2HandedAfforder: Learning Precise Actionable Bimanual Affordances from Human Videos

Anonymous Author(s)
Affiliation
Address
email

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.

Keywords: Affordance Extraction, Affordance Grounding, Egocentric Vision

1 Introduction

When humans perceive objects, they understand different object regions and can predict which object region *affords* which activities [1], 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 [5, 6, 7]. However, there is still a big gap in their ability to detect accurate object affordance regions in images [8]. Moreover, most existing state-of-the-art affordance detection methods [9, 10, 11, 12, 13] use labeled data [14, 10, 15, 16, 17] 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,

Submitted to the 9th Conference on Robot Learning (CoRL 2025). Do not distribute.

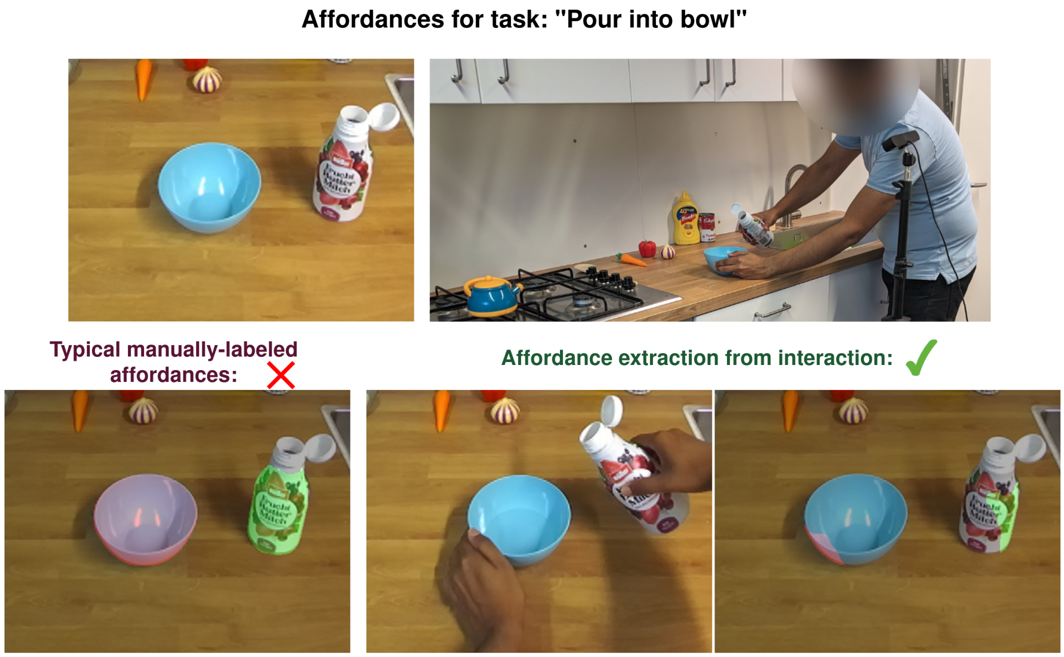

Figure 1: **A motivating example:** When labeling affordances for a task 'Pour into bowl', typical labeled affordances provided by annotators are not precise and reduce the problem to object part segmentation. Alternatively, our affordance extraction method uses the hand-object interaction sequence to get precise bimanual affordance regions that are not just 'conceptual' but also 'actionable'.

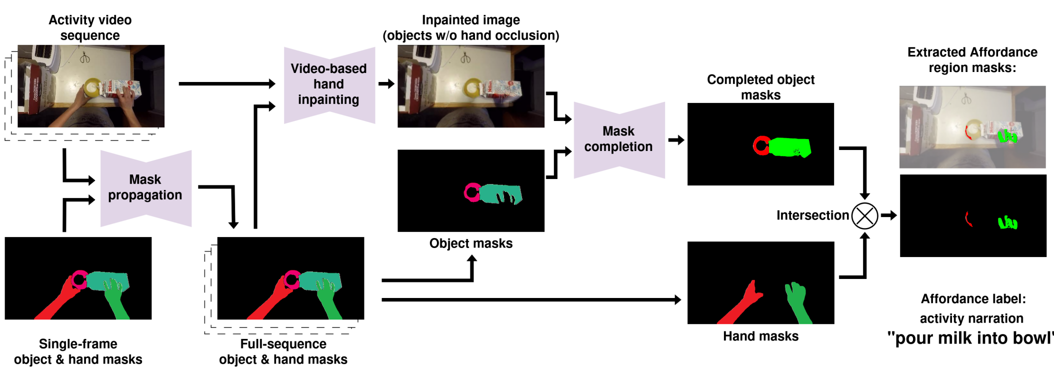

Figure 2: Affordance extraction pipeline. Given a human activity video sequence and a single-frame object and hand masks, we first obtain dense, full-sequence object and hand masks using a video mask-propagation network [2]. We then inpaint out the hands in the RGB images using a video-based hand inpainting model [3]. 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 [4] 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.

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.

We argue that affordances should not be labeled but automatically extracted by observing humans performing tasks, e.g. in activity video datasets. We propose a method that uses hand-inpainting and mask completion to extract affordance regions occluded by human hands. This 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 [18, 19].

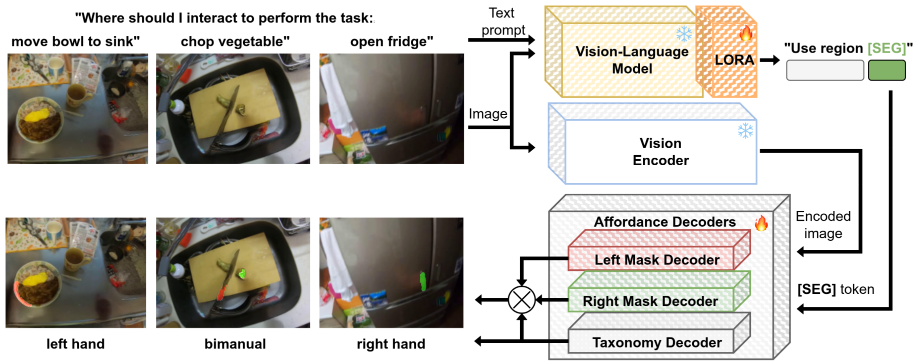

Figure 3: Affordance prediction network. Given an input image and task, we use a question asking where the objects should be interacted for the desired task as a text prompt to a Vision-Language model (VLM). The VLM produces language tokens and a [SEG] token which is passed to the affordance decoders. We also use a SAM [20] 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 [21].

2 Related Work

Fully supervised affordance detection. In fully supervised affordance detection datasets and methods such as by Qian and Fouhey [10], AffordanceLLM [11], the dataset is fixed and hand-annotated such as from IIT-AFF [14] and 3DOI [10]. The affordance classes in these datasets are explicit and annotators guess which affordance class may apply to object regions. Other methods, such as VLPart [12], use a general open vocabulary segmentation pipeline. LISA [13] performs open-vocabulary, prompt-based "reasoning segmentation". However, these methods do not consider actions and typically segment either the whole object [13] or object parts[12], and not precise affordance regions.

Weakly supervised affordance detection. Weakly supervised methods such as Cross-viewAG [15] and Locate [22] learn to predict affordances by observing exocentric images of humans interacting with objects based on the AGD20K dataset [15]. 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 [23, 24]. 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 [25, 26, 27, 28, 29] 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. [18] and Bahl et al. [19] have shown that videos from datasets such as EPIC kitchens [25] and Ego4D [27] 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 [3]. 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 [10] or that only predict heatmaps [15, 19, 30]. 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 [31, 32, 33, 34, 35, 36], 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.

3 Extraction and Learning of Bimanual Affordances from Human Videos

3.1 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. We propose a pipeline to extract affordances that leverages recent advances in hand inpainting [3] and object mask completion [37, 4], providing the first bimanual affordance region segmentation dataset. Moreover, we use the narration of the task being performed as the affordance text label, obtaining a diverse set of affordance classes for various objects.

We use videos from EPIC-KITCHENS [25], containing ∼100 hours of egocentric human videos in kitchens. We use VISOR [26] annotations of the dataset, which contain sparse hand-object mask segmentations and binary labels for whether the hand is in contact with the object.To obtain dense hand-object masks for entire video sequences, we use a video-based mask propagation network [2].

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 [3], that uses 4 frames from the sequence as input to inpaint the missing regions. This sequence-based inpainting better reconstructs the target objects since the objects may be visible in another frame of the 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 [4] to propagate these masks to the new inpainted image.To obtain the final affordance region where the hand and object interact, we can simply compute the intersection of the un-occluded object masks and the hand masks (Fig. 2). For bimanual affordances, we also classify into a bimanual taxonomy [31] of unimanual left, unimanual right, and bimanual actions.

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

3.2 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 [38, 13] and train a VLM-based segmentation model to reason about the required task and predict the relevant affordance region in the input image.

Inspired by reasoning segmentation methods such as by Lai et al. [13], we use a Vision-Language Model (VLM) [39], a LLaVa-13b, 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. Thus, to have a stronger vision-backbone for our segmentation-related task, we use a modified version of SAM [20]. 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

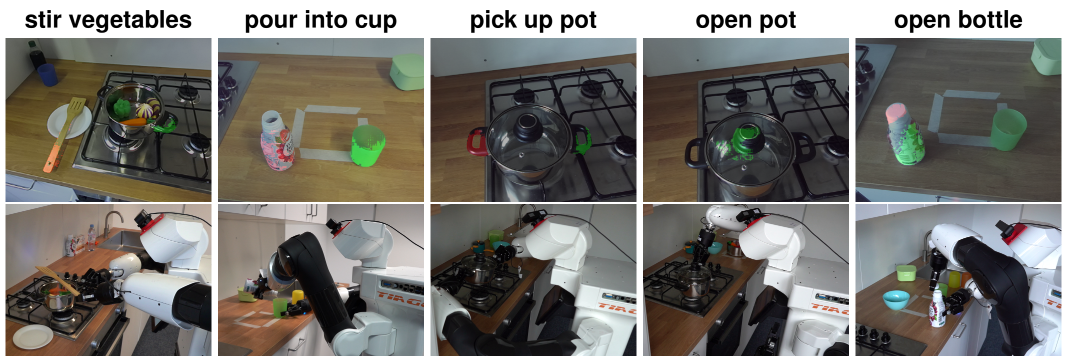

Figure 4: Examples of different manipulation tasks executed on a bimanual Tiago++ robot. Red and green masks denote left and right hand affordance mask predictions, respectively. We segment the task-specific object affordance regions, propose grasps for these regions, and use pre-designed motion primitives to execute manipulation tasks.

mask decoders, generating 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 full-connected classifier decoder. The full architecture is visualized in Figure 3.

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 [13]	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 [11]	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	-

Table 1: Comparison of our models and baseline methods on the ActAffordance benchmark. 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, as a measure of data quality and alignment with the benchmark annotations.

The VLM is trained to generate a specific output token: a segmentation [SEG] token. Specifically, inspired by LISA [13], 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 [10], 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.

4 Experiments

4.1 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. 3.1, when humans label affordances, they often simply label object parts and do not necessarily focus on the precise regions of interaction of the objects [15, 10]. 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 own extracted affordances.

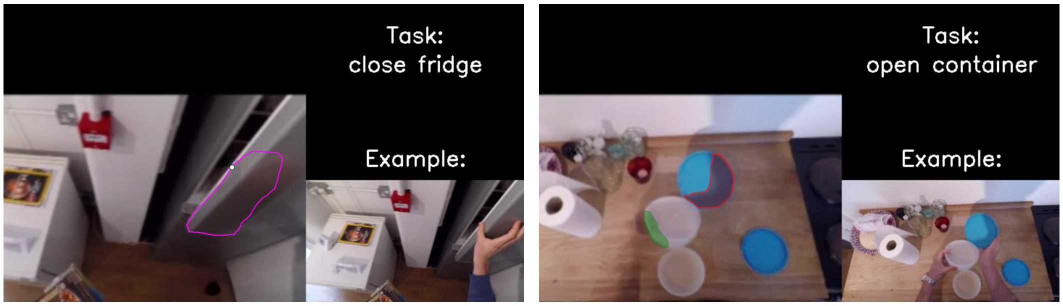

Figure 5: Example annotations for the ActAffordance benchmark. Left: The image to be annotated with the highlighted annotation mask(s). Right: the example interaction provided to the human annotator, along with the task description. The human is asked to annotate ALL the possible regions for the interaction to capture all the different modes.

Therefore, we propose a novel benchmark called "ActAffordance" to evaluate both the dataset quality and the predicted affordances. Specifically, we evaluate the alignment of our affordances with the affordances annotated by humans who are shown the *full interaction video sequence*. 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 contains unimanual and bimanual segmentation masks for 400 activities from EPIC-KITCHENS [25] and Ego4D [27], with no overlap with the data used in 2HANDS.

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. 5 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 contains unimanual and bimanual segmentation masks for 400 activities from EPIC-KITCHENS [25] and Ego4D [27], with no overlap between EPIC-KITCHENS data used in 2HANDS. Details about the benchmark and annotation process are in Appendix Sec. 10.

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 [19] or query points on the object [10] as input.

5 Results

Extraction quality & Benchmark performance:

Table 1 shows the quantitative results, some examples of which are shown in Fig. 6. The high precision of AffExtract shows a reasonably good alignment with the human-annotated segmentations from the benchmark and meaningful affordance region extraction using our extraction pipeline. 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.

Since ours is the first method to perform bimanual affordance mask detection using text prompts, we adapt baselines such as AffordanceLLM [11] and a SOTA text-based reasoning segmentation baseline, LISA [13]. To isolate the effect of the 2HANDS dataset, the comparison with AffLLM and LISA is key since their network architecture is close to ours. Among the trained prediction models, 2HandedAfforder achieves the best results across all metrics. LISA is the next best method

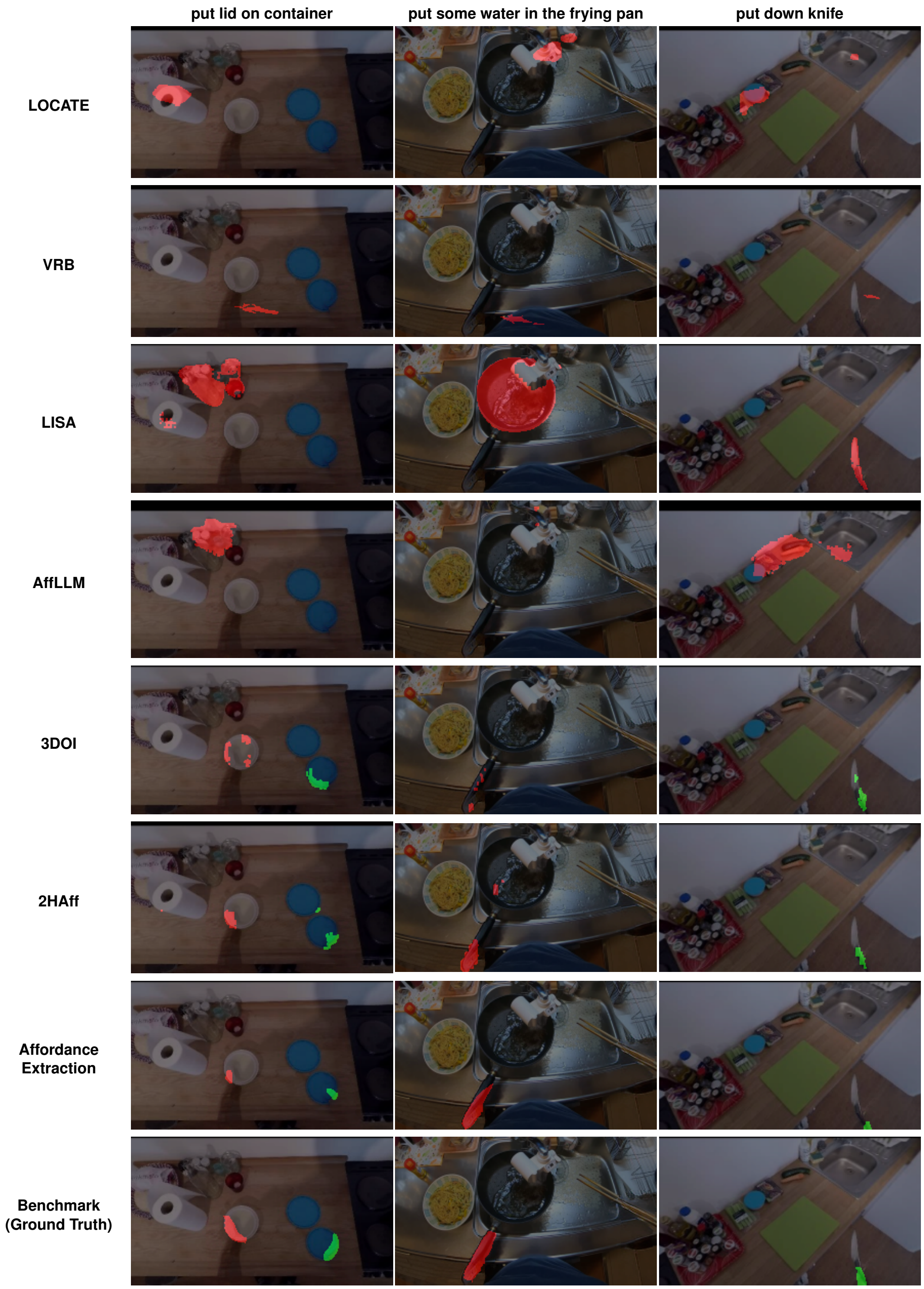

Figure 6: Qualitative affordance prediction results on the ActAffordance benchmark. We compare our 2HandedAfforder model against LOCATE [22], VRB [19], LISA [13], AffordanceLLM [11], 3DOI [10]. We also include an example result if we run our affordance extraction method on the activity sequence to show the quality of the extraction. Red and green masks denote left and right hand affordance mask predictions, respectively.

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. Although our models were not trained on any Ego4D data, their performance on Ego4D is still reasonable and often 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.

Real-world Affordance Prediction on a Robot:

We conduct robotic manipulation experiments with various objects using a bimanual Tiago++ robot in a realistic kitchen environment. We deploy our 2HandedAfforder model for affordance region segmentation inference based on task prompts such as 'pour into cup'. By integrating 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 4.

6 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.

References

[1] J. J. Gibson. The theory of affordances:(1979). In *The people, place, and space reader*, pages 56–60. Routledge, 2014. 1

[2] H. K. Cheng and A. 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. 2, 4

[3] M. Chang, A. Prakash, and S. Gupta. Look ma, no hands! agent-environment factorization of egocentric videos. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 4

[4] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 4, 3

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[6] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, and et al. Gpt-4 technical report, 2024. URL `https://arxiv.org/abs/2303.08774`. 1

[7] M. Deitke, C. Clark, S. Lee, R. Tripathi, Y. Yang, J. S. Park, M. Salehi, N. Muennighoff, K. Lo, L. Soldaini, J. Lu, T. Anderson, E. Bransom, K. Ehsani, and H. N. 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`. 1

[8] G. Li, D. Sun, L. Sevilla-Lara, and V. 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. 1

[9] A. Guo, B. Wen, J. Yuan, J. Tremblay, S. Tyree, J. Smith, and S. 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. 1

[10] S. Qian and D. 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. 1, 3, 5, 6, 7

[11] S. Qian, W. Chen, M. Bai, X. Zhou, Z. Tu, and L. E. 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. 1, 3, 5, 6, 7

[12] P. Sun, S. Chen, C. Zhu, F. Xiao, P. Luo, S. Xie, and Z. Yan. Going denser with open-vocabulary part segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15453–15465, 2023. 1, 3

[13] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. 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. 1, 3, 4, 5, 6, 7

[14] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5908–5915. IEEE, 2017. 1, 3

[15] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2252–2261, 2022. 1, 3, 5

[16] J. He, S. Yang, S. Yang, A. Kortylewski, X. Yuan, J.-N. Chen, S. Liu, C. Yang, Q. Yu, and A. Yuille. Partimagenet: A large, high-quality dataset of parts. In *European Conference on Computer Vision*, pages 128–145. Springer, 2022. 1

[17] J. Lee, A. D. Tjahjadi, J. Kim, J. Yu, M. Park, J. Zhang, Y. Li, S. Kim, X. Liu, J. E. Froehlich, Y. Tian, and Y. 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. 1

[18] M. Goyal, S. Modi, R. Goyal, and S. 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. 3, 1

[19] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. 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. 3, 6, 7

[20] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick. Segment anything. *arXiv:2304.02643*, 2023. 3, 4

[21] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3

[22] G. Li, V. Jampani, D. Sun, and L. 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. 3, 5, 7

[23] W. Zhai, H. Luo, J. Zhang, Y. Cao, and D. Tao. One-shot object affordance detection in the wild. *International Journal of Computer Vision*, 130(10):2472–2500, 2022. 3

[24] D. Hadjivelichkov, S. Zwane, L. Agapito, M. P. Deisenroth, and D. Kanoulas. One-shot transfer of affordance regions? affcorrs! In *Conference on Robot Learning*, pages 550–560. PMLR, 2023. 3

[25] D. Damen, H. Doughty, G. M. Farinella, , A. Furnari, J. Ma, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. 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`. 3, 4, 6

[26] A. Darkhalil, D. Shan, B. Zhu, J. Ma, A. Kar, R. Higgins, S. Fidler, D. Fouhey, and D. 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. 3, 4

[27] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. 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. 3, 6

[28] L. Zhang, S. Zhou, S. Stent, and J. Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *European Conference on Computer Vision*, pages 127–145. Springer, 2022. 3

[29] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, K. Ashutosh, V. Baiyya, S. Bansal, B. 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. 3

[30] M. K. S. et al. Hrp: Human affordances for robotic pre-training. *RSS*, 2024. 3

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

[32] G. Gorjup, A. Dwivedi, N. Elangovan, and M. 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. 4

[33] M. Sánchez-Fibla, S. Forestier, C. Moulin-Frier, J.-Y. Puigbo, and P. F. Verschure. From motor to visually guided bimanual affordance learning. *Adaptive Behavior*, 28(2):63–78, 2020. 4

[34] Y. Liu et al. Taco: Benchmarking generalizable bimanual tool-action-object understanding. *CVPR*, 2024. 4

[35] B. S. Plonka, C. Dreher, A. Meixner, R. Kartmann, and T. 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. 4

[36] R. Fu et al. Gigahands: A massive annotated dataset of bimanual hand activities. *CVPR*, 2025. 4

[37] A. Sargsyan, S. Navasardyan, X. Xu, and H. 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. 4, 3

[38] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. 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. 4

[39] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 4

[40] A. Kar, S. W. Kim, M. Boben, J. Gao, T. Li, H. Ling, Z. Wang, and S. Fidler. Toronto annotation suite. https://aidemos.cs.toronto.edu/toras, 2021. 3