

Worldafford: Affordance Grounding based on Natural Language Instructions

Changmao Chen¹, Yuren Cong², Zhen Kan¹

¹University of Science and Technology of China

²TNT, Leibniz University Hannover

Abstract—Affordance grounding aims to localize the interaction regions for the manipulated objects in the scene image according to given instructions, which is essential for Embodied AI and manipulation tasks. A key challenge in affordance grounding is enabling the agent to understand human instructions, identify usable tools in the environment, and determine how to use them to complete the task. Most recent works primarily support simple action labels as input instructions for localizing affordance regions, failing to capture complex human objectives. Moreover, these approaches typically identify affordance regions of only a single object in object-centric images, ignoring the object context and struggling to localize affordance regions of multiple objects in complex scenes for practical applications. To address this concern, for the first time, we introduce a new task of affordance grounding based on natural language instructions, extending it from previously using simple labels for complex human instructions. For this new task, we propose a new framework, WorldAfford. We design a novel Affordance Reasoning Chain-of-Thought Prompting to reason about affordance knowledge from LLMs more precisely and logically. Subsequently, we use SAM and CLIP to localize the objects related to the affordance knowledge in the image. We identify the affordance regions of the objects through an affordance region localization module. To benchmark this new task and validate our framework, an affordance grounding dataset, LLMaFF, is constructed. We conduct extensive experiments to verify that WorldAfford performs state-of-the-art on the previous AGD20K and the new LLMaFF dataset. In particular, WorldAfford can localize the affordance regions of multiple objects and provide an alternative when objects in the environment cannot fully match the given instruction. Our Project page: <https://worldafford.github.io/>.

Index Terms—affordance grounding, natural language instructions, llms

I. INTRODUCTION

Embodied agents can interact with a physical environment and potentially perform heavy tasks based on human instructions. In order for robots to better manipulate objects in complex scenes, it is urgent to understand which part of the object is the interaction region. Affordance grounding, which aims to localize potential interaction regions for the manipulated objects in the scene image depending on the given instruction, can provide a new experience for Embodied AI and has the potential to significantly increase efficiency and flexibility. As a result, it has recently attracted a significant amount of attention [1]–[3].

A critical challenge in affordance grounding is instruction comprehension, which means that the embodied agent should understand the human instructions and reason about the actions it is going to perform, which emphasizes active interaction

between humans and their environment rather than passive detection. Furthermore, the agent should analyze which tools in the usage environment can accomplish the given instructions and localize the interaction regions (*i.e.*, affordance regions) on the objects. These challenges are expected to be alleviated through using large-scale vision-language foundation models. Unfortunately, the currently available models [4]–[7] have not performed satisfactorily on this particular task.

Most recent works [1]–[3] attempt to transfer knowledge from exocentric images of an object in an active state to egocentric images where the object is not being used. They have achieved impressive progress, making dataset collection easier and learning that the affordance region of an object changes dynamically depending on the different given instructions. Nevertheless, current approaches can only support simple action labels (*e.g.*, “catch” shown in Fig. 1) as input instructions, which cannot express complex human goals. Besides, these methods can only identify the affordance region of a single object in object-centric images, overlook object context, and still fall short in localizing the affordance regions of multiple objects in complex scene images for practical applications in the real world. In this paper, for the first time, we introduce a new task of affordance grounding based on natural language instructions, extending affordance grounding from previously using simple action labels to complex natural language instructions. This new task moves toward real-world applications with significant implications for Embodied AI. For this task, we propose a novel framework, WorldAfford, which integrates the large language model (LLM), Segment Anything model (SAM) [8], CLIP [7] and the affordance region localization module. We first use the LLM to process the natural language instruction. To reason about affordance knowledge from the LLM more precisely and logically, we design a novel Affordance Reasoning Chain-of-Thought Prompting (ARCoT) including Object-Oriented Reasoning Prompting and Action-Oriented Reasoning Prompting. Subsequently, we employ SAM and CLIP to segment and select the objects associated with the actions inferred by the LLM. Moreover, a Weighted Context Broadcasting module (WCB) is proposed and integrated into the affordance region localization module. It allows our framework to focus on more informative objects and to identify affordance regions of multiple objects. To benchmark the new task and validate our framework, we constructed a new dataset, LLMaFF, containing real-world images with natural language instructions and manually la-



Fig. 1: Different from previous works only using naïve action labels for affordance grounding, WorldAfford can derive affordance knowledge from LLMs and precisely localize the affordance regions corresponding to natural language instructions. In this way, our framework can work effectively in complex open-world environments. The results in the second row are from Cross-view-AG+ [3].

beled affordance maps. We train our model on AGD20K using image-level labels as supervision without expensive pixel-level annotation. All LLMaFF images and language instructions are unseen in training. Experimental results demonstrate that our framework outperforms the previous methods both on the existing AGD20K [1] dataset and the new LLMaFF dataset. Our main contributions can be summarized as follows:

- We introduce a new task of affordance grounding based on natural language instructions, extending affordance grounding from using simple action labels to complex natural language instructions.
- We propose a framework for this new task named WorldAfford, which integrates the LLM and other vision models. To reason about affordance knowledge from LLMs, we introduce an Affordance Reasoning Chain-of-Thought Prompting. In addition, we propose a Weighted Context Broadcasting module, allowing WorldAfford to localize affordance regions of multiple objects.
- A new dataset LLMaFF is constructed to benchmark the new task.
- We conduct extensive experiments to validate that our model performs state-of-the-art on both the AGD20K dataset and our new LLMaFF dataset.

II. RELATED WORK

A. Affordance grounding

Visual affordance grounding has been intensively explored in the fields of robotics and computer vision [9]–[13]. Traditional approaches [14] mainly learn the affordance through fully supervised learning. Luo *et al.* [1] propose a Cross-view-AG knowledge transfer framework for affordance grounding, in which the affordance knowledge is acquired from exocentric

human-object interactions, and transfer to egocentric images. Li *et al.* [2] extract object-related information from exocentric images and match it to the objects to localize the affordance regions. However, such methods use only naïve action labels for affordance grounding. In this work, we use flexible natural language as supervision to guide agents in localizing affordance regions of multiple objects in complex scenes images.

B. Large language models and Vision foundation models for affordance grounding

Some studies [15], [16] have used large language models (LLMs) to guide robotic arm object grasping, focusing on basic object perception without addressing fine-grained shapes, functions, or uses. Our work differs by employing LLMs with an affordance reasoning Chain-of-Thought (ARCoT) method to interpret open-world human instructions and reason about diverse objects. Recent studies [17] demonstrate Chain-of-Thought (CoT) can significantly enhance LLM performance on complex reasoning tasks.

Vision-language models have also shown promise in robotics [18], [19]. Some works [20], [21] detect affordances in 3D point clouds but lack human instructions, generalization to unseen objects, and rely on manual annotation. Li *et al.* [22] address one-shot affordance learning. We, however, leverage CLIP for semantic understanding and SAM for spatial recognition to identify objects based on language input, highlighting the synergy between LLM-based reasoning and vision models for affordance grounding.

C. Affordance Grounding Dataset

Affordance grounding [1]–[3], [9], [14], [23] has primarily focused on datasets like AGD20K [1], which target single actionable object scenarios. More recently, Hadjivelichkov *et al.* [24] introduced the UMD-i dataset for one-shot affordance learning with pixel-level labels, while Nguyen *et al.* [25] proposed the IIT-AFF dataset, which lacks semantic affordance information and relies solely on image inputs. To overcome these limitations, we present the LLMaFF dataset.

III. TASK DEFINITION AND LLMaFF DATASET

Given an image I and a natural language instruction t , affordance grounding based on natural language instructions aims to localize the interaction regions of objects in the scene image and the instruction can be completed through these interactions. Compared to the setting in previous works [1]–[3], [24], affordance grounding based on natural language instructions is more oriented towards practical applications in the real world since there is no restriction on the number of objects in the image and complexity of the input instructions.

To facilitate and benchmark this new task, we construct a new dataset, LLMaFF, consisting of 550 complex environmental images with natural language instructions and manually labelled affordance maps. The data collection pipeline is shown in Fig. 2. The source images of our dataset are primarily sampled from IIT-AFF [25]. Due to the limited

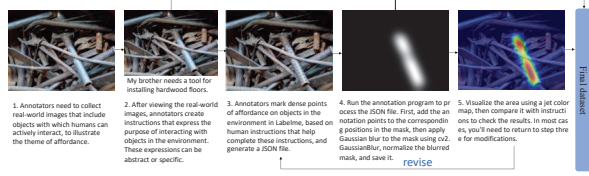


Fig. 2: Data Collection Pipeline for our WorldAfford benchmark.

object categories of IIT-AFF, we augment the dataset with the images sampled from Ego4D [26] and the Internet.

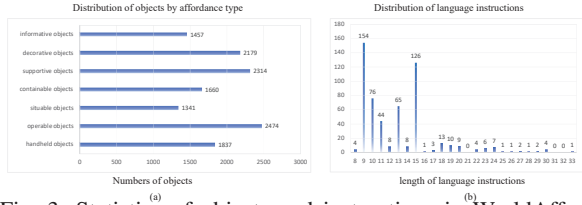


Fig. 3: Statistics of objects and instructions in WorldAfford. (a) Distribution of objects by affordance type. (b) Distribution of language instructions

AGD20K [1] annotates the affordance regions with sparse points and applies a Gaussian kernel to generate ground truth. In contrast, we employ dense points to annotate the affordance map of multiple objects based on the language instructions, which requires careful identification of the objects and their interactions. We find that the density and distribution of the points have a significant impact on the labelling results, thus ensuring a uniform distribution of annotation points across multiple objects is crucial to avoid certain regions in the affordance map appearing blank or with faint heat.

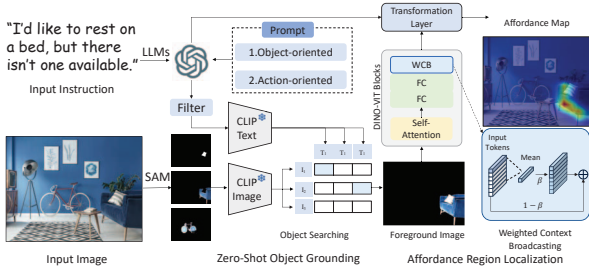


Fig. 4: Framework of WorldAfford.

Based on the affordance of objects in the environment, we categorized them into eight types: handheld objects (1837), operable objects (2474), suitable objects (1341), containable objects (1660), supportive objects (2314), decorative objects (2179), and informative objects (1457). We also conducted a statistical analysis of the length of human language instructions in the dataset, as shown in Fig. 3.

IV. PROPOSED APPROACH

A. WorldAfford Framework

We propose WorldAfford as a general framework for affordance grounding that incorporates complex instruction understanding and multi-object affordance localization with minimal

training cost. First, we use the LLM [27] to analyze instructions and derive affordance knowledge through affordance reasoning chain-of-thought prompting. Then, SAM [8] and CLIP [7] enable zero-shot multi-object grounding, segmenting and selecting objects for sub-actions identified by the LLM. We further integrate a Weighted Context Broadcasting (WCB) into the affordance region localization module for precise localization of multiple objects' affordance regions. The WorldAfford framework is shown in Fig. 4.

B. Affordance Reasoning Chain-of-Thought Prompting

Rather than relying on direct inference, our approach employs a straightforward and effective chain-of-thought prompting to enhance the capabilities of LLMs in affordance reasoning, as shown in Fig. 5. The proposed chain-of-thought prompting consists of two primary strategies: (1) **object-oriented reasoning prompting**, and (2) **action-oriented reasoning prompting**.

1) *Object-Oriented Reasoning Prompting*.: We first utilize the LLM to reason about the possible objects that can afford the given instruction. Considering that multiple objects are likely to be necessary for completing the instruction, the LLM is requested to output a set of object categories \mathcal{O} :

$$\mathcal{O} = LLM(k, t, p_{obj}), \quad (1)$$

where k indicates the size of the object set and t denotes the given natural language instruction. The prompt p_{obj} for object-oriented reasoning is specifically designed as follows:

Prompt: What are the [k] most common objects that can be used if [t]?

Output: Chair..., Hammock..., Blanket and Pillows...

The object-oriented reasoning prompts the LLM to provide diverse objects suitable for an action. Moreover, it associates alternative tools in case the best tool does not exist in the environment, which facilitates the accomplishment of the instruction. These inferred object categories from the LLM are further utilized for subsequent action-oriented reasoning. We

Instruction: I'd like to rest on a bed, but there isn't one available.

What are the [k] most common objects that can be used if [Instruction]?

Chair..., Hammock..., Blanket and Pillows...

Select skills from [lp] to interact with [Chair, Hammock, Blanket and Pillows] if [Instruction]?

sit on the chair. ..., lie on the hammock..., hold the blanket and pillows.....

Fig. 5: The affordance reasoning Chain-of-Thought prompting.

designed a filter function based on the large model's output about object descriptions to filter out excessive explanatory text, which sometimes includes irrelevant objects, hindering the subsequent object search.

2) *Action-Oriented Reasoning Prompting.*: Different from previous work on locating the affordance for single action labels, to address the complex natural language instructions, we utilize the powerful prior knowledge of the LLM to decompose a complex instruction into several simple sub-actions. Given a pre-defined predicate list, we prompt the LLM to select the appropriate predicates from it and assign these predicates to the objects in the object set \mathcal{O} . A set of sub-actions \mathcal{A} is generated as follows:

$$\mathcal{A} = \text{LLM}(\mathcal{O}, l_p, t, p_{act}), \quad (2)$$

where l_p indicates the pre-defined predicate list and t denotes the given instruction. Each sub-action consists of a predicate and an object. The prompt p_{act} for action-oriented reasoning is specifically designed as follows:

Prompt: Select skills from [#l_p] to interact with the above [#t] to help me if [#t]?

Output: sit on the chair..., lie on the hammock..., hold the blanket and pillows...

The inferred sub-actions from the LLM are further utilized as the input of the following affordance region localization module. We use the LLM to extract object-level knowledge and aggregate action-level knowledge. In the inference process of the LLM, irrelevant information in the natural language instruction is ignored and the highly abstract instruction is transformed into a series of executable sub-actions. The powerful reasoning ability and adaptive results of the affordance reasoning chain-of-thought facilitate the subsequent zero-shot object grounding and the affordance region localization.

C. Zero-shot Multiple Object Grounding

To integrate the affordance knowledge provided by the LLM with visual information about the environment, our approach leverages the capabilities of Segment Anything Model (SAM) [8] and CLIP [7] to effectively ground the relevant objects in the scene image according to the given natural language instruction. The impressive zero-shot performance of SAM and CLIP enables our framework to precisely localize objects across the open world without the need for extensive and expensive training on large-scale datasets.

Initially, SAM produces N segmentation masks for the input image. These masks, while precisely segmented, lack semantic labels and unavoidably contain irrelevant objects. In order to obtain the object masks that are relevant to the given instruction, CLIP is integrated to compute the similarity between the visual appearance of the masks and the object categories provided by the LLM. We extract the corresponding regions from the original image I based on the segmentation masks. Subsequently, the cropped regions m are encoded by the CLIP image encoder E_{image} while the textual object categories o are encoded by the CLIP text encoder E_{text} . The probability p of the mask being classified as the i -th object category can be formulated as:

$$p = \frac{\exp(\text{sim}(E_{image}(m), E_{text}(o_i))/\alpha)}{\sum_{o_i \in \mathcal{O}} \exp(\text{sim}(E_{image}(m), E_{text}(o_i))/\alpha)}, \quad (3)$$

where $\text{sim}(\cdot)$ denotes the cosine similarity function and \mathcal{O} indicates the set of object categories from the LLM. The scaling factor α is set to 0.1 in practice. We establish a boundary to determine whether the masks from SAM are valid. The masks with probability p above the boundary are identified as valid masks. With these active masks, we construct a full-view segmentation mask in which the region covered by the valid masks is viewed as foreground, while the remaining area is considered as background. This full-view mask is the same size as the input image and is further used for affordance region localization.

D. Affordance Region Localization

To localize the affordance region of the objects in the image corresponding to the given instruction, we employ LOCATE [2] and enhance the grounding performance through two crucial improvements: (1) We use the full-view mask resulting from zero-shot multi-object grounding to preserve the foreground and mask off the background as the input, rather than the entire image. (2) We propose a weighted context broadcasting (WCB) module, seamlessly integrating it into DINO-ViT [28] to enable the model to prioritize informative objects. With these improvements, our approach outperforms the original LOCATE and can localize multiple affordance regions with the knowledge provided by the LLM.

We utilize the full-view mask from zero-shot multi-object grounding to mask off the irrelevant objects in the image. The relevant objects are preserved and the image is forwarded into DINO-ViT to extract deep part-aware features. We design a Weighted Context Broadcasting (WCB) module inspired by [29] and incorporate it into DINO-ViT as demonstrated in Fig. 4. Given a sequence of N patch tokens, the WCB module combines the average context tokens with the input tokens in a weighted manner as follows:

$$\text{WCB}(x_i) = x_i * \beta + \frac{1}{N} \sum_{j=1}^N x_j * (1 - \beta), \quad (4)$$

where the weight β is an empirically determined hyperparameter. In order to improve the model's capability to perceive multiple objects, it is expected that the attention maps in the self-attention modules of DINO-ViT are dense rather than sparse. It has been discussed in [29] that aggregating the average context token can facilitate the self-attention modules to learn dense attention maps. However, such simple aggregation makes training difficult since the target attention is unknown and uncertain. To solve this issue, we introduce a weight to balance the aggregation. With the proposed WCB, the target attention is easier to learn and the model can focus on more informative objects. The experiment in section V-C also demonstrates that our approach outperforms the previous works [3] in terms of affordance grounding for objects.

The feature maps generated by DINO-ViT are further refined by a transformation layer including a feed-forward layer and two subsequent convolutional layers. We follow the training strategy of LOCATE [2] to transfer affordance

TABLE I: Comparison of WorldAfford with other state-of-the-art affordance grounding methods on AGD20K. The best numbers are highlighted in **bold**.

Approach	Input Instruction	KLD↓	SIM↑	NSS↑
Hotspots [23]	Action Label	1.773	0.278	0.615
Cross-view-AG [1]	Action Label	1.538	0.334	0.927
Affcorr [24]	Action Label	1.407	0.359	1.026
LOCATE [2]	Action Label	1.226	0.401	1.177
Cross-view-AG+ [3]	Action Label	1.213	0.403	1.242
WorldAfford(ours)	Action Label	1.201	0.406	1.255

TABLE II: Comparison on LLMaFF dataset. We manually select labels for the other methods to comparison with them. WorldAfford outperforms all previous methods across all evaluation metrics. The best results are highlighted in **bold**.

Approach	Input Instruction	KLD↓	SIM↑	NSS↑
Cross-view-AG+ [3]	Action Label	2.927	0.123	-0.194
Cross-view-AG [1]	Action Label	2.887	0.119	0.118
LOCATE [2]	Action Label	1.958	0.212	1.713
WorldAfford (ours)	Natural Language	1.163	0.386	2.819

knowledge from exocentric images to egocentric images. To predict the affordance maps, a convolutional layer with a window size of 1×1 is utilized to project the channel number to the total number of the action categories in the pre-defined predicate list l_p . We aggregate the affordance maps corresponding to the action categories provided by the LLM and normalized them to limit the activation values in the map between 0 and 1 as the final output.

V. EXPERIMENTS

A. Datasets and Evaluation Metrics

We conduct experiments on two datasets: AGD20K [1], which includes 20,061 demonstration images and 6,060 object images for training. We evaluate our method on its test set of 1,675 images, focusing on affordance grounding guided by single action labels. Additionally, we use our proposed LLMaFF dataset, consisting of 550 complex environment images with natural language instructions and affordance maps. This dataset allows us to assess our method’s performance in affordance grounding based on natural language instructions. Similar to [1], we employ KLD, SIM, and NSS metrics to quantify the correspondence between predicted affordance maps and ground truth. Only the affordance region localization module is trained, while other modules remain frozen. Training occurs solely on AGD20K, following baseline settings.

B. Implementation Details.

We use GPT-4 [27] as the large language model, while both the CLIP [7] and the Segment Anything Model (SAM) [8] implement object matching and segmentation in a zero-shot fashion. The affordance information is extracted from the output of the large language model by removing most of the irrelevant text to allow the CLIP to more accurately localize the position of objects. The affordance region localization module is trained on a RTX 3090 GPU. We load the pre-trained DINO-ViT [28] model and finetune the features it extracts from images. We set the weight β to 0.88, and the

number k in eq. (1) is set to 3. We use a learning rate of 0.005, a decay factor of $5e-4$, a batch size of 16, and train the affordance region localization module for 35 epochs.

C. Quantitative results

We validate WorldAfford on the AGD20K dataset, commonly used in affordance grounding approaches [1]–[3], which typically use simple action labels for single-object affordance localization in object-centric images. Given WorldAfford’s reliance on natural language instructions, direct comparison is challenging.

To address this, we input only action labels into the affordance region localization module for comparison with these approaches. Results in TABLE I demonstrate our approach outperforms previous methods even in this simplified setup, establishing a new state-of-the-art in affordance grounding. The weighted context broadcasting module enhances object-focused information processing, improving affordance region identification. We further evaluate our method on the LLMaFF

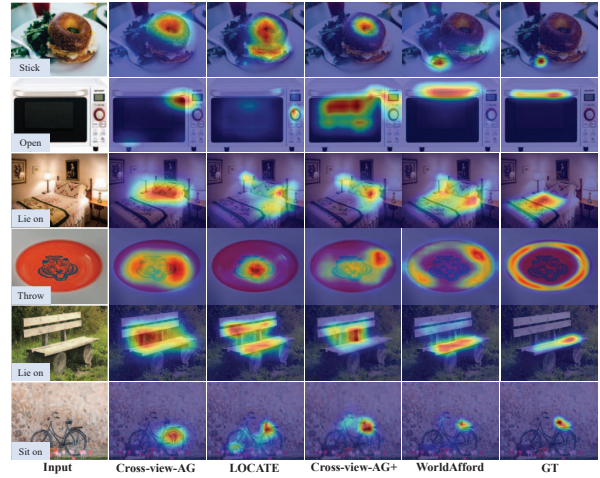


Fig. 6: Visual comparison on the AGD20K dataset. Compared to previous methods, our method can infer more precise affordance maps.

dataset. Since previous methods cannot handle textual instructions directly, we manually select action labels for comparison. Results in TABLE II show our method effectively localizes affordance regions in complex scene images.

Cross-view-AG+ achieves strong results on AGD20K but struggles on LLMaFF, indicated by a negative NSS score (-0.194), suggesting challenges in adapting to complex tasks and potential overfitting to AGD20K. Cross-view-AG and LOCATE also demonstrate decreased performance on LLMaFF, highlighting their limitations in complex scene affordance localization.

D. Qualitative results

Qualitative comparisons on AGD20K are shown in Fig. 6. Cross-view-AG tends to produce overly large affordance re-

gions, sometimes including irrelevant areas. LOCATE predicts smaller regions but often misses parts of the affordance region. Cross-view-AG+ identifies regions associated with action labels but lacks accuracy. In contrast, WorldAfford achieves state-of-the-art performance, providing sharper and more accurate results. The weighted context broadcasting module (WCB) enhances focus on informative objects, improving object knowledge and localization accuracy.

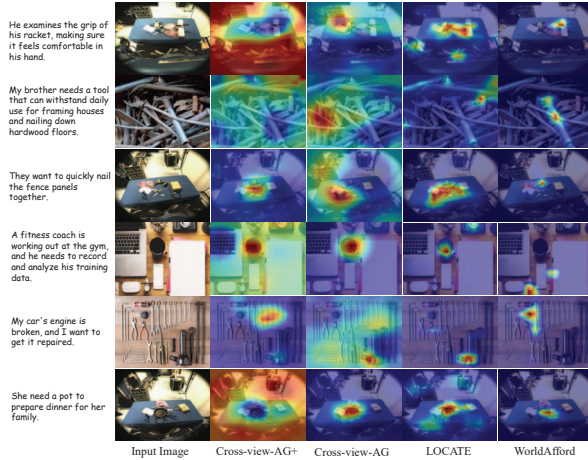


Fig. 7: Visual comparison on the LLMaFF dataset. We manually assign labels to other methods since they cannot adopt the textual input. The labels, "swing", "carry", "catch", "pick up", "catch", and "carry" correspond to the first through sixth rows respectively.

Fig. 7 displays results on the LLMaFF dataset. Cross-view-AG+ struggles to identify affordance regions of multiple objects, leading to disordered color distributions and ineffective visual information. Cross-view-AG shows some success in capturing object information but is biased toward objects with more training samples. LOCATE captures affordance for a few objects but may activate regions of irrelevant objects or interpret multiple objects as one. In comparison, WorldAfford predicts more consistent affordance maps aligned with natural language instructions, offering accurate localization of multiple object affordance regions and richer visual information.

1) *Results on complex language instructions:* We address the challenge of affordance grounding with complex language instructions, as depicted in Fig. 8. Unlike simpler instructions, these require deeper human knowledge, highlighting our method's flexibility and creativity. WorldAfford effectively identifies intricate affordance regions, exemplifying object interactions such as using a knife and an apple for slicing. This provides detailed visual information to enhance the agent's ability in complex tasks. Our approach systematically activates affordance regions for tasks like building a chair, including sawing, measuring, and assembling. This advancement opens new avenues for robotics and AI applications, enriching agent-environment interactions significantly.

Results on complex language instructions



LLMs		
	A man wants an apple and plans to slice it into pieces.	A carpenter wants to use wooden planks and nails to create a chair.
type	multi-objects complementary affordance.	multi-objects sequential affordance.
	He should first pick up the apple itself, and then use a sharp knife to cut the apple into pieces. And then cut the apple on a cutting board and prevent any injury while minimizing mess.	He might need to carry a saw to cut them. And pick up the ruler to measure the length, width, and height of the chair parts, and then hit with hammer to drive the nails into the wooden planks and secure them together.

Fig. 8: Affordance results based on difficult language instructions. While previous methods struggle to infer from difficult language instructions, our method demonstrates the capability to comprehend such instructions and accurately identify the affordance regions of multiple objects.

TABLE III: Generalization ability comparison of WorldAfford with other state-of-the-art affordance grounding methods on AGD20K.

Approach	Input Instruction	KLD↓	SIM↑	NSS↑
Hotspots [23]	Action Label	1.994	0.237	0.577
Cross-view-AG [1]	Action Label	1.787	0.285	0.829
Affcorr [24]	Action Label	1.618	0.348	1.021
LOCATE [2]	Action Label	1.405	0.372	1.157
WorldAfford(ours)	Action Label	1.393	0.38	1.225

E. Generalization ability and learnable parameters

To evaluate the generalization ability of our method, we add the results of the unseen test on AGD20K, which is shown in TABLE III. Additionally, all LLMaFF images, including various scenes and many object categories such as nail gun, smartwatch and so on, are unseen in training, which also demonstrates the superior generalization ability of our method. We use the the knowledge of foundation models, the training cost is very low, and the comparison of learnable parameters: 120.03M(Cross-view-AG)/82.27M (Cross-view-AG+)/6.5M (LOCATE)/6.5M (WorldAfford).

TABLE IV: Ablation results of the proposed modules. LMA denotes the action information associated with the manipulated objects inferred from the LLM. WCB indicates the weighted context broadcasting module. LMO represents the object information inferred from the LLM.

LMA	WCB	LMO	KLD↓	SIM↑	NSS↑
			3.073	0.105	-0.059
	✓		2.729	0.114	0.428
✓			2.768	0.124	0.303
✓	✓		2.335	0.155	0.981
✓	✓	✓	2.336	0.180	1.081
✓	✓	✓	1.700	0.256	2.325
✓	✓	✓	1.163	0.386	2.819

TABLE V: The results of using entire images as input and masking off irrelevant objects on LLMaFF.

Input	KLD↓	SIM↑	NSS↑
entire image	2.752	0.134	0.27
mask off	1.163	0.386	2.819

F. Ablation Study

We conduct the ablative experiments on the LLMaFF dataset to validate the effectiveness of the affordance reasoning chain-of-thought prompting (ARCoT). The results shown in TABLE IV demonstrate that the object information and the action information derived from the LLM via our affordance reasoning chain-of-thought prompting (ARCoT) can both improve the performance for the task of affordance grounding based on language instructions. We also validate that the proposed WCB module can enhance the perception of affordance regions by enabling the model to focus on more informative objects. Overall, our contributions significantly improve the affordance grounding capabilities of the model and establish a new state-of-the-art performance in the affordance grounding based on natural language instructions task. To verify our adjustments for masking off irrelevant objects, we conduct experiments on LLMaFF, the results is shown in TABLE V.

VI. CONCLUSION

In this paper, we introduce a new task of affordance grounding based on natural language instructions and propose a novel framework, WorldAfford. Our framework uses LLMs to process natural language instructions and employs SAM and CLIP for object segmentation and selection. We further propose a Weighted Context Broadcasting module, allowing WorldAfford to localize affordance regions of multiple objects. Additionally, we present a new dataset, LLMaFF, to benchmark this task. The experimental results demonstrate that WorldAfford outperforms the other state-of-the-art methods for affordance grounding on both the AGD20K dataset and the new LLMaFF dataset.

REFERENCES

- [1] 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*, 2022, pp. 2252–2261.
- [2] 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.
- [3] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao, "Grounded affordance from exocentric view," *International Journal of Computer Vision*, pp. 1–25, 2023.
- [4] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 888–12 900.
- [5] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *NeurIPS*, 2023.
- [6] Y. Cong, W. Liao, B. Rosenhahn, and M. Y. Yang, "Learning similarity between scene graphs and images with transformers," *arXiv preprint arXiv:2304.00590*, 2023.
- [7] 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*. PMLR, 2021, pp. 8748–8763.
- [8] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.
- [9] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao, "Learning visual affordance grounding from demonstration videos," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [10] Z. Zhou, S. Wang, Z. Chen, M. Cai, and Z. Kan, "A novel framework for improved grasping of thin and stacked objects," *IEEE Transactions on Artificial Intelligence*, 2023.
- [11] —, "A robotic visual grasping design: Rethinking convolution neural network with high-resolutions," *arXiv preprint arXiv:2209.07459*, 2022.
- [12] S. Wang, Z. Zhou, and Z. Kan, "When transformer meets robotic grasping: Exploits context for efficient grasp detection," *IEEE robotics and automation letters*, vol. 7, no. 3, pp. 8170–8177, 2022.
- [13] S. Egami, S. Nishimura, and K. Fukuda, "A framework for constructing and augmenting knowledge graphs using virtual space: Towards analysis of daily activities," in *2021 IEEE 33rd international conference on tools with artificial intelligence (ICTAI)*. IEEE, 2021, pp. 1226–1230.
- [14] K. Fang, T.-L. Wu, D. Yang, S. Savarese, and J. J. Lim, "Demo2vec: Reasoning object affordances from online videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2139–2147.
- [15] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei, "Voxposer: Composable 3d value maps for robotic manipulation with language models," *arXiv preprint arXiv:2307.05973*, 2023.
- [16] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.
- [17] Z. Li, B. Peng, P. He, M. Galley, J. Gao, and X. Yan, "Guiding large language models via directional stimulus prompting," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [18] S. Huang, Z. Jiang, H. Dong, Y. Qiao, P. Gao, and H. Li, "Instruct2act: Mapping multi-modality instructions to robotic actions with large language model," *arXiv preprint arXiv:2305.11176*, 2023.
- [19] M. Khan, Y. Qiu, Y. Cong, J. Abu-Khalaf, D. Suter, and B. Rosenhahn, "Segment any object model (saom): Real-to-simulation fine-tuning strategy for multi-class multi-instance segmentation," *arXiv preprint arXiv:2403.10780*, 2024.
- [20] 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)*. IEEE, 2023, pp. 5692–5698.
- [21] T. Van Vo, M. N. Vu, B. Huang, T. Nguyen, N. Le, T. Vo, and A. Nguyen, "Open-vocabulary affordance detection using knowledge distillation and text-point correlation," *arXiv preprint arXiv:2309.10932*, 2023.
- [22] G. Li, D. Sun, L. Sevilla-Lara, and V. Jampani, "One-shot open affordance learning with foundation models," *arXiv preprint arXiv:2311.17776*, 2023.
- [23] 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.
- [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*. PMLR, 2023, pp. 550–560.
- [25] A. Nguyen, D. Kanoulas, D. Caldwell, and N. Tsagarakis, "Object-based affordances detection with convolutional neural networks and dense conditional random fields," 09 2017.
- [26] 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*, 2022, pp. 18 995–19 012.
- [27] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
- [28] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.
- [29] N. Hyeon-Woo, K. Yu-Ji, B. Heo, D. Han, S. J. Oh, and T.-H. Oh, "Scratching visual transformer's back with uniform attention," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 5807–5818.