasoning segmentation methods such as by Lai et al. [20], we use a Vision-Language Model (VLM) [27] to jointly process the input text prompt and image and produce language tokens and a segmentation [SEG] token as output. While VLMs excel at tasks such as visual question answering and image captioning, they are not explicitly optimized for vision tasks like segmentation, where accurately predicting pixel-level information is key. Therefore, to have a stronger vision-backbone for our segmentation-related task, we use a modified version of SAM [18]. Given the combined embedding provided by the VLM [SEG] token and SAM image encoder, we use affordance decoders modeled after SAM-style mask decoders to predict the affordances. We use two mask decoders, generating two separate affordance masks for the left and right hands, respectively. Furthermore, we add a prediction head to one of the decoders that takes the output token as input and predicts the bimanual taxonomy: 'unimanual left hand', 'unimanual right hand', and 'bimanual' using a separate fully-connected classifier decoder. An overview of the whole affordance prediction network architecture is visualized in Figure 3.

The VLM is trained to generate a specific output token: a segmentation [SEG] token. Specifically, inspired by LISA [20], we use question-answer templates to encapsulate the narration of the individual tasks in natural language, e.g. "**USER:** [IMAGE] Where would you interact with the objects to perform the action {*action_narration*} in this image? **ANSWER:** Use region: [SEG]." This [SEG] token encapsulates the general-purpose reasoning information from the VLM for the task which is then used by the affordance decoders. For the left and right hand mask decoders, we initialize the decoders with pre-trained SAM weights and train them to predict segmentation masks using the encoded image and [SEG] token as input. For the taxonomy classifier decoder, as in [35], we pass the left mask decoder output token through an MLP to predict whether the action should be performed with the left hand, right hand, or both hands.

We freeze the weights of the image encoder and the VLM, and we apply Low-Rank Adaptation (LoRA) [16] to fine-tune the VLM. By introducing trainable low-rank updates, LoRA enables efficient fine-tuning of the VLM without requiring modifications to its original parameters. This ensures that the pre-trained knowledge of the VLM, a LLaVa-13b, is preserved

while still allowing the model to specialize in segmentation. We do not fine-tune the SAM image encoder as this was shown to reduce performance in reasoning segmentation tasks. For training the mask prediction, we use a combination of dice loss [30] and focal cross-entropy loss [40]. For the taxonomy prediction, we use a cross-entropy loss with the ground truth label. If the task does not require one of the hands, the weight for the corresponding mask loss is set to 0. Similarly, when predicting affordance regions using the network at test time, we use the taxonomy prediction to infer whether left, right, or both mask predictions should be considered.

As an alternative to the VLM-based 2HandedAfforder prediction network, we also train a smaller CLIP-based [26] version of the network that uses CLIP text features instead of the VLM [SEG] token as input to the affordance decoders. We call this network '2HandedAfforder-CLIP'.

IV. Experimental Setup

With our experiments, we aim to answer the following questions:

1) Does our affordance extraction procedure for the 2HANDS dataset provide accurate affordance region segmentation data?
2) Is our 2HandedAfforder model able to predict precise unimanual and bimanual affordances? And how does it compare against baselines?
3) How well does our affordance prediction model generalize to images in-the-wild?
4) Are our affordances actionable, i.e., can they be utilized in real-world scenarios such as for robotic manipulation?

A. ActAffordance Benchmark

To answer the first question of the accuracy of our extracted affordances in the 2HANDS dataset, we evaluate the alignment of our extracted affordance masks with human-annotated affordance regions. As mentioned in Sec. III-A, when humans label affordances, they often simply label object parts and do not necessarily focus on the precise regions of interaction of the objects [28, 35]. Moreover, the second question regarding the accuracy of 2HandedAfforder is non-trivial. Using only the masks in our 2HANDS dataset as "ground truth" leads to a bias towards our extracted affordances. This may not be in alignment with what humans consider accurate affordance regions. Therefore, we propose a novel benchmark called "ActAffordance" to evaluate both the dataset quality and the predicted affordances.

For the "ActAffordance" benchmark, we asked 10 human annotators to label affordance regions with a novel approach: instead of direct segment labeling, we showed them pairs of inpainted and original hand-object interaction images. By *showing annotators example interactions*, we asked them to predict similar affordance regions. Fig. 4 illustrates this annotation pipeline. Annotators predicted ALL possible interaction regions since affordance prediction is inherently multi-modal—for instance, when closing a fridge, a human might choose any point along the door length. The benchmark

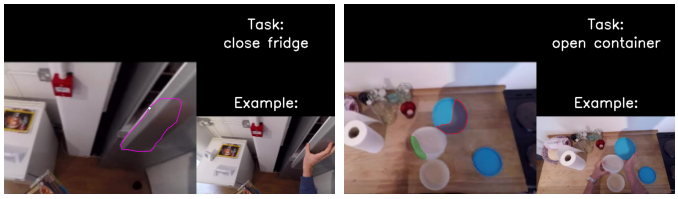

Fig. 4: Example annotations for the ActAffordance benchmark. Left: The image to be annotated with the highlighted annotation mask(s). Right: the example interaction provided to the human annotator, along with the task description. The human is asked to annotate ALL the possible regions for the interaction to capture all the different modes.

contains unimanual and bimanual segmentation masks for 400 activities from EPIC-KITCHENS [5] and Ego4D [11], with no overlap between EPIC-KITCHENS data used in 2HANDS. Details about the benchmark and annotation process are in Appendix Sec. C.

Another point of consideration when evaluating the affordance prediction is that the problem can be divided into two parts: correct identification of the objects based on the text prompt and accurate affordance region segmentation. Since these are two complementary but different capabilities, we further create another version of the benchmark called "ActAffordance-Cropped". Here, we crop the benchmark images to a bounding box containing the target objects. This helps differentiate between the capabilities of segmenting the correct object and segmenting the correct object region. Moreover, it helps evaluate our network predictions against baselines that cannot identify correct objects in images but use bounding-boxes [1] or query points on the object [35] as input.

We note that ActAffordance is a particularly challenging benchmark. To date, reasoning segmentation, i.e., text-prompt-based segmentation of full objects, is an unsolved problem. Prompt-based segmentation of precise object affordance regions is yet more challenging, especially when benchmarked against humans. The addition of bimanual affordances with multiple objects is another step beyond that. However, we feel this challenging benchmark will push the community forward towards building more effective methods for affordance prediction and thus we evaluate all methods and baselines on this benchmark instead of a simpler test set from our dataset.

B. Metrics for Evaluation

We use several metrics to evaluate the performance of the proposed models and baselines. Since we treat the affordance detection problem as a segmentation task, we use metrics widely spread across segmentation problems: precision, intersection over union (IoU), and the directed and general Hausdorff Distance (HD). We train our 2HandedAfforder and 2HandedAfforder-CLIP models on the 2HANDS dataset and evaluate on the "ActAffordance" benchmark. We evaluate performance on both the EPIC-KITCHENS and Ego4D splits of the benchmark. Note that there is no overlap between the

data from the EPIC-KITCHENS data used in 2HANDS. The evaluation on the Ego4D split of the benchmark also helps us answer the generalization question since there is no Ego4D data in 2HANDS.

Note that for the evaluation of our models, the false negative predictions play a reduced role since our models are not trained to predict all of the multimodal solutions in the benchmark but to predict *precise* affordance regions which might only cover a subset of all of the possible solutions. Thus, the key metric for comparison is **precision** over IoU. Another common metric for segmentation is the Hausdorff distance (HD). For each point in each set, the distance to the closest point from the other set is calculated and the Hausdorff distance is defined as the maximum of all of these distances. Similar to the IoU, including the distance from the ground truth to the prediction might distort the results since we aim to predict precise affordances that may only cover a smaller subset of the ground truth. Thus, we additionally provide the **directed Hausdorff distance** that only calculates the maximum distance from the prediction set points to the ground truth set.

To further show the applicability of our approach to real world robotics scenarios, we evaluate our model in-the-wild in a kitchen environment on various household objects. To show that our model can provide useful actionable affordances, we test the predictions on a real-robot system in this kitchen environment. Specifically, we use an RGBD camera mounted on a mobile manipulator robot and use the affordances predicted by our model to segment RGB images and obtain segmented point clouds. These segmented pointclouds denote where the robot should grasp objects to perform a manipulation task. For manipulation, we use pre-designed manipulation primitives for the robot and perform grasping using a 6DoF grasp prediction network.

V. Results

A. Affordance Extraction Quality

We assess the quality of the affordances obtained from our extraction pipeline (Sec. III-A) by evaluating their alignment with the human annotations in the ActAffordance benchmark. The comparison results are shown in Table II, "AffExtract", and Figure 5. As noted previously, the benchmark annotations contain all the possible modes of object interaction, while the extraction process and our models only cover a single interaction mode. Thus, precision is the most important metric to evaluate over IoU. The same principle is true for the Hausdorff distance (HD), which is why we also report Directional Hausdorff distance (Dir. HD), which only calculates the maximum distance from the prediction set points to the ground truth set. We note the precision of AffExtract is better for the Ego4D split (0.541) than the EPIC-KITCHENS split (0.334) with a combined score of 0.42. This shows a reasonably good alignment with the human-annotated segmentations from the benchmark and meaningful affordance region extraction. The IoU scores are relatively lower, with an average of 0.185, showing the challenge of the task when compared against human-level object understanding.

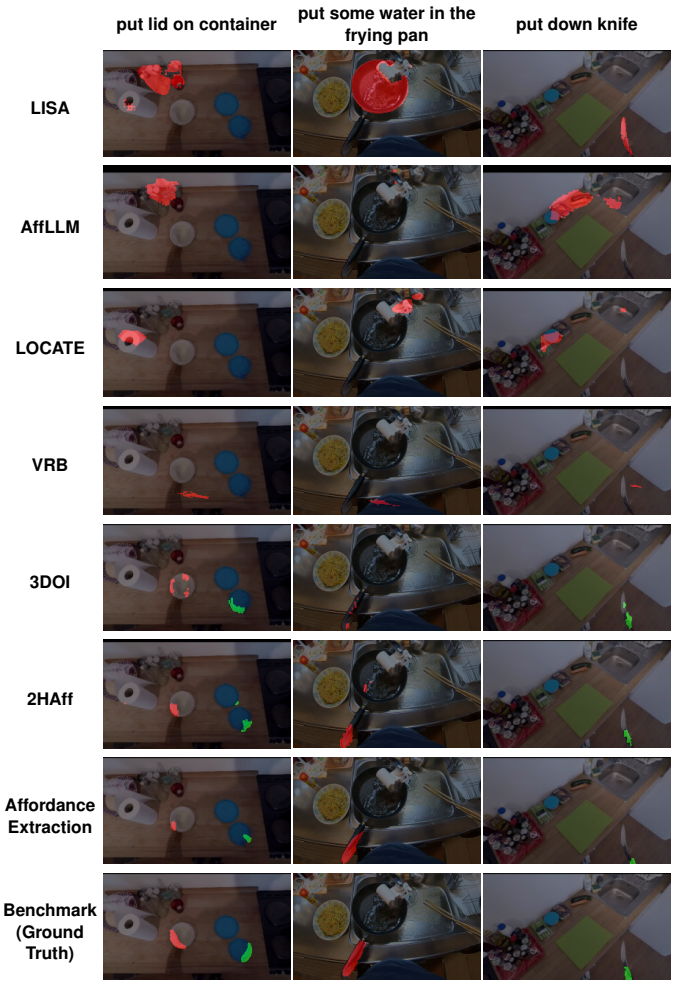

Fig. 5: Qualitative affordance prediction results on the ActAffordance benchmark. We compare our 2HandedAfforder model against LISA [20], AffordanceLLM [37] and 3DOI [35]. We also include an example result if we were to run our affordance extraction method on the activity sequence to show the quality of the extraction. Red and green masks denote the left hand and right hand affordance mask predictions, respectively.

B. Comparison against baselines on ActAffordance benchmark

Since ours is the first method to perform bimanual affordance mask detection using text prompts, there exist no directly comparable baselines. Thus, we adapt affordance detection baselines which includes a SOTA text-based reasoning segmentation baseline. Since several weakly-supervised affordance detection methods [10, 1, 37] represent affordances as only points or points+probabalistic heatmaps around them, we adapt their predictions into segmentation masks by choosing different probability thresholds at which pixels are considered to be part of the affordance region. We use the following baselines for comparison: (i) **LISA [20]**, an object segmentation VLM with text-based reasoning capabilities. (ii) **LOCATE [22]** and (iii) **AffordanceLLM [37]**, which

ActAffordance Benchmark															
Model	EPIC-KITCHENS					EGO4D					Combined				
	IoU ↑	Precision ↑	HD ↓	Dir. HD ↓	mAP ↑	IoU ↑	Precision ↑	HD ↓	Dir. HD ↓	mAP ↑	IoU ↑	Precision ↑	HD ↓	Dir. HD ↓	mAP ↑
LISA [20]	0.048	0.056	298	260	0.053	0.038	0.098	336	257	0.084	0.044	0.050	303	255	0.047
LOCATE [22]	0.010	0.014	274	261	0.007	-	-	-	-	-	-	-	-	-	-
AffLLM [37]	0.010	0.010	267	205	0.010	0.015	0.016	**229**	**226**	0.014	0.012	0.013	287	225	0.012
2HAffCLIP	0.032	0.077	359	317	0.068	0.023	0.050	306	250	0.047	0.026	0.064	341	292	0.059
2HAff	**0.064**	**0.125**	**241**	**185**	**0.104**	**0.051**	**0.137**	292	227	**0.105**	**0.058**	**0.130**	**262**	**202**	**0.104**
AffExtract	0.136	0.334	199	169	-	0.253	0.541	163	121	-	0.185	0.420	184	145	-

ActAffordance – Cropped Benchmark															
Model	EPIC-KITCHENS					EGO4D					Combined				
	IoU ↑	Precision ↑	HD ↓	Dir. HD ↓	mAP ↑	IoU ↑	Precision ↑	HD ↓	Dir. HD ↓	mAP ↑	IoU ↑	Precision ↑	HD ↓	Dir. HD ↓	mAP ↑
LISA [20]	**0.082**	0.115	177	111	0.110	0.097	0.132	205	134	0.125	0.082	0.122	196	130	0.116
LOCATE [22]	0.026	0.097	169	132	0.054	-	-	-	-	-	-	-	-	-	-
AffLLM [37]	0.066	0.092	**155**	**82**	0.088	0.091	0.139	**155**	**66**	0.124	0.076	0.112	**155**	**76**	0.103
VRB [1]	0.020	0.091	161	152	-	0.018	0.083	175	160	-	0.019	0.088	167	155	-
3DOI [35]	0.038	**0.227**	337	289	0.188	0.071	0.221	182	110	0.168	0.082	0.224	168	109	0.180
2HAffCLIP	0.038	0.144	170	108	0.131	0.040	0.202	176	98	0.186	0.039	0.168	172	104	0.154
2HAff	0.074	0.223	188	114	**0.204**	**0.101**	**0.331**	169	80	**0.291**	**0.086**	**0.269**	180	100	**0.240**

TABLE II: Comparison of our models and baseline methods on the ActAffordance Benchmark (top) and the modified version ActAffordance-Cropped (bottom) where images are cropped to a bounding-box around the target objects. Performance is evaluated separately on the EPIC-KITCHENS and EGO4D splits, as well as on the combined benchmark. The reported metrics include IoU (Intersection over Union), Precision, HD (Hausdorff Distance), Dir. HD (Directional Hausdorff Distance), and mAP (Mean Average Precision). For mAP, we average over five different thresholds, and the values for the other metrics correspond to the highest scores obtained across these thresholds. We also run our affordance extraction method, AffExtract, on the activity sequences in the benchmark as a measure of data quality and alignment with the benchmark annotations.

are trained on explicit affordance labels from the AGD20K dataset [28]. (iv) **3DOI [35]**, a fully-supervised method using point-based affordance data from exo and egocentric images and uses query points during inference. (v) **VRB [1]**, which uses bounding boxes to predict affordance hotspots.

All models were evaluated on the ActAffordance benchmark. Additionally, we assessed the methods on a modified version of the benchmark, where all images were cropped to encompass the target objects to enhance comparability with baselines that utilize bounding boxes [1] or query points [35] as prompts instead of language. Since LOCATE [22] uses an explicit affordance class label as input for prediction, we adapt the EPIC-VISOR verb categories used in 2HANDS to fit the AGD20K classes used in LOCATE. Such an adaptation is not possible for Ego4D so we exclude LOCATE from the comparison on the Ego4D split.

Figure 5 shows some qualitative affordance prediction results and Table II shows the quantitative results. On the combined ActAffordance benchmark, 2HandedAfforder achieves the best results across all metrics. LISA is the next best method since it accurately segments the correct object in the scene, resulting in a natural overlap with the ground truth. This demonstrates the power of reasoning segmentation for the challenging task of prompt-based affordance prediction. This reasoning ability is also validated by the 2HandedAfforder-CLIP version being only third-best. Though our models were not trained on Ego4D data, the performance on Ego4D is reasonable and better than the EPIC-KITCHENS split. The IoU scores are low across the board for all methods, indicating further room for improvement on this challenging task.

The results on the cropped version of the benchmark, Table II (lower), show similar results with performance improvements across the board since the uncropped benchmark is more difficult. In this setting, the other baseline models that use prompts or query points as input can be compared as well. 2HandedAfforder again achieves the best performance on the combined benchmark, with significantly better precision and mAP scores than the uncropped benchmark. 3DOI also performs reasonably in terms of precision. Surprisingly, AffordanceLLM achieves good scores in HD and Dir. HD, even though the IoU scores are lower. This stems from the fact that AffordanceLLM is relatively more optimistic and always predicts some affordance regions. The other methods can sometimes not detect any affordance regions and have no mask predictions, which penalizes the HD and dir. HD significantly. LISA is still the third or fourth best method on most metrics, while VRB, being a task-agnostic method, performs poorly.

C. In-the-wild Affordances and Robot Demonstration

We conduct robotic manipulation experiments with various objects using a bimanual Tiago++ robot in a realistic kitchen environment.We deployed our best-performing 2HandedAfforder model for affordance region segmentation inference based on user-defined task prompts such as 'pour into cup'.

To enhance the model's performance for real-world application, we obtain object bounding boxes and masks using a prompt-based segmentation method, LangSAM [29]. We then performed inference on the cropped object images. Moreover, to enhance the stability of our predictions, we only considered the intersection between our inferred affordance masks and the object masks generated by LangSAM. This also allowed us to adjust the prediction threshold to be more optimistic and generate larger affordance masks.

We demonstrate how our affordance prediction method improves the performance of a robot in executing manipulation tasks compared to using standard object or part segmentation approaches, such as the mask output of LangSAM. By inte-

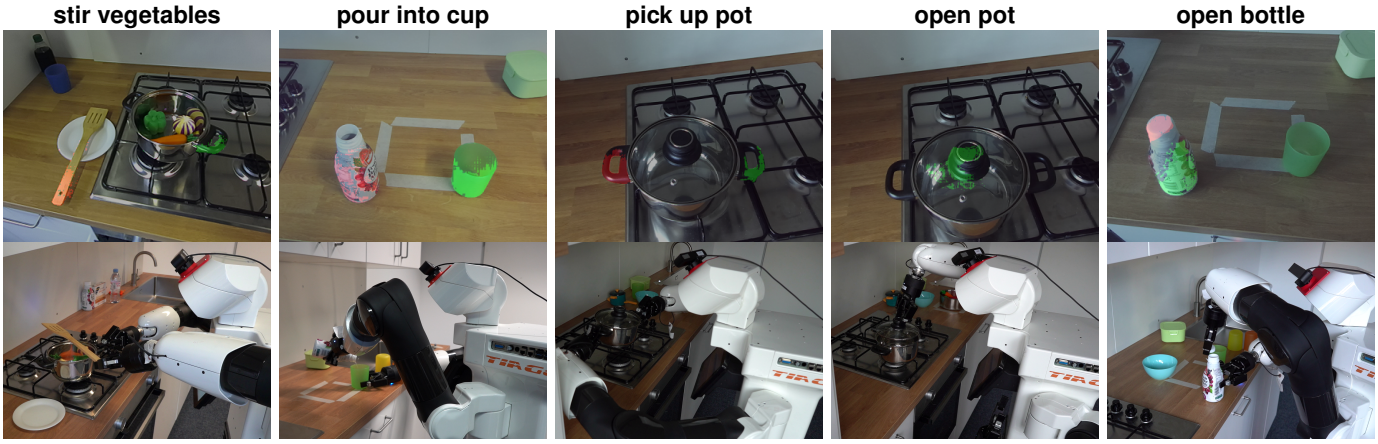

Fig. 6: Examples of different manipulation tasks executed on a bimanual Tiago++ robot. Red and green masks denote left and right hand affordance mask predictions, respectively. We segment the task-specific object affordance regions, propose grasps for these regions, and use pre-designed motion primitives to execute manipulation tasks.

grating our affordance prediction into the grasping pipeline, the robot is able to make more informed grasping decisions, leading to greater task success. Examples of different manipulation tasks are shown in Figure 6.

VI. Conclusion

In this work, we proposed a framework for extracting precise, meaningful affordance regions from human activity videos, resulting in the 2HANDS dataset of actionable bimanual affordances. We further introduced a novel VLM-based task-aware bimanual affordance prediction model, 2HandedAfforder, that predicts actionable affordance regions from task-related text prompts. To evaluate the alignment of the extracted affordances with human-annotated ones, we further proposed a novel ActAffordance benchmark, which is a particularly challenging benchmark for prompt-based segmentation of precise object affordance regions. Our experiments demonstrate that 2HandedAfforder can predict meaningful task-oriented bimanual affordances compared to other works, thereby showcasing the effectiveness of our data extraction pipeline and proposed model. Our real-world robotic manipulation experiments further showcased the efficacy of the predicted affordances compared to other object or part segmentation methods. In the future, we plan to extend our method to other egocentric video datasets and to other tasks beyond the kitchen context. We also plan to use activity recognition methods to automatically extract narrations for the videos.

References

[1] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790, 2023.

[2] Matthew Chang, Aditya Prakash, and Saurabh Gupta. Look ma, no hands! agent-environment factorization of egocentric videos. *Advances in Neural Information Processing Systems*, 36, 2024.

[3] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018.

[4] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022.

[5] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. URL https://doi.org/10.1007/s11263-021-01531-2.

[6] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022.

[7] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, and Huong Ngo et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models, 2024. URL https://arxiv.org/abs/2409.17146.

[8] James J Gibson. The theory of affordances:(1979). In *The people, place, and space reader*, pages 56–60. Routledge, 2014.

[9] Gal Gorjup, Anany Dwivedi, Nathan Elangovan, and Minas Liarokapis. An intuitive, affordances oriented telemanipulation framework for a dual robot arm hand system: On the execution of bimanual tasks. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3611–3616. IEEE, 2019.

[10] Mohit Goyal, Sahil Modi, Rishabh Goyal, and Saurabh Gupta. Human hands as probes for interactive object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3293–3303, 2022.

[11] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[12] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024.

[13] Andrew Guo, Bowen Wen, Jianhe Yuan, Jonathan Tremblay, Stephen Tyree, Jeffrey Smith, and Stan Birchfield. Handal: A dataset of real-world manipulable object categories with pose annotations, affordances, and reconstructions. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11428–11435. IEEE, 2023.

[14] Denis Hadjivelichkov, Sicelukwanda Zwane, Lourdes Agapito, Marc Peter Deisenroth, and Dimitrios Kanoulas. One-shot transfer of affordance regions? affcorrs! In *Conference on Robot Learning*, pages 550–560. PMLR, 2023.

[15] Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts. In *European Conference on Computer Vision*, pages 128–145. Springer, 2022.

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[17] Amlan Kar, Seung Wook Kim, Marko Boben, Jun Gao, Tianxing Li, Huan Ling, Zian Wang, and Sanja Fidler. Toronto annotation suite. https://aidemos.cs.toronto.edu/toras, 2021.

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[19] Franziska Krebs and Tamim Asfour. A bimanual manipulation taxonomy. *IEEE Robotics and Automation Letters*, 7(4):11031–11038, 2022.

[20] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024.

[21] Jaewook Lee, Andrew D. Tjahjadi, Jiho Kim, Junpu Yu, Minji Park, Jiawen Zhang, Yang Li, Sieun Kim, XunMei Liu, Jon E. Froehlich, Yapeng Tian, and Yuhang Zhao. Cookar: Affordance augmentations in wearable ar to support kitchen tool interactions for people with low vision. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, UIST '24, 2024.

[22] Gen Li, Varun Jampani, Deqing Sun, and Laura Sevilla-Lara. Locate: Localize and transfer object parts for weakly supervised affordance grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10922–10931, 2023.

[23] Gen Li, Deqing Sun, Laura Sevilla-Lara, and Varun Jampani. One-shot open affordance learning with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3086–3096, 2024.

[24] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7061–7070, 2023.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

[26] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15305–15314, 2023.

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[28] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2252–2261, 2022.

[29] Luca Medeiros. Lang-segment-anything, 2024. URL https://github.com/luca-medeiros/lang-segment-anything.

[30] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[31] Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis. Object-based affordances detection with convolutional neural networks and dense con-

ditional random fields. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5908–5915. IEEE, 2017.

[32] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, and et al. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

[33] Björn S Plonka, Christian Dreher, Andre Meixner, Rainer Kartmann, and Tamim Asfour. Learning spatial bimanual action models based on affordance regions and human demonstrations. In *2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids)*, pages 234–241. IEEE, 2024.

[34] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild, 2024.

[35] Shengyi Qian and David F Fouhey. Understanding 3d object interaction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21753–21763, 2023.

[36] Shengyi Qian, Linyi Jin, Chris Rockwell, Siyi Chen, and David F Fouhey. Understanding 3d object articulation in internet videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1599–1609, 2022.

[37] Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. Affordancellm: Grounding affordance from vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7587–7597, 2024.

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[39] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[40] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2980–2988, 2017.

[41] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[42] Martí Sánchez-Fibla, Sébastien Forestier, Clément Moulin-Frier, Jordi-Ysard Puigbo, and Paul FMJ Verschure. From motor to visually guided bimanual affordance learning. *Adaptive Behavior*, 28(2):63–78, 2020.

[43] Andranik Sargsyan, Shant Navasardyan, Xingqian Xu, and Humphrey Shi. Mi-gan: A simple baseline for image inpainting on mobile devices. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7335–7345, 2023.

[44] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020.

[45] Peize Sun, Shoufa Chen, Chenchen Zhu, Fanyi Xiao, Ping Luo, Saining Xie, and Zhicheng Yan. Going denser with open-vocabulary part segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15453–15465, 2023.

[46] Wei Zhai, Hongchen Luo, Jing Zhang, Yang Cao, and Dacheng Tao. One-shot object affordance detection in the wild. *International Journal of Computer Vision*, 130 (10):2472–2500, 2022.

[47] Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *European Conference on Computer Vision*, pages 127–145. Springer, 2022.

APPENDIX A
ADDITIONAL FILTERING AND AUGMENTATION STEPS

Between each of the major steps of the affordance extraction pipeline different filtering steps were applied to clean up the data. After calculating the intersection of the completed mask and the hand mask erosion and dilation steps were applied to remove scattered mask pixels and fill gaps to leave only one connected affordance mask. Furthermore, inconsistencies and inaccuracies within the data were detected and the data points were deleted, e.g. the calculated affordance masks are empty, the action is classified as bimanual but only one affordance mask is provided. Lastly, we remove datapoints with narrations that are too vague or do not describe an affordance by blacklisting some expressions, e.g. 'throw **something** into the bin' or '**look** at pan'. We augment the data by flipping all of the actions horizontally, essentially doubling the size of the dataset. By doing that we even out the ratio of left-handed and right-handed actions. Afterwards, we apply common augmenting strategies also used by Goyal et al. [10], i.e., color jittering (randomly changing the brightness, contrast, saturation and hue of the inpainted frame) and cropping.

APPENDIX B
2HANDS DATASET

Each data point of 2HANDS consists of an inpainted frame, two affordance masks where one of them is left empty if it is a unimanual action and the narration. We also provide additional information such as the object masks and object names if needed. In the end, the proposed dataset 2HANDS consists of over 278k datapoints from 25 different kitchen environments from the EPIC-KITCHENS dataset. An overview of the dataset can be found in Table 4.1. This dataset was used to train the models.

	Amount
Left Handed	76,278
Right Handed	76,278
Symmetric	51,684
Asymmetric	73,920
Total	278,160
No. Kitchen Environments	25
No. Videos	47
No. Object Classes	160
No. Verb Classes	73

TABLE III: Overview of the dataset

We collected affordance masks for 160 different object categories and 73 verb class.

The object classes:
alarm, almond, aubergine, bag, banana, basil, bean:green, beer, bin, board:chopping, book, bottle, bowl, box, bread, broccoli, brush, butter, button, can, candle, cap, caper, carrot, chair, cheese, cherry, chicken, chilli, choi:pak, chopstick, cinnamon, cloth, clothes, coffee, colander, container, cooker:slow, cork, corn, cover, cucumber, cumin, cup, cupboard, cutlery, cutter:pizza, dishwasher, dough, drawer, fan:extractor, filter, fish, flour, food, fork, fridge, garlic, ginger, glass, glove, grater, hand, heat, heater, hob, holder, ice, jar, jug, juicer, kettle, knife, knob, label, ladle, leaf, leek, lemon, lettuce, lid, light, lighter, liquid:washing, machine:sous:vide, machine:washing, maker:coffee, mat, meat, microwave, milk, mixture, mushroom, napkin, noodle, nut, oil, onion, opener:bottle, oven, package, pan, pan:dust, paper, paste, peach, peeler:potato, pepper, phone, pin:rolling, pith, pizza, plate, plug, pork, pot, potato, powder:washing, processor:food, rack:drying, rest, rice, roll, rubbish, salt, sauce, sausage, scissors, shell:egg, sink, skin, soda, spatula, sponge, spoon, sprout, stalk, stock, syrup, tap, toaster, tofu, tomato, tongs, top, towel, towel:kitchen, tray, utensil, vegetable, vinegar, wall, water, whetstone, window, wine, wire, wrap, wrap:plastic, yoghurt

The verb classes:
add, adjust, apply, attach, break, brush, carry, check, choose, close, coat, cook, crush, cut, divide, drink, dry, empty, fill, filter, flatten, flip, form, gather, hold, increase, insert, knead, lift, lower, mix, move, open, pat, peel, pour, press, pull, put, remove, rip, roll, rub, scoop, scrape, screw, scrub, season, serve, set, shake, sharpen, slide, soak, sort, spray, sprinkle, squeeze, stab, stretch, take, throw, turn, turn-down, turn-off, turn-on, uncover, unroll, unscrew, unwrap, use, wash, wrap

Fig. 7: Failure cases. In both cases the model is undecisive of what to do. In the top example it predicts affordance regions at different objects that are also not related to the task, i.e., the spatula and the knife. In the bottom example the model predicts a bimanual action to pick up the bowl and predicts affordance regions at multiple bowls even though only one bowl is supposed to be picked up.

APPENDIX C
ACTAFFORDANCE ANNOTATION PROCEDURE

For annotating the images for the ActAffordance Benchmark, we used TORAS [17]. We asked 10 human annotators to highlight all possible interaction regions of the target objects in the image where the hands were already removed with respect to the underlying task, i.e. the narration. This annotation was done for both the left and right hands. Additionally, annotators also had access to the original image to see how the hands interacted with the objects in the scene.

APPENDIX D
REAL ROBOT EXPERIMENTS

A successful example for an affordance prediction as well as the corresponding masks from LangSAM are visualized in Figure 8.

Fig. 8: The affordance detection of our method detects precise affordance regions (left) for the pot and the spatula that can be used to successfully perform the task of stirring within the pot. The right image shows the mask outputs by LangSAM. The text prompts used for this prediction were “wooden spatula” and “cooking pot” for LangSAM and “stir vegetables” for 2HandedAfforder.

Figure 9 shows the robot performing the task of ’stirring vegetables’. The first example illustrates an unsuccessful attempt where the robot, relying on plain object segmentation, attempts to stir within the pot while holding the spatula too close to the middle. This suboptimal grip prevents the robot from reaching into the pot, making it impossible for it to complete the task of stirring the vegetables properly. The second example demonstrates an improvement, as the robot uses our affordance detection method to identify a better grasping region. However, it employs only one arm instead of two, leading to an unintended side effect since the pot is not stabilized. The stirring motion causes it to move, making the task more difficult.

In the final and most successful example, the robot fully utilizes both affordance regions detected by 2HandedAfforder. Here, the left end effector grasps the spatula closer to its edge while the right end effector holds the pot securely in place. This configuration enables a stable and effective stirring motion, demonstrating the advantages of incorporating our affordance predictions in bimanual robotic manipulation tasks.

Fig. 9: Demonstration how different affordance detections determine the success in performing a specific task.

APPENDIX E
ADDITIONAL QUALITATIVE RESULTS

Additional qualitative results showing the performance of our proposed model compared to different baselines and the ground truth are provided in Figure 10.

APPENDIX F
QUALITATIVE ANALYSIS OF MASK COMPLETION APPROACHES

We developed and evaluated two mask completion approaches, i.e., an image reconstruction (IR) based and video segmentation (VS) based approach. The IR based approach uses the image inpainting model MI-GAN [43] to inpaint the missing regions of the mask using the hand mask as inpainting region. The VS based approach creates an image sequence out of the original image and the image with the hands removed. The object mask for the original image is

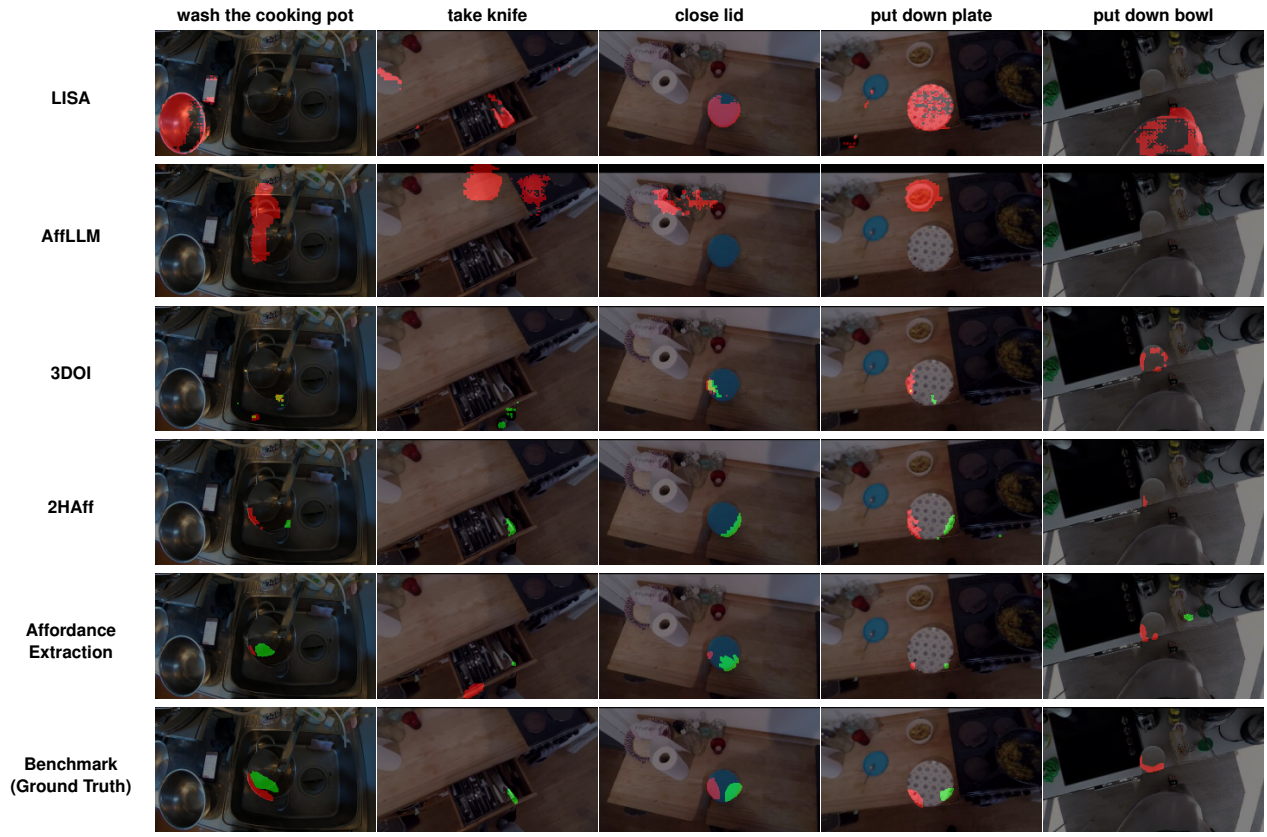

Fig. 10: Additional qualitative results showing the performance of our proposed model compared to different baselines and the ground truth.

then propagated to the inpainted image using SAM2 [39] to create the completed version of the mask. The evaluation of these approaches was conducted qualitatively, and some examples can be seen in Fig. 9. It is clearly visible that the VS based approach performs better on average than the IR based approach. Generally, the VS based approach provides more accurate results, see row 1 and 2, and it does not detect affordances at regions where the object was not reconstructed properly (row 3 and 4). This can be explained intuitively for two reasons: The IR based approach has no information about the underlying RGB image and only processes the object mask itself which is binary by nature. So the image reconstruction model focuses on simple principles such as the continuation of lines and shapes. This leads to the reconstruction of what we call 'ghost handles' where the image reconstruction model still predicts the existence of a handle or object part in general even though the object was not reconstructed successfully. This also reduces the accuracy of the IR based method. The VS based approach, however, has the information about the inpainted image and thus will only complete object parts that are properly inpainted. So, it naturally filters out data points where the object was not inpainted properly since the completed mask will not intersect the hand mask. Thus, it will always be more accurate and never predict 'ghost handles'. There are only a couple of examples where the IR based approach performs better than the VS based approach (row 5). Hence, we decided to use the VS based approach for the creation of 2HANDS.

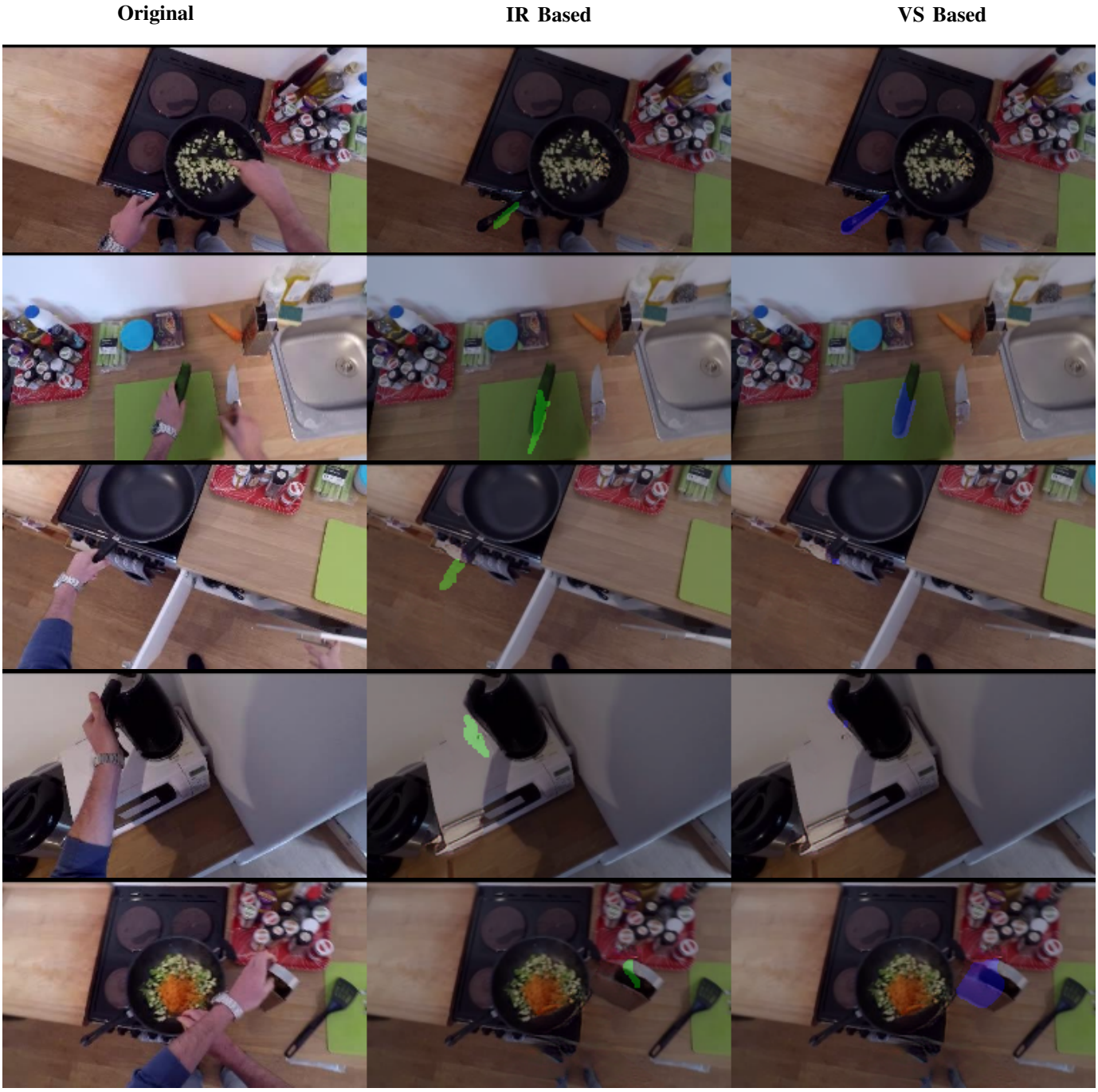

Fig. 11: Examples of the two affordance extraction methods. The left column shows the original image, the center column shows the image reconstruction based approach and the right column shows the video segmentation based approach. The video segmentation based approach outperforms the image reconstruction based approach qualitatively in almost every instance.