RoboAfford: A Dataset and Benchmark for Enhancing Object and Spatial Affordance Learning in Robot Manipulation

Yingbo Tang*
Institute of Automation, CAS
School of Artificial Intelligence, UCAS
Beijing, China
tangyingbo2020@ia.ac.cn

Lingfeng Zhang*
Shenzhen International Graduate School, Tsinghua University
Shenzhen, Guangdong, China
lfzhang715@gmail.com

Shuyi Zhang
Institute of Automation, CAS
School of Artificial Intelligence, UCAS
Beijing, China
zhangshuyi2024@ia.ac.cn

Yinuo Zhao
Beijing Institute of Technology
Beijing, China
ynzhao@bit.edu.cn

Xiaoshuai Hao†
Beijing Academy of Artificial Intelligence (BAAI)
Beijing, China
xshao@baai.ac.cn

Abstract

Robot manipulation is a fundamental capability of embodied intelligence, enabling effective robot interactions with the physical world. In robot manipulation tasks, predicting precise grasping positions and object placement is essential. Achieving this requires ***object recognition*** to localize target object, predicting ***object affordances*** for interaction and ***spatial affordances*** for optimal arrangement. While Vision-Language Models (VLMs) provide insights for high-level task planning and scene understanding, they often struggle to predict precise action positions, such as functional grasp points and spatial placements. This limitation stems from the lack of annotations for object and spatial affordance data in their training datasets. To address this gap, we introduce ***RoboAfford***, a novel large-scale dataset designed to enhance object and spatial affordance learning in robot manipulation. Our dataset comprises 819,987 images paired with 1.9 million question answering (QA) annotations, covering three critical tasks: *object affordance recognition* to identify objects based on attributes and spatial relationships, *object affordance prediction* to pinpoint functional grasping parts, and *spatial affordance localization* to identify free space for placement. Complementing this dataset, we propose ***RoboAfford-Eval***, a comprehensive benchmark for assessing affordance-aware prediction in real-world scenarios, featuring 338 meticulously annotated samples across the same three tasks. Extensive experimental results reveal the deficiencies of existing VLMs in affordance learning, while fine-tuning on the ***RoboAfford*** dataset significantly enhances their affordance prediction in robot manipulation, validating the dataset's effectiveness. The dataset, benchmark and evaluation code will be made publicly available to facilitate future research. Project website: *https://roboafford-dataset.github.io/*.

CCS Concepts

• **Computing methodologies** → **Robotic planning**; **Vision for robotics**; **Cognitive robotics**.

Keywords

Robot Manipulation, Object Affordance, Spatial Affordance.

1 Introduction

Enabling robots to physically interact with objects in 3D environments presents a fundamental challenge in embodied AI, requiring advanced perception, manipulation skills, and adaptability to navigate the complexities of real-world interactions [19, 37]. Recent advancements in Vision-Language Models (VLMs) [14, 17, 33, 35] have significantly improved robots' perceptual and reasoning capabilities, facilitating remarkable progress in scene understanding and task planning. Additionally, these models have shown considerable potential in various embodied tasks, including vision-language navigation [7, 42, 45], embodied manipulation [34, 38, 40], and embodied reasoning [33, 35, 43]. However, despite these advancements in high-level planning and semantic understanding, existing VLMs still encounter significant challenges in achieving fine-grained interaction understanding in real-world scenarios.

In robotic manipulation, VLMs are required to ground target objects from language instructions while identifying feasible grasp points and suitable placement regions. This involves three core capabilities including (1) ***Object Affordance Recognition***: identifying objects based on attributes such as category, color, size, and spatial relations; (2) ***Object Affordance Prediction***: localizing functional parts of objects to support specific actions, such as the handle of a teapot for grasping; and (3) ***Spatial Affordance Localization***: detecting viable placement areas, like an empty shelf in a fridge. Despite recent progress on affordance learning, no unified framework effectively integrates object and spatial affordances. For example, a VLM can infer that a mug can be grasped by its handle and should

*These authors contributed equally to this paper.
†Corresponding Author.

ACM MM, Dublin, Ireland

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

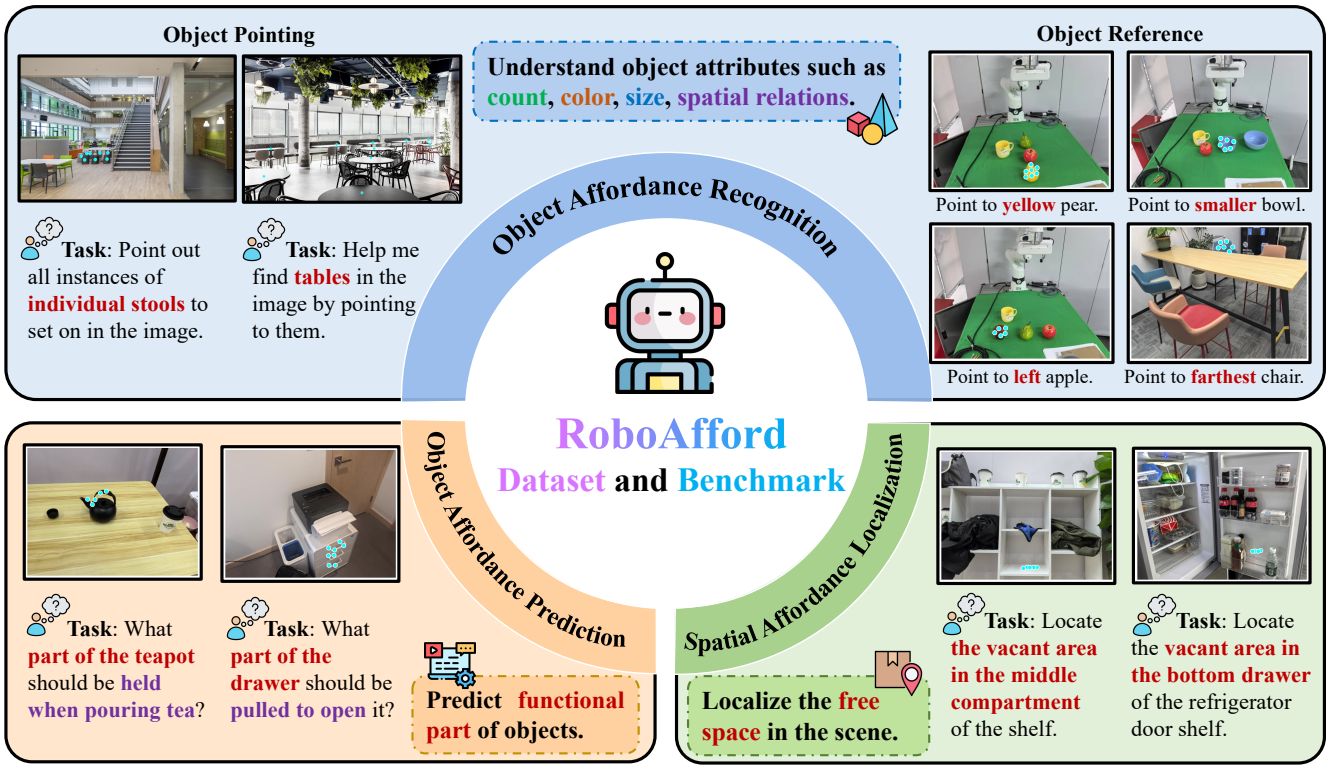

Figure 1: Overview of the RoboAfford Dataset. It encompasses three key dimensions: object affordance recognition, object affordance prediction, and spatial affordance localization.

be placed within the vacant space on the table, however, it may struggle with predicting the exact position for grasping and feasible area for placement that relative to nearby objects. This limitation arises from the lack of fine-grained supervision in current training data, which typically consist of web-scale image–text pairs or coarse scene descriptions lacking affordance-level annotations. As a result, VLMs struggle to map high-level commands (e.g., store the milk on the fridge's top shelf) into low-level actions (e.g., target position of the robot end-effector).

To address these challenges, we present the ***RoboAfford***, a large-scale dataset with *dense, affordance-aware annotations* for instruction-grounded manipulation. It contains 819,987 images and 1.9 million QA pairs, unifying object and spatial affordances to support interaction-centric learning. As shown in Fig. 1, ***RoboAfford*** aims to equip VLMs with three critical capabilities: object affordance recognition, object affordance prediction, and spatial affordance localization. This comprehensive dataset allows models to ground target objects through attributes such as count, color, size, and spatial relations. It also helps the model reason about object affordances, indicating where and how to interact with objects, and identifies spatial affordances to localize available space for placement. ***RoboAfford*** specifically addresses the challenge of translating high-level instructions into actionable positions by providing fine-grained 2D point annotations. Additionally, we introduce ***RoboAfford-Eval***, a rigorously annotated benchmark featuring 338 evaluation questions. RoboAfford-Eval encompasses three tasks: (1) object affordance recognition to assess grounding ability, (2) object affordance prediction to evaluate functional part understanding, and (3) spatial affordance localization to test free space detection. Extensive experiments show that fine-tuning on the RoboAfford dataset markedly enhances affordance prediction in robot manipulation, confirming the dataset's effectiveness.

Our contributions can be summarized as follows:

- We introduce ***RoboAfford***, a novel dataset that unifies object and spatial affordances for robot manipulation tasks. It comprises 1.9 million question answering pairs with precise 2D point annotations, focusing on three key capabilities: object affordance recognition, object affordance prediction, and spatial affordance localization. This comprehensive dataset fills the critical gap in fine-grained affordance annotations necessary for precise physical interactions.
- We present the ***RoboAfford-Eval*** benchmark, featuring 338 manually annotated questions designed to systematically evaluate VLMs on three core tasks: object affordance recognition, object affordance prediction, and spatial affordance localization. This benchmark establishes criteria to assess both object-level and scene-level affordance understanding.
- Extensive experiments reveal the limitations of existing VLMs in affordance learning, while fine-tuning on the ***RoboAfford*** dataset significantly enhances their affordance prediction in robot manipulation, confirming its effectiveness and contribution to the field.

Table 1: Comparison of existing affordance datasets. Obj-Aff: Object Affordance. Spa-Aff: Spatial Affordance.

Dataset	Domain	Format	Obj-Aff	Spa-Aff	#Spatial Relations	#Images	#QAs
UMD [27]	Tabletop	2D Mask	✓	✗	-	30,000	-
IIT-AFF [28]	Generic	2D Mask	✓	✗	-	8,835	-
PAD [25]	Generic	2D Mask	✓	✗	-	4,002	-
AGD-20K [26]	Generic	2D Heatmap	✓	✗	-	26,117	-
3DOI [30]	Generic, Indoor	2D Mask, Point	✓	✗	-	10K	-
RoboPoint [39]	Generic, tabletop	2D point	✗	✓	14	-	1.4M
RoboSpatial [32]	Indoor, tabletop	2D point	✗	✓	6	1M	3M
RoboAfford (Ours)	**Generic, Indoor, tabletop**	**2D point**	✓	✓	**20**	**820K**	**1.9M**

2 Related Work

Datasets for affordance learning Affordance learning is a key area in computer vision and robotics, focusing on understanding object and environment properties for interaction. Prior works [18, 20, 29] primarily aimed to identify functional object parts enabling specific interactions through semantic segmentation, using pixel-wise annotated datasets like UMD [27], IIT-AFF [28], PAD [25, 41], AGD-20K [26], and 3DOI [30]. Another approach utilizes videos to transfer learned affordances to target images via interactive heatmap predictions [2, 6]. Some works also explored 3D affordance datasets [12, 13, 21], modeling affordances as grasping contact maps and scene-aware motion synthesis. However, existing datasets often focus on either object-centric or scene-centric annotations, rarely unifying both, as shown in Table 1. This paper proposes the RoboAfford dataset, which bridges this gap by integrating interactive point predictions that combine functional part localization with language-grounded spatial affordances in real-world scenes.

Spatial reasoning with VLMs Spatial reasoning is crucial for robots to interact effectively with the physical world. Building on 2D VLMs, several works have extracted 3D spatial information from 2D images to generate large-scale, spatially-related question answering pairs for fine-tuning. For example, SpatialVLM [5] converts 2D images into object-centric 3D point clouds using metric depth estimation, synthesizing VQA data with 3D spatial reasoning supervision. SpatialRGPT [8] enhances region-level spatial reasoning through region proposals and 3D scene graph construction. RoboPoint [39] introduces a synthetic dataset for free space references, predicting points for precise action spaces. Recent studies like SpatialBot [4] and RoboSpatial [32] utilize both RGB and depth images to improve spatial understanding. Additionally, SpatialCoT [24] leverages language models to generate rationales for complex spatial reasoning. Despite these advancements, existing VLMs have limitations in spatial reasoning for real-world interactions. Their insufficient coverage of object functionality and spatial relations restricts accurate predictions of contact positions for affordances. This paper explores ways to enhance VLMs' spatial reasoning capabilities to improve robot interactions in the real world.

3 RoboAfford Dataset and Benchmark Construction

3.1 Data collection and Filtering

The RoboAfford dataset integrates annotations from five data sources, as shown in Table 2. By combining real-world data from the internet with synthetic data from robot manipulation, we create a diverse dataset for modeling object and spatial interactions. This dataset consists of three main components: Object Affordance Recognition, Object Affordance Prediction, and Spatial Affordance Localization.

Table 2: Data source information for RoboAfford dataset.

Data Source	#Selected Images	#Generated QA Pairs
LVIS [15]	152,152	513K
Pixmo-Points [11]	63,907	190K
PACO-LVIS [31]	45,790	561K
Object Reference [39]	287,956	347K
Region Reference [39]	270,182	320K

Object Affordance Recognition To equip VLMs with general object recognition knowledge, we utilize LVIS [15], which offers 2.2M high-quality instance segmentation masks across over 1,000 categories. We convert 152K images from LVIS into an object detection format with bounding box coordinates (x_1, y_1, x_2, y_2) to establish visual grounding. Additionally, we leverage the Pixmo-Points dataset [11], containing 2.3M point annotations from 223K images, along with 288K synthetic images from RoboPoint [39] for object reference learning. To address the issue of densely repeated object instances in Pixmo-Points, we implement a two-step filtering process: first, we discard annotations with more than 10 point labels for training simplicity, and second, we use GPT-4o [16] to retain only relevant indoor objects (e.g., furniture, kitchenware), resulting in 63,907 images suitable for object pointing.

Object Affordance Prediction For object affordance prediction, we utilize the PACO-LVIS dataset [31] to provide part-level annotations for reasoning. We extract bounding box and part segmentation masks from 45,790 images across 75 object categories and 200 part categories, transforming these into ground truth labels for object affordance. This structured data enables precise predictions of how objects can be interacted with based on their affordances.

Spatial Affordance Localization For spatial affordance localization, we source region reference data from RoboPoint [39], utilizing 270K images across 8K instances and 262 categories. Each image includes one or two colored bounding boxes to indicate objects, with ground truths formulated as series of points $[(x_1, y_1), (x_2, y_2), ...]$ for free space referencing. We convert normalized coordinates to absolute positions and sample ground truths to a maximum of ten points per answer, enhancing model optimization for spatial tasks.

3.2 Question Answering Pairs Generation

We create specialized question-answer generation pipelines for each task in RoboAfford, as shown in Fig. 2. By transforming the collected

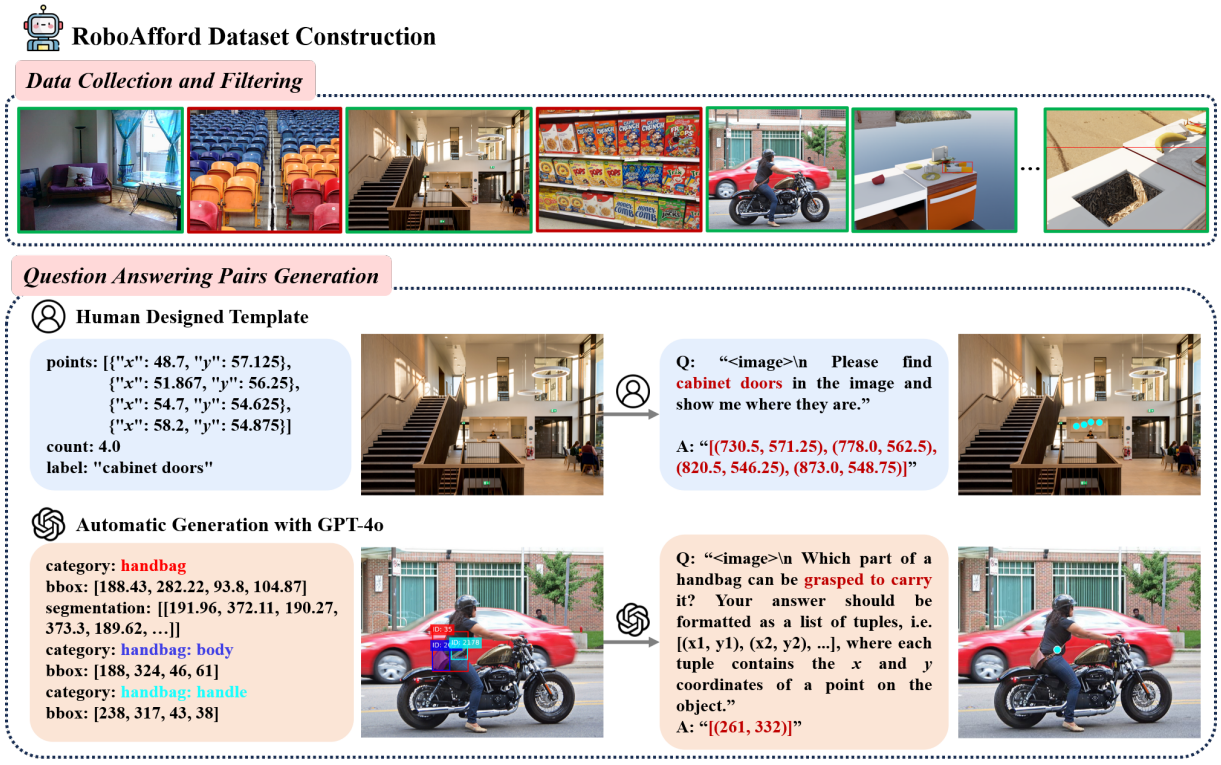

Figure 2: Pipeline for constructing the RoboAfford dataset. We first discard the image with densely repeated objects, and then generate question answering pairs using human designed template or GPT-4o.

data into affordance-aware QAs, we enhance the engagement of visual language models (VLMs) with the dataset, allowing them to learn and infer relationships between objects and their affordances.

• For object affordance recognition, we use GPT-4o [16] to analyze scenes and filter out irrelevant outdoor images. We create templates for generating questions and answers, such as "Point to all occurrences of <label> in the image" and "Can you see any <label> in the image? Point to them," where "<label>" refers to ground truths from Pixmo-Points [11]. We design 28 templates and randomly select one for each object pointing question.

• For object affordance prediction, we generate QAs by prompting GPT-4o with images, object and part categories, and ground truth bounding boxes. Questions for whole objects focus on functional use without naming the object (*e.g.*, "What appliance can be used to heat food quickly?" for a microwave). For object parts, we ask for part identification (*e.g.*, "Which part of a knife should be held to cut safely?" for the handle). Ground truth answers include two formats: (1) bounding boxes for the target object or part, and (2) points sampled from the ground truth segmentation mask. This dual representation ensures accurate part grounding and enhances fine-grained object affordance prediction.

• For spatial affordance localization, we generate QAs by modifying annotations from RoboPoint [39]. Specifically, we convert normalized coordinates to absolute positions for accurate real-world object scales and spatial relationships. Ground-truth points are resampled to a maximum of ten per question, and instructions are adjusted for answer format consistency. This method preserves the spatial relationships defined in RoboPoint while integrating them into our unified affordance framework.

3.3 RoboAfford-Eval Benchmark

To evaluate object affordance recognition and prediction, we manually annotated 114 questions for recognition and 124 for prediction using images from the Where2Place dataset [39]. For spatial affordances, we retained the original 100 questions but adjusted the prompts to shift predictions from normalized to absolute coordinates. Ground-truth for each question consists of one or more human-annotated polygon masks corresponding to the parts or instances in the answers.

For each predicted point, we check if it falls within the ground-truth masks. The accuracy for a question is the ratio of correctly located points to total predicted points, with overall accuracy being the average across all questions. To enforce stricter criteria, we penalize points outside the image boundaries, encouraging the model to better learn absolute interactive positions.

4 Methodology

Framework Using the RoboAfford dataset, we fine-tune a model named RoboAfford-Qwen. This model employs a multimodal architecture comprising a vision encoder, an MLP projector, a language tokenizer, and the Qwen2.5 LLM, as illustrated in Figure 3. The vision encoder extracts visual features from input images, which

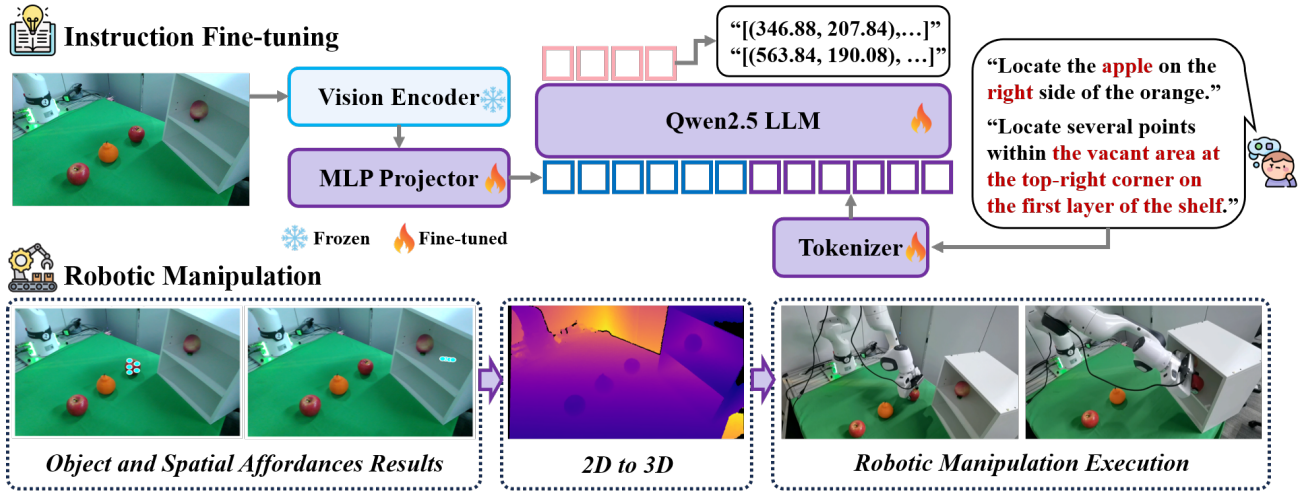

Figure 3: Our RoboAfford-Qwen Framework. We fine-tune the model on the RoboAfford dataset to enhance object and spatial affordance capabilities. For robotic manipulation, we use depth information to transform 2D points representing objects and spatial affordances into 3D coordinates, which are then converted to end-effector positions for robotic manipulation.

are transformed into the same embedding space as the language tokens through the MLP projector. These visual embeddings are concatenated with the embedded textual instructions and input to the LLM for joint reasoning across modalities.

Instruction Fine-tuning We employ a multi-stage training strategy based on the LLaVA-1.5 instruction tuning framework [23], consisting of two phases: General Localization Learning and Object-Spatial Affordance Enhancement. The first phase uses the LVIS and Pixmo-Points datasets, totaling 216K images and 703K QA pairs, to enhance basic object affordance recognition. The second phase incorporates Object Reference, PACO-LVIS, and Region Reference data, amounting to 604K images and 1.23M QA pairs, to optimize object affordance prediction and spatial affordance localization. This multi-stage approach, leveraging diverse datasets, allows the model to progressively develop hierarchical affordance reasoning, evolving from basic recognition to advanced prediction tasks.

Real-world Robotic Manipulation Fig. 3 shows that the fine-tuned RoboAfford-Qwen can be effectively applied to downstream robotic manipulation tasks. For the instruction "Place the apple in the vacant area of the shelf," RoboAfford-Qwen predicts the object's affordance for the specified apple and the spatial affordance of the shelf for valid placement. The predicted 2D affordance points are then transformed into 3D coordinates using a depth map, serving as target positions for robotic execution.

5 Experiments

5.1 Experimental Setup

Implementation Details. Our RoboAfford-Qwen model is initialized with the pre-trained Qwen2.5-VL-7B-Instruct weights and undergoes full-parameter supervised fine-tuning as described in [44]. The experiments are conducted on 8 H100 GPUs using AdamW as the optimizer, with a learning rate of 10^{-5} and a training duration of 1 epoch. Each device processes a batch size of 4, with gradient accumulation set to 2 steps.

Baseline Models. We evaluate a range of state-of-the-art vision-language models (VLMs), including both closed-source and open-source models, using the proposed RoboAfford-Eval benchmark. Closed-source models include GPT-4o [16], Claude-3.5-Sonnet [1], Gemini-2.5-Flash [9], and Gemini-2.5-Pro [10]. Open-source models encompass general-purpose VLMs such as LLaVA-Next [22], Molmo [11], Qwen2-VL [36], and Qwen2.5-VL [3]. We also evaluate VLMs with spatial awareness, including SpaceMantis (a community implementation of SpatialVLM [5]) and RoboPoint [39].

Evaluation Metrics. We evaluate the proposed RoboAfford-Eval benchmark across three tasks: object affordance recognition, object affordance prediction, and spatial affordance localization. The evaluation metric used is accuracy (Acc), defined as the ratio of correctly located predicted points within the ground truth masks to the total number of predicted points.

5.2 Quantitative Results

Tab. 3 presents a quantitative comparison of various baseline models on the RoboAfford-Eval benchmark. The zero-shot generalization ability of generic VLMs is limited across all object-spatial affordance settings, with top competitors like Gemini-2.5-Flash [9] and Gemini-2.5-Pro [10] achieving only 23.5 and 23.4 average accuracy, respectively. In contrast, specialized VLM with spatial reasoning capabilities such as RoboPoint [39], shows significantly improved performance, resulting in an average accuracy of 44.7. This underscores the critical role of spatial reasoning in affordance tasks.

Our fine-tuned Qwen2.5-VL-7B model on RoboAfford dataset (RoboAfford-Qwen) outperforms all baseline models, achieving accuracies of 66.1, 54.3, and 57.9 on three tasks, resulting in an average accuracy of 59.3. Compared to baseline model Qwen2.5-VL-7B [3], our model improves the accuracies by 46.6, 46.0, 36.0 on three tasks, respectively. It shows an improvement of 14.6 in average accuracy over the specialized VLM RoboPoint [39], demonstrating the effectiveness of the RoboAfford dataset in enhancing object-spatial affordance understanding.

Table 3: Comparison results of various VLMs on RoboAfford-Eval benchmark.

Type	Models	Parameters	Object Affordance Recognition↑	Object Affordance Prediction↑	Spatial Affordance Localization↑	Average↑
Closed-source Models	GPT-4o [16]	-	21.2	15.9	25.4	20.5
	Claude-3.5-Sonnet [1]	-	20.4	13.1	22.6	18.4
	Gemini-2.5-Flash [9]	-	20.4	21.7	29.4	23.5
	Gemini-2.5-Pro [10]	-	17.0	14.5	41.8	23.4
Open-source Models	Qwen2.5-VL [3]	3B	7.1	2.0	12.5	6.8
	Molmo [11]	7B	5.7	4.7	4.7	5.0
	Qwen2-VL [36]	7B	15.6	10.5	14.3	13.4
	LLaVA-Next [22]	8B	2.9	0.8	0.6	1.4
	SpaceMantis [5]	8B	3.6	4.8	12.0	6.5
	RoboPoint [39]	13B	55.7	35.0	44.2	44.7
	Qwen2.5-VL [3] (Baseline)	7B	19.5	8.3	21.9	16.1
	RoboAfford-Qwen (Ours)	**7B**	**66.1 (+46.6↑)**	**54.3 (+46.0↑)**	**57.9 (+36.0↑)**	**59.3 (+43.2↑)**

Figure 4: Qualitative results of RoboAfford-Qwen, where cyan points indicate the object and spatial affordances.

5.3 Qualitative Results

To intuitively demonstrate the capabilities of RoboAfford-Qwen, we present qualitative results in Fig. 4. The visualizations illustrate how RoboAfford-Qwen adapts to various real-world scenarios, perspectives, and tasks. The model effectively identifies object parts and aligns spatial semantics with task instructions, showcasing its understanding from object-level affordances to scene-level spatial reasoning. Additionally, we apply RoboAfford-Qwen to downstream manipulation tasks, as shown in Fig. 5. In these scenarios, the model provides useful visual cues that assist in real-world manipulation, highlighting its practical application in dynamic environments.

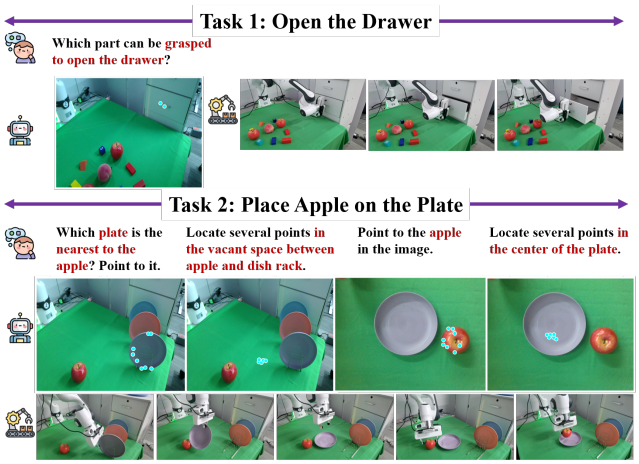

Figure 5: Results of deploying RoboAfford-Qwen to downstream robotic manipulation tasks.

6 Conclusion

In this paper, we introduce ***RoboAfford***, a large-scale dataset that integrates object and spatial affordance learning, comprising 1.9 million question-answer pairs with fine-grained 2D point annotations. This dataset enhances VLMs by improving their ability to reason about object grasping and spatial placement affordances. We also present ***RoboAfford-Eval***, a benchmark with 338 manually annotated questions to evaluate VLMs on three core tasks: object affordance recognition, object affordance prediction, and spatial affordance localization. Our findings highlight the limitations of current VLMs in affordance reasoning, while the model fine-tuning with RoboAfford yields significant performance improvements. This work advances affordance-aware learning for robotic manipulation, effectively bridging high-level reasoning with low-level interactions in real-world scenarios. Future efforts will expand ***RoboAfford*** to include a broader range of tasks, further enhancing VLM capabilities and supporting intelligent robotic systems development.

References

[1] Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. (2024).

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

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923* (2025).

[4] Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. 2024. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642* (2024).

[5] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14455–14465.

[6] Joya Chen, Difei Gao, Kevin Qinghong Lin, and Mike Zheng Shou. 2023. Affordance grounding from demonstration video to target image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6799–6808.

[7] Jiaqi Chen, Bingqian Lin, Xinmin Liu, Lin Ma, Xiaodan Liang, and Kwan-Yee K Wong. 2025. Affordances-oriented planning using foundation models for continuous vision-language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 23568–23576.

[8] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. SpatialRGPT: Grounded Spatial Reasoning in Vision Language Models. *arXiv preprint arXiv:2406.01584* (2024).

[9] Google Deepmind. 2025. Gemini 2.5 Flash Preview Model Card. (2025).

[10] Google Deepmind. 2025. Gemini 2.5 Pro Preview Model Card. (2025).

[11] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146* (2024).

[12] Alexandros Delitzas, Ayca Takmaz, Federico Tombari, Robert Sumner, Marc Pollefeys, and Francis Engelmann. 2024. SceneFun3D: fine-grained functionality and affordance understanding in 3D scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14531–14542.

[13] Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 2021. 3d affordancenet: A benchmark for visual object affordance understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1778–1787.

[14] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. 2023. Palm-e: An embodied multimodal language model. (2023).

[15] Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5356–5364.

[16] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

[17] Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, et al. 2025. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. *arXiv preprint arXiv:2502.21257* (2025).

[18] Dengyang Jiang, Mengmeng Wang, Teli Ma, Hengzhuang Li, Guang Dai, Lei Zhang, et al. 2025. AffordanceSAM: Segment Anything Once More in Affordance Grounding. *arXiv preprint arXiv:2504.15650* (2025).

[19] Dingzhe Li, Yixiang Jin, Yuhao Sun, Hongze Yu, Jun Shi, Xiaoshuai Hao, Peng Hao, Huaping Liu, Fuchun Sun, Jianwei Zhang, et al. 2024. What foundation models can bring for robot learning in manipulation: A survey. *arXiv preprint arXiv:2404.18201* (2024).

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

[21] Yicong Li, Na Zhao, Junbin Xiao, Chun Feng, Xiang Wang, and Tat-seng Chua. 2024. Laso: Language-guided affordance segmentation on 3d object. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14251–14260.

[22] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26296–26306.

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.

[24] Yuecheng Liu, Dafeng Chi, Shiguang Wu, Zhanguang Zhang, Yaochen Hu, Lingfeng Zhang, Yingxue Zhang, Shuang Wu, Tongtong Cao, Guowei Huang, et al. 2025. SpatialCoT: Advancing Spatial Reasoning through Coordinate Alignment and Chain-of-Thought for Embodied Task Planning. *arXiv preprint arXiv:2501.10074* (2025).

[25] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. 2021. One-shot affordance detection. *arXiv preprint arXiv:2106.14747* (2021).

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

[27] Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. 2015. Affordance detection of tool parts from geometric features. In *IEEE International Conference on Robotics and Automation*. 1374–1381.

[28] Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis. 2017. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*. 5908–5915.

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

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

[31] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. 2023. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7141–7151.

[32] Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. 2024. RoboSpatial: Teaching Spatial Understanding to 2D and 3D Vision-Language Models for Robotics. *arXiv preprint arXiv:2411.16537* (2024).

[33] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. 2025. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752* (2025).

[34] Yingbo Tang, Shuaike Zhang, Xiaoshuai Hao, Pengwei Wang, Jianlong Wu, Zhongyuan Wang, and Shanghang Zhang. 2025. Affordgrasp: In-context affordance reasoning for open-vocabulary task-oriented grasping in clutter. *arXiv preprint arXiv:2503.00778* (2025).

[35] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. 2025. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020* (2025).

[36] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* (2024).

[37] Yujie Wu, Huaihai Lyu, Yingbo Tang, Lingfeng Zhang, Zhihui Zhang, Wei Zhou, and Siqi Hao. 2025. Evaluating GPT-4o's Embodied Intelligence: A Comprehensive Empirical Study. *TechRxiv preprint techrxiv.174495686.69962588/v1* (2025).

[38] Rongtao Xu, Jian Zhang, Minghao Guo, Youpeng Wen, Haoting Yang, Min Lin, Jianzheng Huang, Zhe Li, Kaidong Zhang, Liqiong Wang, et al. 2025. A0: An Affordance-Aware Hierarchical Model for General Robotic Manipulation. *arXiv preprint arXiv:2504.12636* (2025).

[39] Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. 2024. Robopoint: A vision-language model for spatial affordance prediction for robotics. *arXiv preprint arXiv:2406.10721* (2024).

[40] Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. 2024. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693* (2024).

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

[42] Lingfeng Zhang, Xiaoshuai Hao, Qinwen Xu, Qiang Zhang, Xinyao Zhang, Pengwei Wang, Jing Zhang, Zhongyuan Wang, Shanghang Zhang, and Renjing Xu. 2025. Mapnav: A novel memory representation via annotated semantic maps for vlm-based vision-and-language navigation. *arXiv preprint arXiv:2502.13451* (2025).

[43] Wenqi Zhang, Mengna Wang, Gangao Liu, Xu Huixin, Yiwei Jiang, Yongliang Shen, Guiyang Hou, Zhe Zheng, Hang Zhang, Xin Li, et al. 2025. Embodied-reasoner: Synergizing visual search, reasoning, and action for embodied interactive tasks. *arXiv preprint arXiv:2503.21696* (2025).

[44] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

[45] Gengze Zhou, Yicong Hong, and Qi Wu. 2024. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 7641–7649.