OVAL-Prompt: Open-Vocabulary Affordance Localization for Robot Manipulation through LLM Affordance-Grounding

Edmond Tong[1], Anthony Opipari[1], Stanley Lewis[1], Zhen Zeng[2], and Odest Chadwicke Jenkins[1]

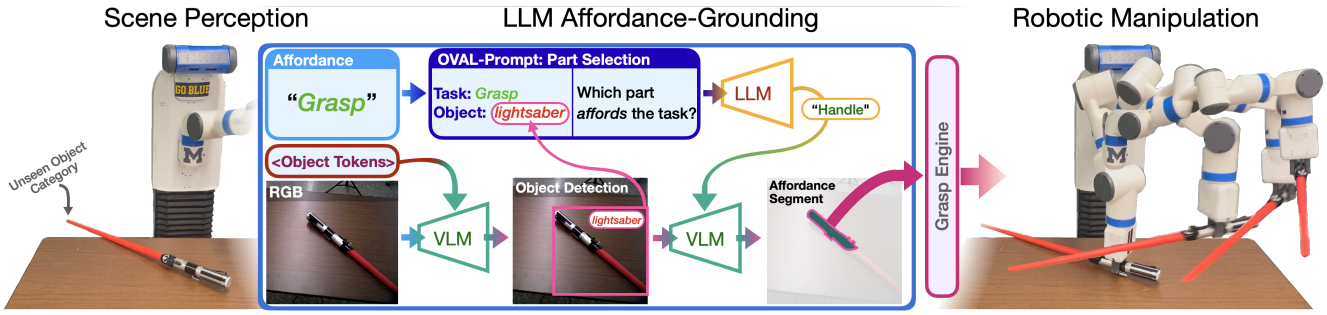

Fig. 1: Overview of Affordance-Prompting, a pipeline for open-vocabulary affordance segmentation and localization. Left: Given an RGB image and list of object tokens, a pre-trained Vision Language Model (VLM) provides object detections. Middle: The detected object is composed with a natural language affordance to form an OVAL-Prompt that is fed into a pre-trained Large Language Model (LLM) to identify which part of the detected object supports the task affordance. Right: The identified object part is fed back into the VLM to generate a part-affordance-segment that is used by a grasp engine (e.g. Dex-Net [1]) to enable robot manipulation.

***Abstract*— In order for robots to interact with objects effectively, they must understand the form and function of each object they encounter. Essentially, robots need to understand *which* actions each object affords, and *where* those affordances can be acted on. Robots are ultimately expected to operate in unstructured human environments, where the set of objects and affordances is not known to the robot before deployment (i.e. the open-vocabulary setting). In this work, we introduce OVAL-Prompt, a prompt-based approach for open-vocabulary affordance localization in RGB-D images. By leveraging a Vision Language Model (VLM) for open-vocabulary object part segmentation and a Large Language Model (LLM) to ground each part-segment-affordance, OVAL-Prompt demonstrates generalizability to novel object instances, categories, and affordances without domain-specific finetuning. Quantitative experiments demonstrate that without any finetuning, OVAL-Prompt achieves localization accuracy that is competitive with supervised baseline models. Moreover, qualitative experiments show that OVAL-Prompt enables affordance-based robot manipulation of open-vocabulary object instances and categories.**

I. Introduction

For robots to function effectively in unstructured settings like homes and offices, they must be adept at identifying objects in their surroundings and utilizing them appropriately. The potential uses for each object can be understood in terms of the object's "affordances." Originally termed by psychologist James J. Gibson [2], an object's affordance describes what actions are possible for a robot to perform using the given object – thereby enabling the robot to reason about effective object use. For example, knives afford cutting solids and spoons afford scooping liquids. Once the robot has detected which object affordances are necessary for task completion, it must then be able to localize the relevant object geometry that enables said affordances. For this work, we refer to the combined detection and localization task as 'affordance grounding.' Traditional approaches to affordance grounding have relied heavily on extensive labeled datasets and task-specific training, limiting their flexibility and scalability for robotic applications. Roboticists have increasingly sought to address this limitation by developing approaches for open-vocabulary settings, where the set of objects [3] and affordances [4] is not constrained to the same fixed set that was used during training. In particular, Large Language Models (LLMs) and Vision Language Models (VLMs) have been proposed as sources of knowledge to address the open-vocabulary setting [4, 5, 6].

There is a growing interest in using LLMs and VLMs for robotic applications [5, 7, 8, 9], with notable examples including CLIP [10, 11] and Say-Can [5]. Specifically, LLMs and VLMs have been proposed as knowledge bases that can be tapped into for solving classical robotics challenges. For example, within planning, Valmeekam et al. and Kambhampati et al. proposed using LLMs as a source of knowledge within 'LLM-Modulo' planning frameworks but find that LLMs struggle to plan in isolation [9, 12]. For affordance grounding and localization specifically, Qian et al. propose finetuning a LLM to generate affordance mask-tokens, which can be decoded into localized affordance segments [6]. Nguyen et al. propose finetuning a text-encoder to be used for open-vocabulary feature correlations within an affordance localization pipeline [4]. In contrast, we set out to understand the potential for pre-trained LLMs and VLMs to be used without domain-specific finetuning.

Our work is motivated by two simple questions: *Can*

[1]E. Tong, A. Opipari, S. Lewis, and O.C. Jenkins are with the Department of Robotics, University of Michigan, Ann Arbor, MI, USA, 48109.
[2]Z. Zeng is with J.P. Morgan AI Research.
*E. Tong is the corresponding author: ekjt@umich.edu

pre-trained LLMs and VLMs be used to ground object-part affordances without domain-specific finetuning and do they enable robots to manipulate open-vocabulary affordances? To answer these questions, we develop an LLM and VLM-based affordance-grounding pipeline to detect and localize part-based object affordances in images. By using pre-trained foundation models, the developed pipeline is applicable to the open-vocabulary setting where object instances, categories, and affordances are evaluated without having been trained for in a supervised fashion. The proposed pipeline is quantitatively evaluated through multiple experiments to measure its accuracy relative to state-of-the-art supervised methods. Furthermore, real robot experiments are performed using the affordance-grounding pipeline to establish whether the pipeline is applicable to real-world robotic scenarios.

In summary, this paper sets out to advance affordance-based robotic manipulation with the following contributions:

1) First, the OVAL-Prompt algorithm is introduced to perform open-vocabulary affordance localization. OVAL-Prompt uses a LLM to ground part-segments generated by a VLM with corresponding action-affordances.
2) Second, we explore the importance of prompt structure in affordance-prompting to ground visual-affordances with natural language tokens. Results from these experiments suggest using a LLM to translate affordance to object part and prompt it for similar names.
3) Third, we demonstrate that OVAL-Prompt, which uses no domain finetuning, is competitive with multiple finetuned state-of-the-art baseline models on a real-world affordance localization dataset.
4) Fourth, robot experiments are performed showing OVAL-Prompt can be successfully used on real robot platforms for successful affordance-based object and tool grasping.

II. Related Works

A. Closed-Vocabulary Affordance Localization

Early approaches to affordance grounding relied on hand-crafted features such as geometric shapes [13]. However, recent research has shifted towards deep learning techniques [14]. Traditional deep learning approaches use a convolutional backbone (e.g. VGG16 [15]) for feature extraction and use the resulting latent features to generate a final prediction [16, 17, 18, 19, 20, 21]. These approaches rely on supervised training with a fixed set of object affordances that do not change during inference, thereby limiting their applicability to robotic applications. In contrast, the present paper sets out to address affordance localization in the open-vocabulary setting where the object categories and affordances are not fixed after training.

B. Open-Vocabulary Affordance Localization

'Open-vocabulary' refers to machine learning tasks in which models are trained using a fixed set of category labels and evaluated using a non-fixed set of category labels [22]. In the context of affordance localization, the open-vocabulary setting has received increasing interest [3, 4, 23, 24]. Luo et al. introduce a large-scale image dataset for open-vocabulary affordance localization and develop a feature co-relation strategy to achieve open-vocabulary inference [3]. Li et al. propose a vision transformer and exocentric human demonstrations to identify affordances in egocentric images with weakly-supervised training [23]. Similarly, Li et al. propose a vision transformer trained with a one-shot approach in which only one labeled example per object category is used for training [24]. Nguyen et al. focus on affordance localization within 3D point clouds with a contributed dataset and use feature correlations along with a pre-trained text-encoder for the open-vocabulary setting [4]. In contrast, the present paper sets out to avoid expensive finetuning strategies and instead to explore the potential for pre-trained foundation models to be used for open-vocabulary affordance localization.

C. Foundation Models

Several Large Language Models (LLMs), such as GPT-4 [25] and Gemini [26], have demonstrated exceptional capabilities in tasks like test-taking, gaining significant traction in both research and practical applications. These advancements in LLMs, have paved the way for their widespread adoption across various domains. In the field of robotics, LLMs are increasingly being utilized for a broad spectrum of functions. These include enhancing human-robot interaction [5], planning [27], and navigation [28].

This work aims to answer the questions: Do LLMs understand affordances? While LLMs are known for their remarkable recall and comprehension abilities, they have also been criticized for concocting infeasible plans [29], and demonstrating flawed reasoning [30].

III. Affordance Localization

A. Problem Definition

We tackle zero-shot open-vocabulary affordance localization, by identifying actionable object parts from images and descriptions without prior specific training. We focus on the task of open-vocabulary affordance localization in images. Given a RGB image, $I \in \mathcal{R}^{H \times W \times 3}$, the goal is to detect and segment the affordance labels, $\mathcal{L} = \{l_1, l_2, \ldots, l_N\}$, present in the image. Crucially, for the open-vocabulary setting, the set of test labels can differ from the one used during model training.

B. Evaluation Metric

To evaluate zero-shot affordance localization, we use the weighted F1-score defined by 1 that has been used in previous affordance localization papers [13].

$$F_\beta^w = \frac{(1+\beta^2) \cdot P_r^w \cdot R_c^w}{\beta^2 \cdot P_r^w + R_c^w}, \text{ with } \beta = 1 \tag{1}$$

Where P_r^w and R_c^w are weighted versions of the standard precision $P_r = \frac{TP}{TP+FP}$ and recall $R_c = \frac{TP}{TP+FN}$ measures. Where TP = true positive, FP = false positive, FN = false negative. Precision indicates how many of the items identified as relevant are actually relevant, while recall measures how many relevant items were identified out of all relevant

items. For more information see [13]. In zero-shot evaluation, an affordance that is not predicted but exist in the ground truth sample is assigned an F-score of 0.

IV. OVAL-Prompt: Affordance-Prompting

Our method processes a RGB image and task description, using a VLM to detect and list objects within the image. This list informs a natural language prompt for a LLM, which specifies the object part related to the task. The VLM then segments this part, creating a mask for manipulation tasks. If segmentation fails, the LLM is reprompted for alternative part names, and the VLM tries again.

This approach is highly adaptable, using off-the-shelf components. The LLM and VLM are interchangeable, allowing for easy updates in the future, promising significant enhancements as LLMs and VLMs progress.

A. VLM

We employ the Vision-Language Model (VLM) for object detection and part segmentation. Initially, we supply the VLM with an image and a list of potential objects to identify, discarding detections below 50% confidence. This filtered list is then provided to the LLM where it is processed and returns an object part. Upon receiving an object part from the LLM, the VLM segments it and generates a binary mask.

We chose VLpart [31] for its part segmentation capability, specifically the "swinbase_cascade_lvis_paco_pascalpart_partimagenet_in" model, because of its high mAP score in their evaluation.

B. LLM

Current research highlights the significance of crafting precise prompts for Large Language Models (LLMs). Our method uses straightforward natural language prompts with organized outputs, engaging LLMs at three key stages: identifying relevant objects, pinpointing the object part related to the task, and seeking alternative part names.

First, we provide the LLM with a list of items and the task, from which it selects objects suited to the task.

Object Prompt

Your task is to <task>. Which of these objects can do this task: <list of objects>
Answer like the following:
Objects:
Reason:

Our prompts for the LLM include "Object" and "Reason" sections to specify the chosen object and its justification, enhancing response validity through a logical "chain of thought." Following object identification, the LLM is further queried for the task-related object part, using a similar prompt structure. If the part segmentation by the VLM fails, we seek alternative part names from the LLM, like "cup side" for "cup body." We use GPT-4 [25] API with a zero temperature setting for consistent outputs.

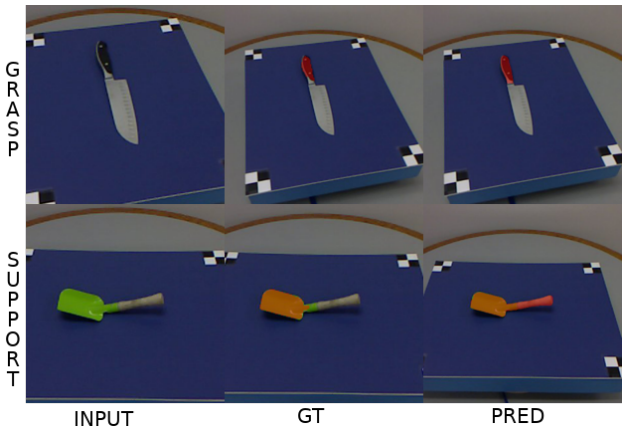

Fig. 3: Segmentation Failures

V. Experiments

A. Affordance Localization

In our experiments, we aim to determine the LLM's understanding of affordances by testing our pipeline on the UMD [13] dataset.

We selected the UMD [13] dataset for its extensive usage and the availability of numerous benchmarks for comparison. This dataset features RGB-D images with ground truth pixel-wise affordance labels for 105 items encompassing different viewpoints, 7 affordances, and 17 object classes. We use the weighted F-score as used in the original publication.

The UMD [13] dataset, includes the following objects [knife, saw, scissors, shears, scoop, spoon, trowel, bowl, cup, ladle, mug, pot, shovel, turner, hammer, mallet, tenderizer] and affordances [grasp, cut, scoop, contain, pound, support, wrap-grasp]. We did not finetune or train on this list, but we use the list of objects and affordances as inputs to the system. If a given object was not detected or an object part was not detected, a f-score of 0 was assigned to that case. We ran this method on the test set images and the results are in table I.

OVAL-Prompt achieves a F-score +0.154 higher than HMP on average but -0.144 lower than the highest-performing GSE model. Despite its sub state-of-the-art performance it achieves a level of performance that is functional for real-world application.

1) Failure mode: The segmentation from the VLM resulted in lower F-scores primarily due to misaligned ground truth and imprecise part isolation, as depicted in Figure 3. Ground truth often mismatched the actual object parts, and VLM occasionally segmented entire objects instead of a part. The system exhibited weaknesses at multiple stages:

- Class names: VLM struggled with class names like "turner" and "pot," causing confusion. Using more common names such as "spatula" instead of "turner" and "plant pot" for "pot" might improve its performance, as VLM better recognizes common names.
- Spatial relations: The LLM produced part descriptions involving spatial relationships (e.g. cup top), which were difficult for the VLM to process. To overcome

TABLE I: Performance on UMD Dataset

	HMP [13]	DeepLab [32]	AffordenceNet [18]	Attention-CNN [33]	RelaNet [34]	GSE [35]	OVAL-Prompt(Ours)
Grasp	0.367	0.620	0.731	**0.784**	0.737	0.779	0.650
W-Grasp	0.373	0.730	0.814	0.822	0.824	**0.840**	0.718
Cut	0.415	0.600	0.762	0.761	0.755	0.776	**0.823**
Contain	0.810	0.900	0.833	0.840	**0.941**	0.924	0.688
Support	0.643	0.600	0.821	0.844	0.866	**0.893**	0.538
Scoop	0.524	0.800	0.793	**0.862**	0.802	0.856	0.753
Pound	0.767	0.880	0.836	0.847	0.879	**0.918**	0.809
Average	0.557	0.733	0.799	0.823	0.829	**0.855**	0.711

this, the LLM was prompted once more for alternative names when the VLM failed to recognize the initially described part. This approach improved the VLM's detection accuracy when the reprompted part was a physical part (e.g. cup rim).

B. Robot Demonstration

To demonstrate the concept with an actual robot, we utilized a Fetch and Freight Research Edition robot to grasp various random objects that were placed on a table directly in front of it. We captured both RGB and depth images using the onboard cameras. The RGB image was then processed through the previously described pipeline, where we segmented the image area that affords grasping, resulting in a binary mask. This mask, along with the depth information, was fed into DexNet [1] to suggest a potential grasp. To simplify the grasping process, we modified the grasp suggested by DexNet [1] and converted it into a top-down pick-up. This was achieved by using the x, y, z coordinates, setting the gripper orientation to face downwards along the negative z-axis, and then calculating the z-axis rotation for the wrist rotation. The finalized grasp configuration was then executed by the robot using the MoveIt [36]. A image of the setup of shown in figure 4.

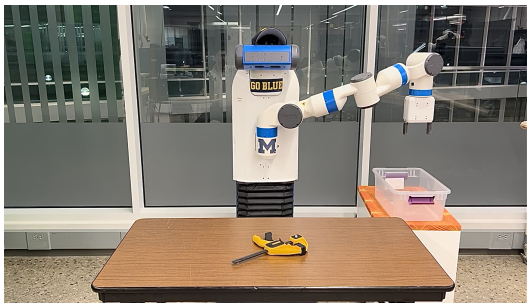

Fig. 4: Robot Experimental Setup
Video: https://youtu.be/fXvYo-0AJek

The experiment involved a spatula, walkie-talkie, toy lightsaber, mug, clamp, and brush, repeated ten times each. Table II shows successful segmentations where segments matched the correct object parts, the rate of feasible grasp proposals at suggested parts, and successful object pickups by the robot at these parts.

Failures are mostly caused by imprecise segmentation that include whole objects instead of specific parts. However, with accurate segmentation, grasp and pickup success rates approach 100%. Grasping failures occurred for the mug because its shapes was not considered in grasp planning.

TABLE II: Grasping Performance Metrics

Object	Suggested Part	Segmentation Success	Grasp Success	Pick Success
Brush	Handle	10/10	10/10	10/10
Clamp	Handle	6/10	6/6	5/6
Mug	Handle	10/10	10/10	7/10
Lightsaber	Handle	5/10	5/5	5/5
Spatula	Handle	7/10	7/7	6/7
Walkie Talkie	Body	4/10	4/4	4/4

C. Ablation

To validate our method, we removed components from our network to evaluate their impact on performance. We tested two variations: 1) using only the VLM with a list of affordances and 2) excluding the reprompting step.

TABLE III: Ablation on UMD dataset

	VLM only	No Reprompting	Full Network
grasp	0.002	0.498	0.650
cut	0.013	0.401	0.823
scoop	0.062	0.345	0.753
contain	0	0.437	0.688
pound	0	0.317	0.809
support	0	0.436	0.538
wrap-grasp	0	0.312	0.718
Average	0.011	0.392	0.711

Table III shows that a standalone VLM struggles with affordances, likely because VLMs are not typically trained on such tasks. The necessity of reprompting for alternative part names is also highlighted, emphasizing how the VLM's open vocabulary can respond variably to different words, showing that using synonyms can improve results.

VI. Conclusion and Future Work

Our experiments demonstrate OVAL-Prompt's potential for open-vocabulary affordance localization, achieving results in the UMD dataset that is comparable to supervised models without fine-tuning. Furthermore, OVAL-Prompt's efficacy in real-world robotic tasks highlight its practicality.

Our method efficiently performs zero-shot affordance localization but struggles with scalability for numerous items due to individual LLM queries. Future improvements could enhance simultaneous multi-item and affordance identification. Testing in cluttered environments and eliminating the VLM's need for predefined object lists are also potential development areas.

Acknowledgment

This work was supported in part by a Qualcomm Innovation Fellowship, J.P. Morgan AI Research, Amazon, and Ford Motor Company.

References

[1] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," 2017. 1, 4

[2] J. J. Gibson, *The ecological approach to visual perception: classic edition*, ser. Psychology Press classic editions. Psychology Press. 1

[3] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao, "Learning affordance grounding from exocentric images," in *CVPR*, 2022. 1, 2, 7

[4] T. Nguyen, M. N. Vu, A. Vuong, D. Nguyen, T. Vo, N. Le, and A. Nguyen, "Open-vocabulary affordance detection in 3d point clouds," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 5692–5698. 1, 2

[5] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, K.-H. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao, J. Rettinghouse, D. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, M. Yan, and A. Zeng, "Do as i can and not as i say: Grounding language in robotic affordances," in *arXiv preprint arXiv:2204.01691*, 2022. 1, 2

[6] S. Qian, W. Chen, M. Bai, X. Zhou, Z. Tu, and L. E. Li, "Affordancellm: Grounding affordance from vision language models," *arXiv preprint arXiv:2401.06341*, 2024. 1

[7] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," in *arXiv preprint arXiv:2209.07753*, 2022. 1

[8] K. Kawaharazuka, T. Matsushima, A. Gambardella, J. Guo, C. Paxton, and A. Zeng, "Real-world robot applications of foundation models: A review," *arXiv preprint arXiv:2402.05741*, 2024. 1

[9] S. Kambhampati, K. Valmeekam, L. Guan, K. Stechly, M. Verma, S. Bhambri, L. Saldyt, and A. Murthy, "Llms can't plan, but can help planning in llm-modulo frameworks," *arXiv preprint arXiv:2402.01817*, 2024. 1

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. 1

[11] A. Rashid, S. Sharma, C. M. Kim, J. Kerr, L. Y. Chen, A. Kanazawa, and K. Goldberg, "Language embedded radiance fields for zero-shot task-oriented grasping," in *7th Annual Conference on Robot Learning*, 2023. [Online]. Available: https://openreview.net/forum?id=k-Fg8JDQmc 1

[12] K. Valmeekam, M. Marquez, S. Sreedharan, and S. Kambhampati, "On the planning abilities of large language models-a critical investigation," *Advances in Neural Information Processing Systems*, vol. 36, 2024. 1

[13] A. Myers, C. L. Teo, C. Fermüller, and Y. Aloimonos, "Affordance detection of tool parts from geometric features," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 1374–1381. 2, 3, 4

[14] D. Chen, D. Kong, J. Li, S. Wang, and B. Yin, "A survey of visual affordance recognition based on deep learning," *IEEE Transactions on Big Data*, vol. 9, no. 6, pp. 1458–1476, 2023. 2

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 2

[16] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis, "Detecting object affordances with convolutional neural networks," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 2765–2770. 2

[17] ——, "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)*, 2017, pp. 5908–5915. 2

[18] T.-T. Do, A. Nguyen, and I. Reid, "Affordancenet: An end-to-end deep learning approach for object affordance detection," 2018. 2, 4

[19] T. Nagarajan, C. Feichtenhofer, and K. Grauman, "Grounded human-object interaction hotspots from video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8688–8697. 2, 7

[20] K. Fang, Y. Zhu, A. Garg, A. Kurenkov, V. Mehta, L. Fei-Fei, and S. Savarese, "Learning task-oriented grasping for tool manipulation from simulated self-supervision," *The International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 202–216, 2020. 2

[21] S. Deng, X. Xu, C. Wu, K. Chen, and K. Jia, "3d affordancenet: A benchmark for visual object affordance understanding," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1778–1787. 2

[22] J. Wu, X. Li, S. Xu, H. Yuan, H. Ding, Y. Yang, X. Li, J. Zhang, Y. Tong, X. Jiang, B. Ghanem, and D. Tao, "Towards open vocabulary learning: A survey," *T-PAMI*, 2024. 2

[23] 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*, 2023, pp. 10 922–10 931. 2, 7

[24] G. Li, D. Sun, L. Sevilla-Lara, and V. Jampani, "One-shot open affordance learning with foundation models," 2023. 2

[25] OpenAI, "Gpt-4," https://openai.com/research/gpt-4, 2024, accessed: 2024-03-04. 2, 3

[26] "Gemini: Google deepmind's multimodal language model," https://deepmind.google, 2024, accessed: 2024-03-04. 2

[27] C. H. Song, J. Wu, C. Washington, B. M. Sadler, W.-L. Chao, and Y. Su, "Llm-planner: Few-shot grounded planning for embodied agents with large language models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023. 2

[28] D. Shah, B. Osinski, B. Ichter, and S. Levine, "Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action," 2022. 2

[29] K. Valmeekam, M. Marquez, S. Sreedharan, and S. Kambhampati, "On the planning abilities of large language models : A critical investigation," 2023. 2

[30] S. Kambhampati, "Can llms really reason and plan?" *Communications of the ACM*, Sep 2023. [Online]. Available: https://cacm.acm.org/blogcacm/can-llms-really-reason-and-plan/ 2

[31] P. Sun, S. Chen, C. Zhu, F. Xiao, P. Luo, S. Xie, and Z. Yan, "Going denser with open-vocabulary part segmentation," 2023. 3

[32] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," 2017. 4

[33] Q. Gu, J. Su, and L. Yuan, "Visual affordance detection using an efficient attention convolutional neural network," vol. 440, pp. 36–44. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0925231221000278 4

[34] X. Zhao, Y. Cao, and Y. Kang, "Object affordance detection with relationship-aware network," vol. 32, no. 18, pp. 14 321–14 333. [Online]. Available: http://link.springer.com/10.1007/s00521-019-04336-04

[35] Y. Zhang, H. Li, T. Ren, Y. Dou, and Q. Li, "Multi-scale fusion and global semantic encoding for affordance detection," in *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1–8. 4

[36] D. T. Coleman, I. A. Sucan, S. Chitta, and N. Correll, "Reducing the barrier to entry of complex robotic software: a MoveIt! case study," publisher: [object Object]. [Online]. Available: https://aisberg.unibg.it//handle/10446/87657 4

[37] J. Mai, M. Yang, and W. Luo, "Erasing integrated learning: A simple yet effective approach for weakly supervised object localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 7

[38] X. Pan, Y. Gao, Z. Lin, F. Tang, W. Dong, H. Yuan, F. Huang, and C. Xu, "Unveiling the potential of structure preserving for weakly supervised object localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 11 642–11 651. 7

[39] Y. Yao, F. Wan, W. Gao, X. Pan, Z. Peng, Q. Tian, and Q. Ye, "Ts-cam: Token semantic coupled attention map for weakly supervised object localization," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2022. 7

[40] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao, "Grounded affordance from exocentric view," 2023. 7

[41] D. Hadjivelichkov, S. Zwane, M. P. Deisenroth, L. Agapito, and D. Kanoulas, "One-shot transfer of affordance regions? affcorrs!" 2022. 7

VII. Supplementary Material

In addition to the UMD dataset we evaluated our method on the AGD20K dataset.

1) AGD20K: The AGD20K dataset is intended for affordance grounding tasks and includes over 20,000 images across 36 affordance categories. It utilizes three evaluation metrics: Kullback-Leibler Divergence (KLD), similarity metric (SIM), and Normalized Scanpath Saliency (NSS). Detailed information on their design and implementation can be found in the supplementary material of the Luo et al. paper [3]. For the performance of other models, we referenced the data from the paper by Li et al. [23].

Our method (see table IV) has higher KLD than the best seen model by 9.429 points and the unseen by 7.427. Our SIM and NSS scores were competitive with other models beating several models and only being surpassed by LOCATE [23] for both seen and unseen categories, suggesting our segmentation closely aligns with the ground truth. The high KLD stems from our method producing a binary mask for the affordance area, in contrast to the ground truth comparison that uses a value distribution, the peak of which indicates the affordance region. Consequently, our non-distributional output incurs a significant penalty in KLD calculation.

TABLE IV: AGD20K Dataset Evaulation

Method		**Seen**			**Unseen**	
	KLD↓	**SIM↑**	**NSS↑**	**KLD↓**	**SIM↑**	**NSS↑**
EIL [37]	1.931	0.285	0.522	2.167	0.227	0.330
SPA [38]	5.528	0.221	0.357	7.425	0.169	0.262
TS-CAM [39]	1.842	0.260	0.336	2.104	0.201	0.151
Hotspots [19]	1.773	0.278	0.615	1.994	0.237	0.577
Cross-view-AG [3]	1.538	0.334	0.927	1.787	0.285	0.829
Cross-view-AG+ [40]	1.489	0.342	0.981	1.765	0.279	0.882
AffCorrs† [41]	1.407	0.359	1.026	1.618	0.348	1.021
LOCATE [23]	**1.226**	**0.401**	**1.177**	**1.405**	**0.372**	**1.157**
OVAL-Prompt(Ours)	10.649	0.339	1.044	8.832	0.365	0.925