Affordance-R1: Reinforcement Learning for Generalizable Affordance Reasoning in Multimodal Large Language Model

Hanqing Wang[1,8,*], Shaoyang Wang[2,*], Yiming Zhong[3], Zemin Yang[3], Jiamin Wang[3] Zhiqing Cui[5], Jiahao Yuan[4], Yifan Han[7], Mingyu Liu[6,8], Yuexin Ma[3,†]

[1]The Hong Kong University of Science and Technology (GZ), [2]National University of Singapore, [3]ShanghaiTech University
[4] East China Normal University, [5] Nanjing University of Information Science & Technology, [6]Zhejiang University
[7]Institute of Automation, Chinese Academy of Science, [8]Shanghai AI Laboratory
hwang201@connect.hkust-gz.edu.cn

Abstract

Affordance grounding focuses on predicting the specific regions of objects that are associated with the actions to be performed by robots. It plays a vital role in the fields of human-robot interaction, human-object interaction, embodied manipulation, and embodied perception. Existing models often neglect the affordance shared among different objects because they lack the Chain-of-Thought(CoT) reasoning abilities, limiting their out-of-domain (OOD) generalization and explicit reasoning capabilities. To address these challenges, we propose Affordance-R1, the first unified affordance grounding framework that integrates cognitive CoT guided Group Relative Policy Optimization (GRPO) within a reinforcement learning paradigm. Specifically, we designed a sophisticated affordance function, which contains format, perception, and cognition rewards to effectively guide optimization directions. Furthermore, we constructed a high-quality affordance-centric reasoning dataset, ReasonAff, to support training. Trained exclusively via reinforcement learning with GRPO and without explicit reasoning data, Affordance-R1 achieves robust zero-shot generalization and exhibits emergent test-time reasoning capabilities. Comprehensive experiments demonstrate that our model outperforms well-established methods and exhibits open-world generalization. To the best of our knowledge, Affordance-R1 is the first to integrate GRPO-based RL with reasoning into affordance reasoning. The code of our method and our dataset is released on *https://github.com/hq-King/Affordance-R1*.

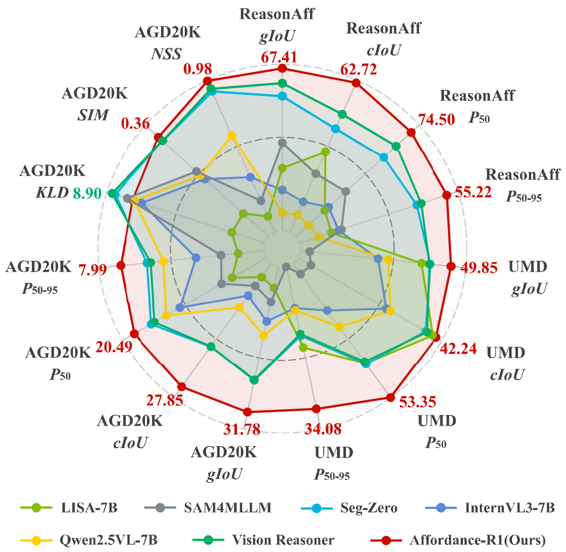


Figure 1: Affordance-R1 demonstrates extraordinary affordance reasoning ability and powerful generalization ability.

1 Introduction

Affordance is a crucial lens through which humans and embodied agents interact with various objects of the physical world, reflecting the possibility of where and how to act. Given an open-ended, complex, and implicit task instruction specified in natural language, affordance grounding aims to highlight the actionable possibilities of these objects, linking visual perception with robotic manipulation.

Recent efforts have made remarkable progress in affordance learning, such as extracting affordance knowledge from human-object-interaction (HOI) images (Yang et al. 2023; Wang et al. 2025b; Yang et al. 2024; Luo et al. 2024; Wang et al. 2025a; Rai, Buettner, and Kovashka 2024), human videos (Ma et al. 2025; Luo et al. 2023; Chen et al. 2023), and 3D perception modeling approaches such as object and scene point clouds (Deng et al. 2021; Chu et al. 2025; Nguyen et al. 2023; Delitzas et al. 2024) and 3D Gaussian Splatting (wei et al. 2025). However, these methods cannot actively reason about complex and implicit user intentions. Real-world physical interactions often require models to understand the human intention and reason about: *"What object can afford this? Why can this object afford such an affordance? Where is the affordance area?"*. Specifically, given a kitchen scene and the question "How would you reheat the food?", the model must reason deeply to identify that the oven can heat food and requires the "openable" affordance. This lack of affordance reasoning creates a gap in real-world applications. Some research (Yu et al. 2025; Qian et al. 2024) has utilized MLLM reasoning abilities to assist affordance grounding, but they only provide final affordance areas without the reasoning process—they cannot explain

*These authors contributed equally.
†Corresponding Author.

why an object affords such capabilities. To address this limitation, reinforcement learning offers a promising solution by enabling step-by-step reasoning through reward feedback, helping models understand both the answer and the reasoning process. Recent advances (OpenAI 2024; Guo et al. 2025; Liu et al. 2025a; Shen et al. 2025; Liu et al. 2025b; Huang et al. 2025) have demonstrated this capability through verifiable reward mechanisms. However, these models focus primarily on object-level reasoning and cannot handle embodied perception tasks requiring fine-grained analysis, such as affordance reasoning.

To fill this gap, we propose **Affordance-R1**, a reinforcement learning framework that enhances affordance grounding models with deep reasoning capabilities. We employ GRPO to fine-tune MLLMs without supervised training, investigating their self-evolution potential to develop reasoning abilities rather than relying on explicitly annotated processes. To closely link reasoning with affordance grounding, we design rewards from cognitive and perceptual perspectives: perception rewards and affordance recognition rewards. Inspired by *"Think twice before you act"*, we add a rethinking reward to help the model verify its reasoning process, addressing the transparency issue in current affordance models. Additionally, a box-num reward ensures the model outputs all possible affordance areas. Through these integrated rewards, Affordance-R1 achieves comprehensive reasoning at both perceptual and cognitive levels.

To facilitate such reasoning capabilities, existing datasets are insufficient for complex affordance reasoning. They are overly simplistic, lack real-world contextual complexity, and are specifically tailored for training visual segmentation models, making them unsuitable for MLLM instruction tuning.To address these limitations, we construct ReasonAff, a high-quality dataset with fine-grained affordance masks and reasoning-based implicit instructions that promote deep affordance understanding, specifically tailored for MLLM training. We utilize GPT-4o (Achiam et al. 2023) to construct the implicit instructions by providing it with an HOI image related to the affordance and the original instruction to help the agent better understand *"affordance"* and alleviate hallucination problems.

Through the synergy of our reinforcement learning framework and reasoning-oriented dataset, **Affordance-R1** demonstrates exceptional performance on both in-domain and out-of-domain data, which is crucial for real-world deployment. Furthermore, **Affordance-R1** maintains robust visual QA capabilities without the need for VQA training data. Experimental results show that **Affordance-R1** exhibits strong test-time reasoning capabilities and achieves superior generalization performance compared to models of the same scale. To summarize, our contributions are as follows:

- We introduce **Affordance-R1**, which is capable of generating explicit reasoning alongside the final answer. With the help of proposed affordance reasoning reward, which contains *format*, *perception*, and *affordance recognition* reward, it achieves robust zero-shot generalization and exhibits emergent test-time reasoning capabilities.
- We construct a high-quality affordance dataset **ReasonAff** for MLLM-based instruction-tuning, which is crucial for embodied perception and reasoning.
- We implement extensive experiments to demonstrate the effectiveness of our learning pipeline and observe noticeable gains over baselines with strong generalization capability, which highlights the effectiveness and adaptability of our approach in real-world applications.

2 Related Work

2.1 Affordance Learning

The concept of affordance was popularized by psychologist James Gibson (Gibson 1977), which reveals how embodied agents should interact with objects in dynamic, complex, and physical environments. Many researchers have made great efforts in affordance learning. Specifically, some works utilize affordance to link perception with robotic manipulations (Tang et al. 2025; Tong et al. 2024; Ma et al. 2025; Ju et al. 2024) and grasping (Wei et al. 2025; Zhang et al. 2023). Other studies, from a perceptual perspective, focus on endowing robots with an understanding of the affordance of objects and have explored numerous methods to obtain affordance knowledge from demonstrations, such as learning from HOI images (Yang et al. 2023; Gao et al. 2024; Shao et al. 2024a), human videos (Ma et al. 2025), and 3D perception modeling approaches including object (Deng et al. 2021; Qian et al. 2024; Yu et al. 2025; Chu et al. 2025; Nguyen et al. 2023) and scene (Delitzas et al. 2024) point clouds and 3DGS (wei et al. 2025). With the remarkable progress of LLMs, impressive reasoning capabilities have been demonstrated that can simulate human thinking. Some studies have explored how to transfer the inherent reasoning ability of LLMs to affordance learning. These works (Qian et al. 2024; Yu et al. 2025; Chu et al. 2025) adopt the strategy of introducing a special token into the vocabulary of LLMs and then utilize the embedding of this special token to perform affordance grounding. However, they still fail in generalization and cannot perform well when encountering OOD data, because they only establish a mapping between the affordance areas and the special token and cannot grasp general affordance knowledge. To address this issue, we utilize the GRPO (Shao et al. 2024b) algorithm to conduct a post-training process on the multimodal large language model, enabling the model to think and reason like humans to perform affordance perception.

2.2 Multimodal Large Language Models

MLLMs (Yang et al. 2025; Achiam et al. 2023) have made remarkable progress, which can achieve human-like or even superhuman intelligence in many aspects, such as visual understanding, generation, and multimodal reasoning. However, for many practical applications, such as segmentation and grounding, these models lack the necessary fine-grained perception required for detailed visual tasks. To address this issue, research efforts (Wang et al. 2024; Lan et al. 2024; Wu et al. 2024) enable the localization of specific regions within images by encoding spatial coordinates as tokens, improving the models' ability to reason about precise areas within the visual data. Moreover, OpenAI o1 (OpenAI 2024)

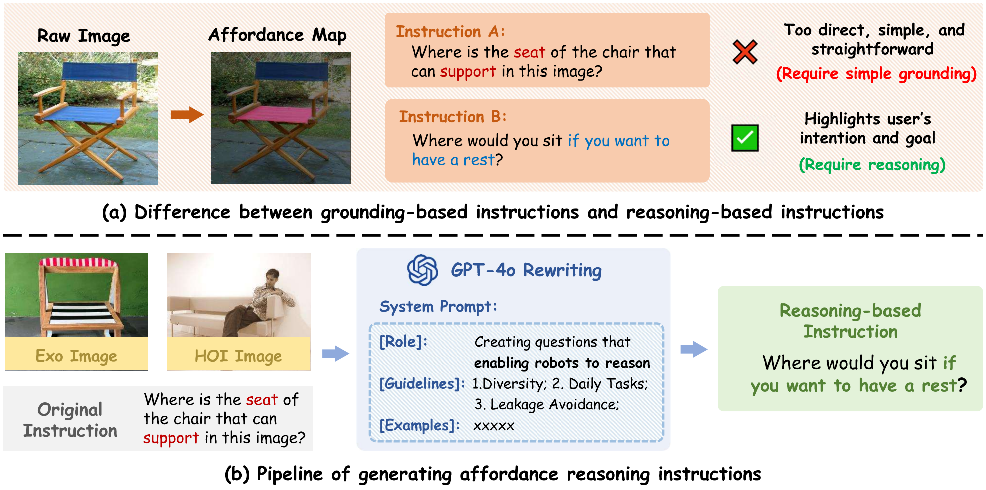


Figure 2: Affordance reasoning instruction generation and comparison. (a) Comparison between grounding-based and reasoning-based instructions. Instruction A directly asks for the faucet handle location (simple grounding), while Instruction B asks how to interact with the faucet to achieve opening (requires reasoning). (b) Pipeline for generating affordance reasoning instructions using GPT-4o to rewrite original instructions based on exo images, HOI images, and system prompts with guidelines for diversity, daily tasks, and leakage avoidance. The used prompt and statistical information of ReasonAff can be seen in our Appendix.

Dataset	#Object	#Aff	#Diversity	#Reasoning	#Q&A
UMD	17	7	✗	✗	✗
IIT-AFF	10	9	✗	✗	✗
ADE-Af	150	7	✗	✗	✗
PAD	72	31	✗	✗	✗
PADv2	103	39	✗	✗	✗
AGD20K	50	36	✗	✗	✗
InstructPart	48	30	✗	✗	✗
Ours	48	30	✓	✓	✓

Table 1: Comparison of Existing 2D Affordance Dateset with Ours. $\#Diversity$: diverse contextual instructions. $\#Obj$: number of object categories. $\#Aff$: number of affordance categories. $\#Q\&A$: Q&A instruction-tuning for MLLM.

introduces inference-time scaling by extending the Chain-of-Thought (CoT) reasoning process, significantly enhancing its multimodal reasoning performance. DeepSeek-R1 (Guo et al. 2025) further utilizes the GRPO (Shao et al. 2024b) algorithm to advance the reasoning ability, achieving superior performance with only a few thousand RL training steps. Several recent works (Shen et al. 2025; Liu et al. 2025a; Huang et al. 2025; Liu et al. 2025b; Song et al. 2025; Ouyang et al. 2025; Zhou et al. 2025; Pan et al. 2025; Zhang et al. 2025; Feng et al. 2025) have expanded this success into fine-grained visual tasks. However, these works primarily address high-level object reasoning and do not consider fine-grained part-level, especially affordance-level understanding.

Addressing this limitation, this paper aims to endow MLLMs with general affordance-aware perception by enabling them to interpret and interact with objects through reasoning in context-sensitive scenarios.

3 Dataset

Previous affordance-centric datasets fall short in supporting complex affordance reasoning. Moreover, these datasets are specifically designed for training visual segmentation models (e.g., SAM (Ravi et al. 2024)), making them difficult to seamlessly integrate into the instruction fine-tuning of multimodal large language models (MLLMs). As a result, models trained on such datasets tend to rely on grounding rather than in-depth reasoning. This prevents them from acquiring generalizable affordance knowledge, severely undermining their generalization capabilities.

Source	Results on AGD20K			Results on UMD			
	KLD↓	SIM↑	NSS↑	gIoU↑	cIoU↑	P_{50-95}↑	P_{50}↑
Instruct-Part	10.79	0.30	0.89	44.37	38.06	26.24	47.13
ReasonAff	**9.73**	**0.36**	**0.98**	**49.85**	**42.24**	**34.08**	**53.35**

Table 2: Evaluating Cross-Dataset Generalization for Affordance reasoning.

To better enhance the affordance grounding ability of MLLMs and improve their generalization performance, we have constructed the high-quality dataset ReasonAff, which can be utilized for MLLM instruction tuning. Specifically, we construct ReasonAff based on Instruct-Part (Wan et al. 2025). As shown in Figure 2 (b), we rewrite the instructions in the Instruct-Part dataset because we find the instructions are too direct and simple, and there are many sentences with **consistent structures** and many sentences are completely **identical**, which may limit the reasoning ability of the model. We utilize GPT-4o (Achiam et al. 2023) to rewrite the instructions by providing it with an HOI image related to the affordance and the original instruction to alleviate hallucination issues

Figure 3: Comparison of instructions and reasoning outputs between ReasonAff and Instruct-Part datasets on the same images.

and avoid identical instructions to enhance **diversity**. Specifically, for a given binary mask of affordance, we determine its bounding box $(x1, y1, x2, y2)$ by extracting the leftmost, topmost, rightmost, and bottommost pixel coordinates. Additionally, we compute the centroid of the mask as point coordinates (x_p, y_p). We show the comparison of ReasonAff with previous datasets in Table 1, and more dataset details are provided in the Appendix.

As can be seen in Figure 3, we present the different reasoning output (highlight areas) between the original Instruct-Part Affordance-related instructions and our reasoning-based instructions. Our implicit instructions based on reasoning can better enhance the reasoning ability of the model compared to previous instructions, enabling the model to learn more general affordance knowledge through the reasoning process and improve its generalization ability, as demonstrated by our experimental results shown in Table 2. The model trained on the reasoning-based **ReasonAff** dataset shows better performance and generalization on OOD datasets.

4 Affordance-R1 Framework

4.1 Overview

We provide an overview of our proposed method **Affordance-R1**. The task we address is a reasoning-based visual affordance grounding problem, where the model is tasked with localizing functional areas on objects based on implicit and complex instructions. Formally, given a textual instruction T and a target image I, the model $\mathcal{F}$ is expected to output the affordance area $\mathcal{A}ff$, defined as $\mathcal{A}ff = \mathcal{F}(T, I)$. Our method consists of two stages as shown in Figure 4. In the first stage, we directly employ rule-based reinforcement learning GRPO (Shao et al. 2024b) without SFT to enhance the model's inherent reasoning abilities. Additionally, we introduce a carefully designed affordance reward containing format, perception, and recognition components, to encourage the model to think and rethink about the image before providing final answers. In the second stage, we extract the output bounding boxes and points from **Affordance-R1**, which are then used as prompts for state-of-the-art segmentation models to produce fine-grained affordance masks.

4.2 Architecture

Following Seg-Zero (Liu et al. 2025a), **Affordance-R1** adopts a two-stage strategy comprising a reasoning model and a segmentation model. The overall architecture is illustrated in Figure 4. Specifically, given an image $\mathbf{I}$ and a high-level text instruction $\mathbf{T}$, **Affordance-R1** $\mathcal{F}$ generates an interpretable reasoning process and subsequently produces the expected output corresponding to $\mathbf{T}$. The model output is represented in a structured format, from which we extract the bounding boxes $\mathbf{B}$ and points $\mathbf{P}$ to serve as input to segmentation models such as SAM (Kirillov et al. 2023). This process can be formulated as follows:

$$(\{\mathbf{B}_i, \mathbf{P}_i\})_{i=1}^{N} = \mathcal{F}(\mathbf{I}, \mathbf{T}). \tag{1}$$

Subsequently, the affordance masks A_{ff} are predicted by the segmentation model $\mathcal{M}$ using the extracted bounding boxes $\mathbf{B}$ and points $\mathbf{P}$:

$$\mathbf{A}_i = \mathcal{M}(\mathbf{B}_i, \mathbf{P}_i). \tag{2}$$

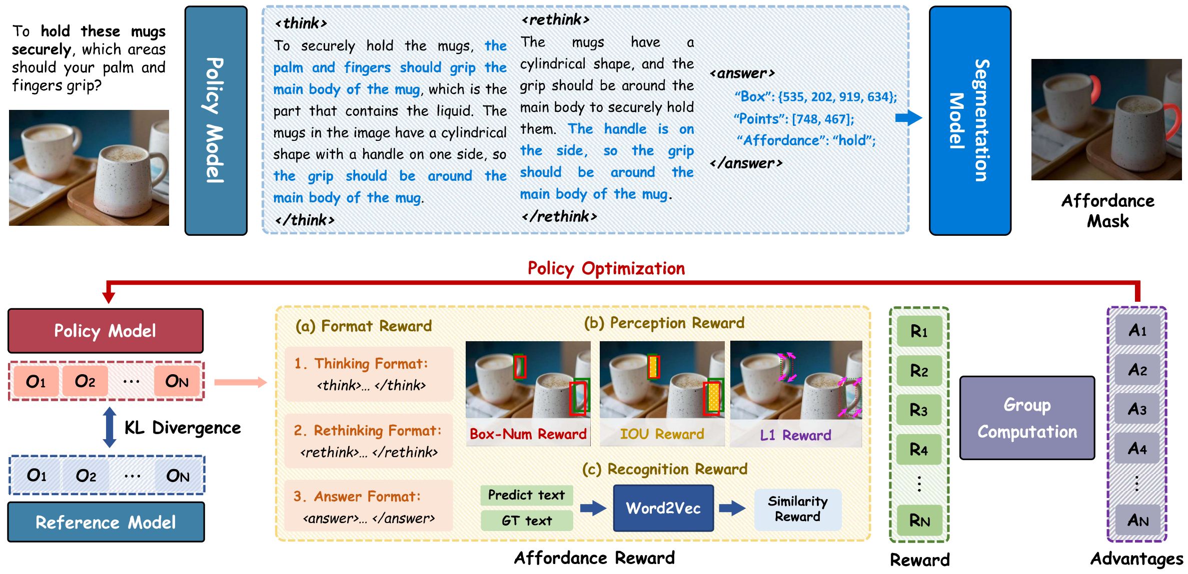


Figure 4: **Affordance-R1 framework overview**. The model processes queries through policy-based reasoning with $< think >$ and $< rethink >$ stages to generate affordance predictions. The policy optimization uses a sophisticated reward system comprising (a) format rewards for reasoning structure, (b) perception rewards for spatial accuracy (Box-Num, IOU, L1), and (c) recognition rewards for semantic similarity, enabling effective GRPO-based training for affordance reasoning.

4.3 Group Relative Policy Optimization (GRPO)

Unlike reinforcement learning algorithms such as PPO (Schulman et al. 2017), which require an additional critic model to estimate policy performance, GRPO (Shao et al. 2024b) directly compares groups of candidate responses, thereby eliminating the need for a separate critic network. Given a question q, GRPO (Shao et al. 2024b) samples N candidate responses $\{o_1, o_2, \ldots, o_N\}$ from the policy π_θ and evaluates each response o_i using a reward function $R(q, o_i)$, which quantifies the quality of the candidate response in the context of the given question. To determine the relative quality of these responses, GRPO (Shao et al. 2024b) normalizes the rewards by computing their mean and standard deviation, and subsequently derives the advantage as:

$$A_i = \frac{r_i - \text{mean}\{r_1, r_2, \ldots, r_N\}}{\text{std}\{r_1, r_2, \ldots, r_N\}}, \tag{3}$$

where A_i represents the advantage of candidate response o_i relative to other sampled responses within the group.GRPO (Shao et al. 2024b) encourages the model to generate responses with higher advantages by optimizing the policy π_θ through the following objective:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[\{o_i\}_{i=1}^{N} \sim \pi_{\theta_{old}}(q)] \tag{4}$$

$$\frac{\sum_{i=1}^{N} \{\min[s_1 A_i,\ s_2 A_i] - \beta \mathbb{D}_{KL}[\pi_\theta || \pi_{ref}]\}}{N} \tag{5}$$

$$s_1 = \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}; \quad s_2 = \text{clip}\left(s_1, 1+\epsilon, 1-\epsilon\right). \tag{6}$$

Reward Functions. As can be seen in Figure 4, we designed a sophisticated affordance reward system that contains *format, perception, and recognition* rewards to better guide the optimization of affordance reasoning.

Format Reward. We utilize the format reward to ensure the model's response strictly adheres to the required format. It can be divided into three parts: **1) Thinking Reward:** To force the model to think deeply before answering, we add the format $< think >$ *Thinking Process Here* $< /think >$ to constrain the model; **2) Rethinking Reward:** Inspired by the proverb: *"Think twice before you act"*, we add the rethinking reward $< rethink >$ *Rethinking Process Here* $< /rethink >$ to force the model to evaluate the thinking process itself, which double-checks the correctness of the reasoning process; **3) Answer Reward:** $< answer >$ *Final Answer Here* $< /answer >$.

Perception Reward. To help the model ground the affordance area, we utilize the perception reward, which mainly contains: **1) IoU Reward:** We calculate the Intersection over Union (IoU) between output bounding boxes and ground truth bounding boxes. If IoU > 0.5, the reward is 1; otherwise, the reward is 0; **2) L1 Reward:** We compute the L1 distance between output and ground truth bounding boxes (including points). If the L1 distance < 10, the reward is 1; otherwise, the reward is 0; **3) Box-Num Reward:** We introduce the box-num reward to ensure the model outputs all possible affordance areas.

Affordance Recognition Reward. As the ancient wisdom states, *"to know what it is and to know why it is"*, affordance reasoning requires not only perception but also recognition. Specifically, we use the word2vec model to calculate affordance text similarity. If similarity > 0.8, the reward is 1; otherwise, the reward is 0.

Model	LLM	Reasoning	gIoU↑	cIoU↑	P_{50-95}↑	P_{50}↑
VLPart	×	×	4.21	3.88	1.31	0.85
OVSeg	×	×	16.52	10.59	9.89	4.12
SAN	×	×	10.21	13.45	7.18	3.17
LISA-7B	✓	×	38.17	40.58	33.62	19.69
SAM4MLLM	✓	×	45.51	33.64	43.48	22.79
AffordanceLLM	✓	×	48.49	38.61	42.11	20.19
InternVL3-8B	✓	×	31.79	24.68	35.41	21.93
Qwen2.5VL-7B	✓	×	25.18	20.54	26.00	15.82
Seg-Zero	✓	✓	59.26	48.03	61.33	45.87
Vision Reasoner	✓	✓	63.04	52.70	67.33	47.23
Affordance-R1(Ours)	✓	✓	**67.41**	**62.72**	**74.50**	**55.22**

Table 3: Affordance reasoning comparison on ReasonAff.

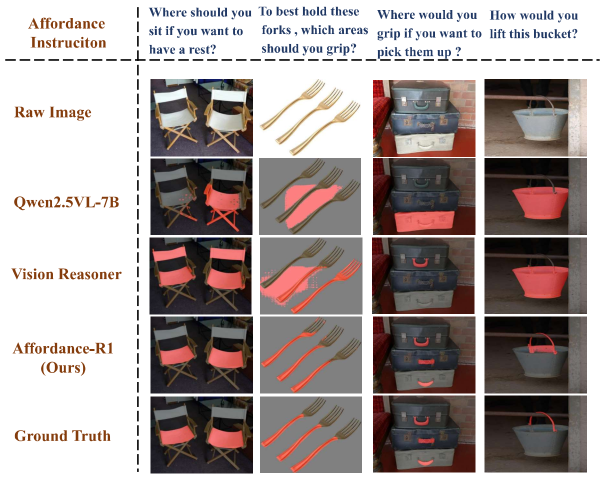


Figure 5: Qualitative Comparison of Affordance Reasoning

5 Experiment

This section provides a comprehensive evaluation of our proposed framework, **Affordance-R1**. We first describe the experimental settings, including datasets, baseline methods, evaluation metrics, and implementation details. Next, we present the quantitative analysis of the experimental results. Additionally, we conduct ablation studies to demonstrate the effectiveness of each component of our method.

5.1 Experimental Settings

Dataset and Out-of-Domain Datasets. As mentioned in Section 3, we construct a high-quality dataset **ReasonAff** based on the Instruct-Part (Wan et al. 2025) dataset. We train our model on this dataset, and to assess our model's generalization capability, we conduct experiments to evaluate its performance under OOD scenarios. Specifically, we leverage subsets from the UMD Part Affordance dataset (Myers et al. 2015) and AGD20K (Luo et al. 2022) as our OOD benchmarks for affordance task evaluation. For the UMD Part Affordance dataset (Myers et al. 2015), to better assess the zero-shot performance of different models, we select all objects from all categories. Since one in every three frames is manually annotated, we sample one-tenth of these annotated frames as our test split, resulting in a total of 1,922 test images. For the AGD20K (Luo et al. 2022) dataset, we use the test split of the *Seen* partition for zero-shot evaluation, which comprises 1,710 object-affordance pairs.

Model	Reasoning	gIoU↑	cIoU↑	P_{50-95}↑	P_{50}↑
LISA-7B	×	41.90	41.23	19.33	39.65
SAM4MLLM	×	12.40	8.41	0.05	4.12
AffordanceLLM	×	43.11	38.97	22.36	41.56
Qwen2.5VL-7B	×	33.21	29.83	10.45	25.17
InternVL3-7B	×	30.46	28.73	9.94	18.67
Seg-Zero	✓	44.26	39.30	16.53	39.93
Vision Reasoner	✓	44.00	39.71	16.10	39.04
Affordance-R1(Ours)	✓	**49.85**	**42.24**	**34.08**	**53.35**

Table 4: MLLM based zero-shot affordance reasoning comparison results on UMD dataset.

Baselines. For a thorough comparison, we evaluate our method against several representative baselines, including open-vocabulary segmentation methods such as VLPart (Sun et al. 2023), OVSeg (Liang et al. 2023), and SAN (Xu et al. 2023); and powerful open-source MLLMs such as LISA (Lai et al. 2024), SAM4MLLM (Chen et al. 2024), AffordanceLLM (Qian et al. 2024), Qwen2.5-VL (Bai et al. 2025), InternVL3 (Zhu et al. 2025), Seg-Zero (Liu et al. 2025a), and Vision Reasoner (Liu et al. 2025b) to compare their affordance reasoning capabilities with **Affordance-R1**.

Evaluation Metrics and Implementation Details. Following Instruct-Part, we use standard metrics gIoU, cIoU, Precision@50 (P@50), and Precision@50:95 (P@50:95). We employ Qwen2.5-VL-7B (Bai et al. 2025) and SAM2-Large (Ravi et al. 2024) as our default configuration. **Affordance-R1** is trained on a 4×A100 GPU server using the DeepSpeed library. During training, we use a total batch size of 8 with a sampling number of 8 per training step. The initial learning rate is set to 1e-6, the weight decay is 0.01, and the KL loss coefficient is set to 5e-3. The entire training process takes approximately 7 hours.

5.2 Quantitative Analysis

We conducted extensive experiments to comprehensively evaluate the affordance reasoning ability of **Affordance-R1**, including both in-domain and OOD datasets.

Results on ReasonAff. As presented in Table 3, **Affordance-R1** establishes a new SOTA on our **ReasonAff** benchmark, consistently outperforming all baseline methods across every evaluation metric. The performance gains are particularly pronounced on the high-precision metrics, P@50 and P@50:95, underscoring the high quality and accuracy. We show some qualitative comparison results of affordance reasoning in Figure 5. More results can be seen in Appendix.

We attribute this superior performance directly to our novel framework. Unlike conventional methods that rely on supervised fine-tuning, **Affordance-R1** leverages GRPO (Shao et al. 2024b) to unlock the MLLM's intrinsic reasoning capabilities. This approach is uniquely suited for the challenges posed by **ReasonAff**, which demands deep reasoning over implicit, complex, and real-world contextual instructions. The core of our success lies in the meticulously designed affordance reward function. Specifically, the Format Reward, which encourages a thinking and rethinking process, compels

Model	Reasoning	gIoU↑	cIoU↑	P_{50-95}↑	P_{50}↑	KLD↓	SIM↑	NSS↑
LISA-7B	×	13.18	11.96	1.45	5.31	13.68	0.16	0.46
SAM4MLLM	×	15.27	13.22	2.40	6.95	9.51	0.27	0.52
Qwen2.5VL-7B	×	20.28	16.35	5.61	15.49	9.81	0.26	0.77
InternVL3-7B	×	18.18	14.63	3.79	13.37	10.09	0.25	0.61
Seg-Zero	✓	26.99	22.01	6.52	17.82	9.02	0.35	0.94
Vision Reasoner	✓	26.98	21.98	6.31	17.31	**8.90**	0.35	0.95
Affordance-R1(Ours)	✓	**31.78**	**27.85**	**7.99**	**20.49**	9.73	**0.36**	**0 .98**

Table 5: MLLM based zero shot affordance reasoning comparison results on AGD20K dataset.

the model to build a coherent reasoning chain and self-correct before committing to an answer. This iterative refinement process, guided by the Perception and Affordance Recognition rewards, allows **Affordance-R1** to deconstruct complex problems and accurately ground abstract instructions to visual evidence, a capability where other baselines fall short.

Results on Out-of-Domain Datasets. To assess the generalization power of **Affordance-R1**, we performed a zero-shot evaluation on the AGD20K (Luo et al. 2022) and UMD (Myers et al. 2015) datasets. The results, summarized in Table 5 and Table 4, reveal that **Affordance-R1** maintains its significant performance edge, demonstrating exceptional generalization to unseen object types and visual domains. This strong generalization is a direct outcome of our methodology. By forgoing traditional SFT in favor of GRPO (Shao et al. 2024b), **Affordance-R1** learns a robust and generalizable policy for affordance reasoning, rather than merely memorizing patterns from the training data. The reinforcement learning process, guided by our comprehensive reward signals, teaches the model the fundamental principles of identifying functional regions based on reasoning. Consequently, this learned policy is less sensitive to domain-specific visual features and translates effectively to novel scenarios presented in OOD datasets. In contrast, competing models show a more significant performance drop, indicating a degree of overfitting to their training distributions and a weaker grasp of the underlying affordance concepts. This confirms that **Affordance-R1** learns a more fundamental and transferable understanding of object affordance.

Visualization Results on Web Image. To evaluate the generalization ability of **Affordance-R1**, we collect some kitchen and household scene pictures from the EPIC-KITCHENS dataset (Damen et al. 2018) and the internet. As can be seen in Figure 6, **Affordance-R1** can still maintain strong affordance reasoning ability and effectively handle complex scenarios. More results can be seen in Appendix.

5.3 Ablation Study Results

We conduct various ablation studies to assess the impact of different components on our model **Affordance-R1**'s performance, including the proposed rethinking reward, the affordance recognition reward, and the Box-Num reward.

Rethinking Reward. As ancient wisdom states: *"Think twice before you act"*. The results Table 6 demonstrate that the introduction of the rethinking reward can force the model to reconsider and re-examine the question and image, making it think twice before giving final answers, resulting in an improvement over the baseline.

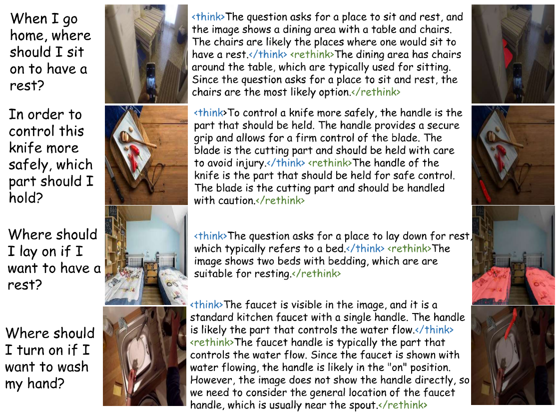


Figure 6: **Visualization on Web Image. Affordance-R1** can understand complex scenarios and shows well generalization.

Rethinking	Recognition	Box-Num	gIoU↑	cIoU↑	P_{50-95}↑	P_{50}↑
×	×	×	60.58	51.94	66.89	45.55
✓	×	×	63.04	56.33	67.02	51.55
✓	✓	×	65.25	61.22	68.33	50.07
✓	✓	✓	**67.41**	**62.72**	**74.50**	**55.22**

Table 6: **Ablation Study.** We investigate the improvement of Rethinking Reward and Affordance Reward on the model performance based on the baseline.

Affordance Recognition Reward. As the saying goes, *"to know what it is and to know why it is"*, affordance reasoning not only requires the model to know where the affordance area is but also the type of affordance this object affords. Table 6 presents the performance comparison with and without the affordance recognition reward. The model achieves better results when trained using the affordance recognition reward, which means the affordance recognition reward can help the model understand the concept of affordance and general affordance knowledge.

Box-Num Reward. As can be seen in Table 6, we conducted ablation experiments to study the influence of the box-num reward. We found that without this reward function, the model would tend to output a single affordance reasoning answer and ignore other possibilities, resulting in performance degradation.

6 Conclusion and Future Work

In this paper, we introduce the first affordance-centric reasoning model **Affordance-R1** and a high-quality affordance-centric reasoning dataset **ReasonAff**, which can be integrated into the instruction-tuning training process of multimodal large language models. With the help of the proposed sophisticated affordance reasoning reward function, we adopt pure reinforcement learning, specifically GRPO, to fine-tune the MLLM without supervised fine-tuning (SFT). **Affordance-R1** advances affordance reasoning by integrating LLM capabilities, enhancing the model's ability to handle complex and real-world contexts. It not only achieves state-of-the-art performance on **ReasonAff** but also shows superior generalization on out-of-domain datasets. For future work, we will explore how to utilize the excellent affordance reasoning abilities of **Affordance-R1** to construct an automatic data engine pipeline for affordance reasoning, thereby advancing the scaling law of embodied perception.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; et al. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923.

Chen, J.; Gao, D.; Lin, K. Q.; and Shou, M. Z. 2023. Affordance grounding from demonstration video to target image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6799–6808.

Chen, Y.-C.; Li, W.-H.; Sun, C.; Wang, Y.-C. F.; and Chen, C.-S. 2024. SAM4MLLM: Enhance Multi-Modal Large Language Model for Referring Expression Segmentation. In *European Conference on Computer Vision*, 323–340. Springer.

Chu, H.; Deng, X.; Chen, X.; Li, Y.; Hao, J.; and Nie, L. 2025. 3D-AffordanceLLM: Harnessing Large Language Models for Open-Vocabulary Affordance Detection in 3D Worlds. *arXiv preprint arXiv:2502.20041*.

Damen, D.; Doughty, H.; Farinella, G. M.; Fidler, S.; Furnari, A.; Kazakos, E.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; et al. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, 720–736.

Delitzas, A.; Takmaz, A.; Tombari, F.; Sumner, R.; Pollefeys, M.; and Engelmann, F. 2024. SceneFun3D: Fine-Grained Functionality and Affordance Understanding in 3D Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14531–14542.

Deng, S.; Xu, X.; Wu, C.; Chen, K.; and Jia, K. 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.

Feng, K.; Gong, K.; Li, B.; Guo, Z.; Wang, Y.; Peng, T.; Wu, J.; Zhang, X.; Wang, B.; and Yue, X. 2025. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*.

Gao, X.; Zhang, P.; Qu, D.; Wang, D.; Wang, Z.; Ding, Y.; Zhao, B.; and Li, X. 2024. Learning 2d invariant affordance knowledge for 3d affordance grounding. *arXiv preprint arXiv:2408.13024*.

Gibson, J. J. 1977. The theory of affordances. *Hilldale, USA*, 1(2): 67–82.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Huang, W.; Jia, B.; Zhai, Z.; Cao, S.; Ye, Z.; Zhao, F.; Xu, Z.; Hu, Y.; and Lin, S. 2025. Vision-R1: Incentivizing Reasoning Capability in Multimodal Large Language Models. *ArXiv*, abs/2503.06749.

Ju, Y.; Hu, K.; Zhang, G.; Zhang, G.; Jiang, M.; and Xu, H. 2024. Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation. In *European Conference on Computer Vision*, 222–239. Springer.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9579–9589.

Lan, M.; Chen, C.; Zhou, Y.; Xu, J.; Ke, Y.; Wang, X.; Feng, L.; and Zhang, W. 2024. Text4seg: Reimagining image segmentation as text generation. *arXiv preprint arXiv:2410.09855*.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7061–7070.

Liu, Y.; Peng, B.; Zhong, Z.; Yue, Z.; Lu, F.; Yu, B.; and Jia, J. 2025a. Seg-Zero: Reasoning-Chain Guided Segmentation via Cognitive Reinforcement. *ArXiv*, abs/2503.06520.

Liu, Y.; Qu, T.; Zhong, Z.; Peng, B.; Liu, S.; Yu, B.; and Jia, J. 2025b. VisionReasoner: Unified Visual Perception and Reasoning via Reinforcement Learning.

Luo, H.; Zhai, W.; Wang, J.; Cao, Y.; and Zha, Z.-J. 2024. Visual-Geometric Collaborative Guidance for Affordance Learning. *arXiv preprint arXiv:2410.11363*.

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

Luo, H.; Zhai, W.; Zhang, J.; Cao, Y.; and Tao, D. 2023. Learning visual affordance grounding from demonstration videos. *IEEE Transactions on Neural Networks and Learning Systems*.

Ma, T.; Zheng, J.; Wang, Z.; Gao, Z.; Zhou, J.; and Liang, J. 2025. GLOVER++: Unleashing the Potential of Affordance Learning from Human Behaviors for Robotic Manipulation. *arXiv preprint arXiv:2505.11865*.

Myers, A.; Teo, C. L.; Fermüller, C.; and Aloimonos, Y. 2015. Affordance detection of tool parts from geometric features. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 1374–1381.

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

OpenAI. 2024. OpenAI o1. https://openai.com/o1/.

Ouyang, R.; Li, H.; Zhang, Z.; Wang, X.; Zhu, Z.; Huang, G.; and Wang, X. 2025. Motion-R1: Chain-of-Thought Reasoning and Reinforcement Learning for Human Motion Generation. *arXiv preprint arXiv:2506.10353*.

Pan, J.; Liu, C.; Wu, J.; Liu, F.; Zhu, J.; Li, H. B.; Chen, C.; Ouyang, C.; and Rueckert, D. 2025. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*.

Qian, S.; Chen, W.; Bai, M.; Zhou, X.; Tu, Z.; and Li, L. E. 2024. Affordancellm: Grounding affordance from vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7587–7597.

Rai, A.; Buettner, K.; and Kovashka, A. 2024. Strategies to Leverage Foundational Model Knowledge in Object Affordance Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 1714–1723.

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

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shao, Y.; Zhai, W.; Yang, Y.; Luo, H.; Cao, Y.; and Zha, Z.-J. 2024a. GREAT: Geometry-Intention Collaborative Inference for Open-Vocabulary 3D Object Affordance Grounding. *arXiv preprint arXiv:2411.19626*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024b. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shen, H.; Liu, P.; Li, J.; Fang, C.; Ma, Y.; Liao, J.; Shen, Q.; Zhang, Z.; Zhao, K.; Zhang, Q.; Xu, R.; and Zhao, T. 2025. VLM-R1: A Stable and Generalizable R1-style Large Vision-Language Model. *ArXiv*, abs/2504.07615.

Song, Z.; Ouyang, G.; Li, M.; Ji, Y.; Wang, C.; Xu, Z.; Zhang, Z.; Zhang, X.; Jiang, Q.; Chen, Z.; et al. 2025. Maniplvm-r1: Reinforcement learning for reasoning in embodied manipulation with large vision-language models. *arXiv preprint arXiv:2505.16517*.

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

Tang, Y.; Huang, W.; Wang, Y.; Li, C.; Yuan, R.; Zhang, R.; Wu, J.; and Fei-Fei, L. 2025. UAD: Unsupervised Affordance Distillation for Generalization in Robotic Manipulation. *arXiv preprint arXiv:2506.09284*.

Tong, E.; Opipari, A.; Lewis, S.; Zeng, Z.; and Jenkins, O. C. 2024. OVAL-prompt: Open-vocabulary affordance localization for robot manipulation through LLM affordance-grounding. *arXiv preprint arXiv:2404.11000*.

Wan, Z.; Xie, Y.; Zhang, C.; Lin, Z.; Wang, Z.; Stepputtis, S.; Ramanan, D.; and Sycara, K. 2025. InstructPart: Task-Oriented Part Segmentation with Instruction Reasoning. *arXiv preprint arXiv:2505.18291*.

Wang, C.; Zhai, W.; Yang, Y.; Cao, Y.; and Zha, Z. 2025a. GRACE: Estimating Geometry-level 3D Human-Scene Contact from 2D Images. *arXiv preprint arXiv:2505.06575*.

Wang, H.; Zhang, Z.; Ji, K.; Liu, M.; Yin, W.; Chen, Y.; Liu, Z.; Zeng, X.; Gui, T.; and Zhang, H. 2025b. DAG: Unleash the Potential of Diffusion Model for Open-Vocabulary 3D Affordance Grounding. *arXiv preprint arXiv:2508.01651*.

Wang, W.; Chen, Z.; Chen, X.; Wu, J.; Zhu, X.; Zeng, G.; Luo, P.; Lu, T.; Zhou, J.; Qiao, Y.; et al. 2024. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36.

Wei, Y.-L.; Lin, M.; Lin, Y.; Jiang, J.-J.; Wu, X.-M.; Zeng, L.-A.; and Zheng, W.-S. 2025. Afforddexgrasp: Open-set language-guided dexterous grasp with generalizable-instructive affordance. *arXiv preprint arXiv:2503.07360*.

wei, Z.; Lin, J.; Liu, Y.; Chen, W.; Luo, J.; Li, G.; and Lin, L. 2025. 3DAffordSplat: Efficient Affordance Reasoning with 3D Gaussians. arXiv:2504.11218.

Wu, J.; Zhong, M.; Xing, S.; Lai, Z.; Liu, Z.; Chen, Z.; Wang, W.; Zhu, X.; Lu, L.; Lu, T.; et al. 2024. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. *Advances in Neural Information Processing Systems*, 37: 69925–69975.

Xu, M.; Zhang, Z.; Wei, F.; Hu, H.; and Bai, X. 2023. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2945–2954.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, Y.; Zhai, W.; Luo, H.; Cao, Y.; Luo, J.; and Zha, Z.-J. 2023. Grounding 3d object affordance from 2d interactions in images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10905–10915.

Yang, Y.; Zhai, W.; Wang, C.; Yu, C.; Cao, Y.; and Zha, Z.-J. 2024. Egochoir: Capturing 3d human-object interaction regions from egocentric views. *arXiv preprint arXiv:2405.13659*.

Yu, C.; Wang, H.; Shi, Y.; Luo, H.; Yang, S.; Yu, J.; and Wang, J. 2025. Seqafford: Sequential 3d affordance reasoning via multimodal large language model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1691–1701.

Zhang, X.; Wang, D.; Han, S.; Li, W.; Zhao, B.; Wang, Z.; Duan, X.; Fang, C.; Li, X.; and He, J. 2023. Affordance-driven next-best-view planning for robotic grasping. *arXiv preprint arXiv:2309.09556*.

Zhang, X.; Wen, S.; Wu, W.; and Huang, L. 2025. Tinyllava-video-r1: Towards smaller lmms for video reasoning. *arXiv preprint arXiv:2504.09641*.

Zhou, H.; Li, X.; Wang, R.; Cheng, M.; Zhou, T.; and Hsieh, C.-J. 2025. R1-Zero's" Aha Moment" in Visual Reasoning on a 2B Non-SFT Model. *arXiv preprint arXiv:2503.05132*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; et al. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. arXiv:2504.10479.

Supplementary Material

This document contains supplementary materials for our main paper. We provide further technical details and more qualitative examples to complement the findings presented in the main text. We hope this supplementary information will help readers better understand our approach and results.

The remainder of this supplementary material is organized as follows. In Section A, we provide the hardware specifications used in our experiments. In Section B, we list the hyperparameters employed. Section C presents the detailed prompts for training and inference. Section D shows more details about the dataset. Section E gives a detailed picture of the future affordance data engine. Section F presents more visualizations.

A Computational Resources

To ensure reproducibility, we provide detailed information on the computational resources used in our experiments. For all experiments, including training and inference, we used 4 NVIDIA RTX A800 GPUs. The base model we used is Qwen-2-VL-7B, consuming approximately 78GB of memory during operation.

B Hyperparameters

In Table 7, we present the hyperparameters used in our experiments.

Hyperparameter	Value
Batch Size	8
Experience	16
Tempture	1
KL	0.05
epoch	5
RL Steps	750
Optimizer	AdamW
Learning Rate	1.e-4
seed	42
max pixels	12845056
min pixels	3136
max response length	2048
weight decay	1.0e-2
max promptlength	1300

Table 7: Details of Hyperparameters

C Prompts

We have carefully designed prompts to enable MLLM to build datasets and complete affordance reasoning.

C.1 Prompts for Data Construction

Prompt for generating reasoning-based instructions. We hope that the model can provide complex instructions based on different contextual scenarios. We collect Human-Object-Interaction images to prompt the GPT-4o to relieve hallucination problems. The details of our prompt are shown in Figure 7.

Prompt for generating instructions

[Role]
You are an analytical assistant specializing in robotic affordance grounding. Your expertise is in creating questions that facilitate the training of robotic affordance grounding, enabling robots to reason about task execution, such as determining the appropriate part of an object to grasp.
[Task]
Description: You will be provided with the name of an object. The robot is expected to use this tool to perform a variety of everyday tasks. Along with the tool name, you will receive an affordance type that the object can afford, a human object interaction image which show that how people use this object and a list of questions that have already been generated for this tool.
[Guidelines]
Diversity: Aim for a wide range of tasks, ensuring that there is no overlap with previous ones.
Daily Tasks: questions should be common and representative of the ones encountered in daily life.
Leakage Avoidance: Ensure that the generated questions tasks can only afford the given affordance.
Examples: When cutting a rope with these scissors, which part of it should your palm touch?
If you want to lift the bag, at which point is your finger most likely to carry it?
Point out the areas on the microwave ideal for opening. How would you grasp the hat to best maintain its condition? If you want to boil water, at which points on the tap would you open the water valve?
Instruction: Use the HISTORY of generated tasks as a reference to ensure compliance with the diversity guideline. The output should be in the JSON format

Figure 7: Full prompt for generating reasoning instruction.

Prompt for affordance reasoning. In order to better stimulate the reasoning ability of the large model during training, we have carefully designed prompts to guide the model and provide a specific answer example to help the model understand the task. The details of our prompt are shown in Figure 8.

D Datasets

In this section, we give more details about the proposed dataset, ReasonAff.

D.1 Dataset Details

In order to more intuitively demonstrate the diversity and superiority of our proposed dataset, ReasonAff, we calculated the distribution of word frequency and instruction length for the data instruction. Figure 9 shows the word cloud of the generated instructions, and Figure 10 shows the comparison

```
Prompt for Affordance R1
QUESTION_TEMPLATE = \
        "Please answer \"{Question}\" with bboxs
and points." \
        "Analyze the functional properties of
specific parts of each object in the image and
carefully find all the part(s) that matches the
problem." \
        "Output the thinking process in <think>
</think>, rethinking process in <rethink> </rethink>
and final answer in <answer> </answer> tags." \
        "Output the bbox(es) and point(s) and
affordance tpye(s) inside the interested object(s)
in JSON format." \
        "i.e., <think> thinking process here
</think>" \
        "<rethink> rethinking process here
</rethink>" \
        "<answer>{Answer}</answer>"
Answer="[{\"bbox_2d\": [10,100,200,210],
\"point_2d\": [30,110], \"affordance\": \"hold\"},
{\"bbox_2d\": [225,296,706,786], \"point_2d\":
[302,410], \"affordance\": \"grasp\"}]"
```

Figure 8: Full prompt for affordance training and inference.

Figure 9: Word could of ReasonAff

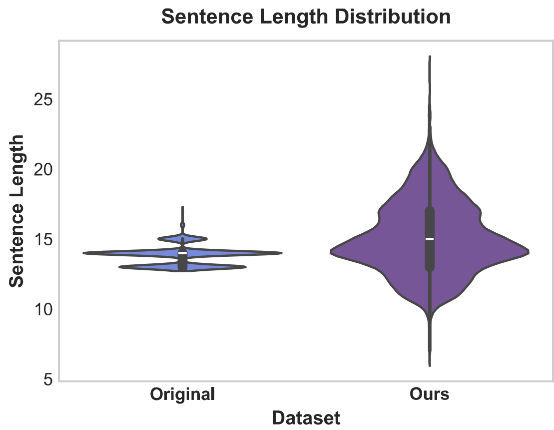


Figure 10: Comparison of violin plots of instruction length between raw data and Reasonaff

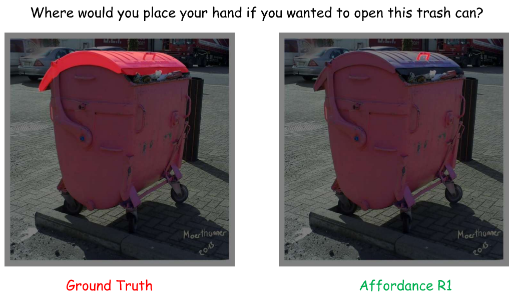


Figure 11: Ground truth is coarser in some cases, and Affordance R1 may predict a finer-grained affordance mask.

of violin plots of instruction length between raw data and Reasonaff. The superiority of our data is reflected in the more diverse reasoning-based affordance instructions, which take into account rich scene contexts and make MLLM more suitable for real-world scenarios

E Future work:Affordance Data Engine

Although ReasonAff provides a good source of affordance reasoning data, its scale and granularity are still insufficient. We built ReasonAff based on InstructParts, but it focuses more on part-level affordance. During the experiment, we found that its annotation of partial affordance data was coarser. As can be seen in Figure 11, in this case, the instruction is *Where would you place your hand if you wanted to open this trash can?* , the ground truth highlights the lid of the trash can, and our Affordance-R1 predicts more accurately. This enables us to utilize the excellent affordance reasoning ability of Affordance-R1 to make it an affordance data engine, which can predict the finer-grained affordance mask of the given, in-the-wild image. We hope to adopt the Affordance as the affordance mask generator, and an advancing VLM (e.g, GPT-4o) as the judge, inspired by the *MLLM-as-a-Judge* strategy. The judge VLM may not own such wonderful affordance grounding ability, but it is reasonable to adopt it as the **verifier** to filter the output of Affordance-R1. We would like to further explore this by utilizing the strategy of **Model-in-the-Loop** to construct a large-scale affordance reasoning dataset, which is suitable for MLLM to perform affordance reasoning instruction-tuning, to scale up the scaling law in embodied perception. As shown in Figure 12, we would use the filter data to further train the model, improving the affordance reasoning ability. The data may come from the internet or even data generated by generative models.

F More Visualization

In this section, we provide more visualization of Affordance-R1, the results are as follows

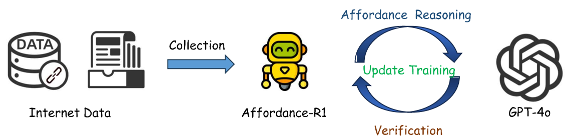
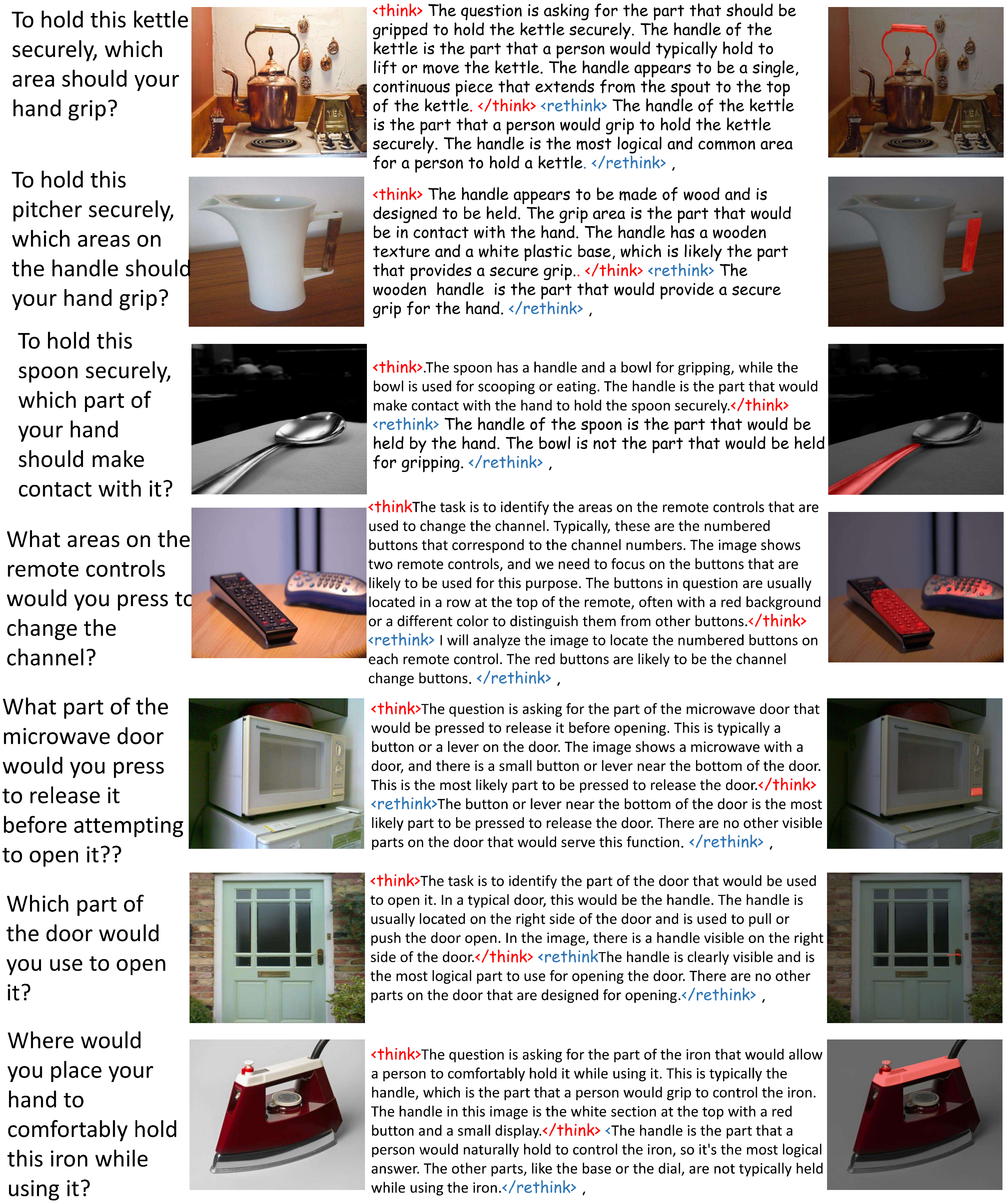
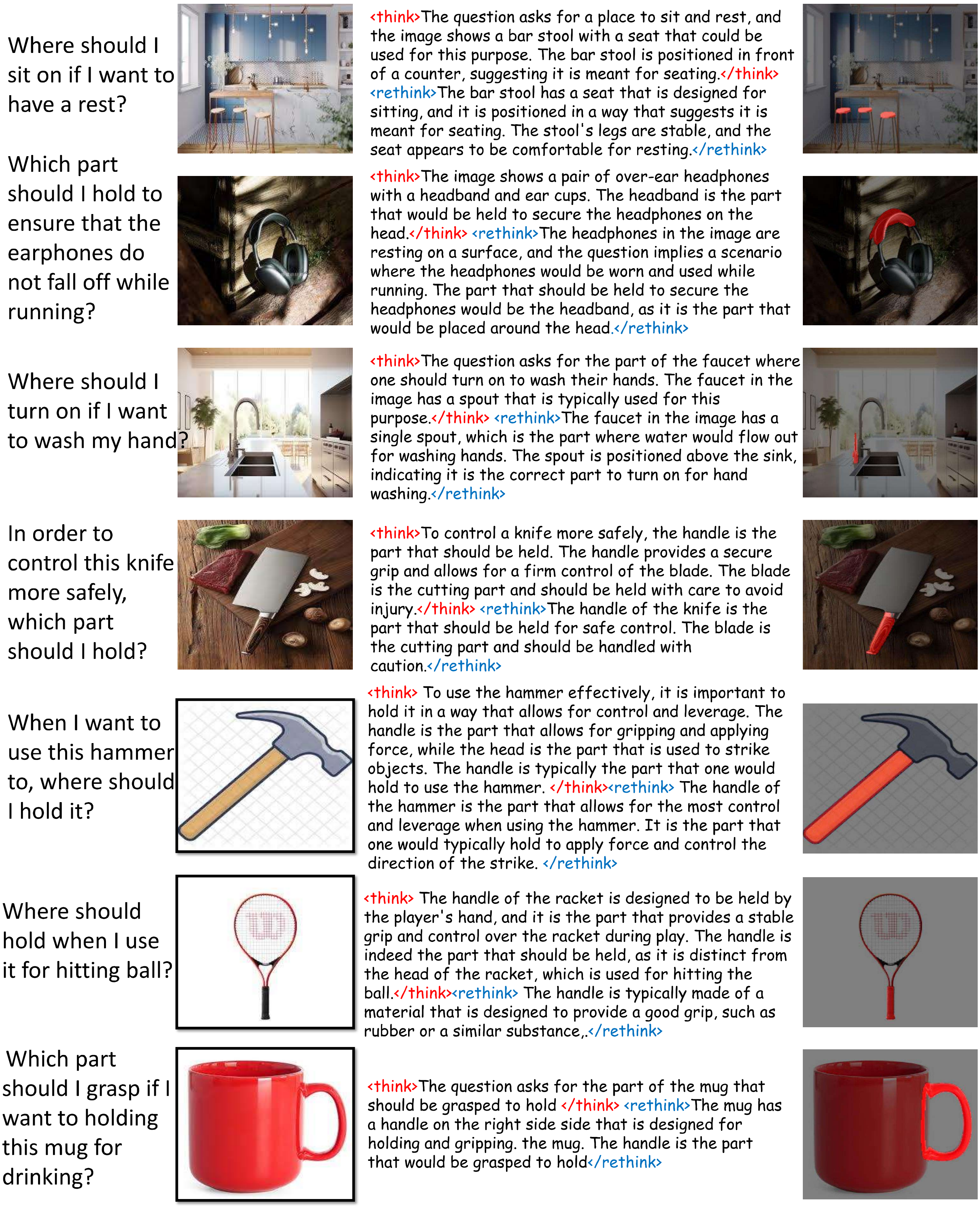


Figure 12: Model in the loop scaling up for affordance reasoning.



s

Figure 13: Visualization on ReasonAff. **Affordance-R1** can understand complex scenarios and shows well generalization.



s

Figure 14: Visualization on Web Image. **Affordance-R1** can understand complex scenarios and shows well generalization.