

A4-Agent: An Agentic Framework for Zero-Shot Affordance Reasoning

Zixin Zhang^{1,4*} Kanghao Chen^{1,4*} Hanqing Wang^{1*} Hongfei Zhang¹
 Harold H. Chen^{1,4} Chenfei Liao^{1,3} Litao Guo¹ Ying-Cong Chen^{1,2†}

¹HKUST(GZ) ²HKUST ³SJTU ⁴Knowin

*Equal contribution †Corresponding author

[\[Project Page\]](#) [\[Github Repo\]](#)

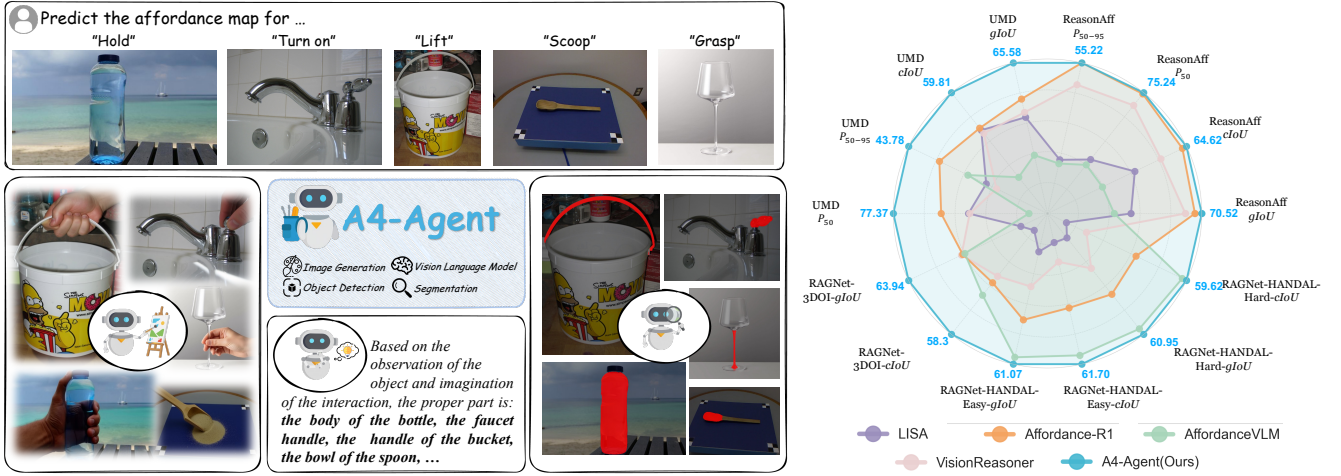


Figure 1. **Left:** Overview of **A4-Agent**, an affordance-centric vision-language agent that predicts actionable regions based on complex task instruction. Given an observed object, A4-Agent integrates image generation, object detection, segmentation, and a vision-language model to imagine plausible interactions and localize the proper action-specific part. **Right:** A4-Agent achieves state-of-the-art performance across multiple benchmarks with **zero-shot** setting, surpassing baseline models that are specifically trained for affordance prediction task.

Abstract

Affordance prediction, which identifies interaction regions on objects based on language instructions, is critical for embodied AI. Prevailing end-to-end models couple high-level reasoning and low-level grounding into a single monolithic pipeline and rely on training over annotated datasets, which leads to poor generalization on novel objects and unseen environments. In this paper, we move beyond this paradigm by proposing A4-Agent, a training-free agentic framework that decouples affordance prediction into a three-stage pipeline. Our framework coordinates specialized foundation models at test time: (1) a **Dreamer** that employs generative models to visualize how an interaction would look; (2) a **Thinker** that utilizes large vision-language models to decide what object part to interact with; and (3) a **Spotter** that orchestrates vision foundation models to precisely locate where the interaction area is. By lever-

aging the complementary strengths of pre-trained models without any task-specific fine-tuning, our zero-shot framework significantly outperforms state-of-the-art supervised methods across multiple benchmarks and demonstrates robust generalization to real-world settings.

1. Introduction

Affordance, a concept describing the action possibilities that objects offer to agents, serves as a crucial bridge between visual perception and physical interaction. In the context of embodied AI and robotic manipulation, affordance prediction aims to identify specific regions of objects that enable task-relevant interactions based on natural language instructions. For instance, given the instruction “open the refrigerator”, a model must recognize the handle as the actionable region. This capability is fun-

damental to downstream applications including task planning [21], robotic grasping [2, 17], and human-robot collaboration [4, 13], where understanding not just *what* objects are present, but *where* and *how* to interact with them becomes essential for successful task execution.

Affordance prediction fundamentally requires two complementary capabilities: ❶ **high-level reasoning**, interpreting natural language instructions and identifying task-relevant object parts, and ❷ **low-level grounding**, precisely localizing these parts in pixel coordinates. Traditional approaches [24, 30, 31] mainly focused on grounding, treating it as a regression problem: given an affordance type, the model predicts an affordance map. However, such approaches lack high-level reasoning capabilities and therefore struggle to handle complex instructions. More recent studies [36, 45, 49] attempt to incorporate large language models (LLMs) and trained unified models that perform both reasoning and grounding. By fine-tuning on affordance datasets, these models are endowed with the ability to output affordance maps. However, such tightly coupled designs introduce several issues, including a trade-off between reasoning and grounding, limited generalization, and reduced flexibility, which ultimately hinder their applicability in real-world scenarios. This leads us to question: *despite the appeal of end-to-end systems, is entangling high-level reasoning and low-level grounding truly the right path forward for affordance prediction?*

In this paper, we present a preliminary exploration, **A4-Agent**, an agentic framework tailored to affordance prediction through training-free coordination of foundation models. Our key insight lies in decoupling the reasoning and grounding processes. We decompose the task into a three-stage pipeline, with each stage managed by a specialized expert leveraging powerful foundation models: 1) **Dreamer**: Drawing inspiration from human cognitive processes, the Dreamer initiates an imagination phase. It employs generative models to synthesize visual scenarios depicting *how* an interaction would look (e.g., a hand grasping a handle, a door partially opening). 2) **Thinker**: The Thinker utilizes leading Vision-Language Models (VLMs) to interpret task instructions. Integrating visual observations with the imagined scenarios, it generates structured textual descriptions that specify *what* to interact with. 3) **Spotter**: The Spotter orchestrates robust vision foundation models to execute precise spatial localization, pinpointing exactly *where* the interaction area is within the visual input.

Remarkably, as shown in Fig. 1, by coordinating powerful pre-trained models without any task-specific training, our zero-shot framework A4-Agent significantly outperforms current state-of-the-art supervised methods across multiple benchmarks and demonstrates robust generalization to real-world settings. To summarize, our main contributions are as follows:

- We introduce A4-Agent, a training-free agentic framework that achieves superior performance and demonstrates strong zero-shot generalization capabilities.
- We validate a novel approach for affordance prediction by decoupling the reasoning and grounding processes. This allows for the integration of state-of-the-art models for each respective task, and we experimentally demonstrate the effectiveness of this method.
- We propose an Imagination-assisted affordance reasoning paradigm, showcasing the critical role of explicit imagination in the affordance reasoning process.

2. Related Work

Affordance Learning. The concept of affordance, introduced by Gibson [12], describes how agents perceive and interact with objects in their environment based on action possibilities. This foundational concept has inspired extensive research in affordance learning for robotic systems. Traditional approaches have explored various learning paradigms, including learning from human-object interaction (HOI) images [11, 39, 53], human demonstration videos [32], and 3D perception through point clouds [9, 10, 34, 36, 54] or 3D Gaussian Splatting [47].

Recent advances have leveraged multimodal large language models (MLLMs) to enhance affordance understanding. For example, AffordanceLLM [36] and Seqafford [54] introduce special tokens into the vocabulary and map affordance regions to token embeddings for segmentation outputs. LISA [22] extends this paradigm by incorporating reasoning capabilities for language-driven segmentation tasks. More recently, Affordance-R1 [45] employs reinforcement learning to enhance affordance reasoning and bounding box and key point grounding in MLLMs through process rewards. However, most of these methods adopt an end-to-end training paradigm that jointly optimizes reasoning and grounding capabilities within a single model architecture. They often face inherent trade-offs between reasoning complexity and spatial precision and exhibit poor generalization to novel scenarios. In contrast, our method proposes a training-free agentic framework that coordinates foundation models to achieve zero-shot affordance prediction through explicit reasoning and grounding.

Multimodal Reasoning in MLLMs. MLLMs [1, 26, 52] have demonstrated remarkable capabilities in visual understanding, generation, and multimodal reasoning. Recent advances have significantly enhanced their reasoning abilities through inference-time scaling. OpenAI o1 [35] extends the Chain-of-Thought (CoT) [46] reasoning process to achieve superior performance, while DeepSeek-R1 [14] leverages reinforcement learning with GRPO [40] to further advance reasoning capabilities. Building on these successes, several works [18, 28, 41] have expanded these rea-

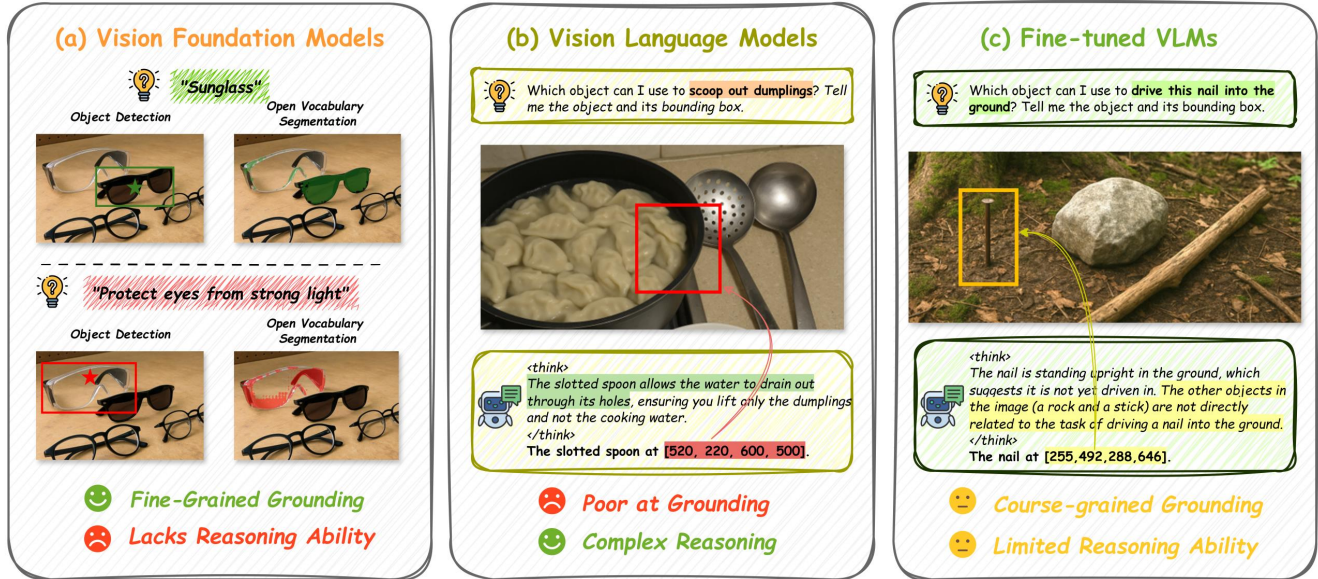


Figure 2. Vision Foundation Models are good at fine-grained grounding, but are poor at reasoning. Vision Language Models are good at reasoning, but are poor at visual grounding. Some works finetuned VLMs for better grounding ability, but both abilities are underwhelming.

soning paradigms to vision tasks, demonstrating the potential of enhanced reasoning in multimodal contexts.

Beyond text-based reasoning, emerging paradigms have explored reasoning with visual representations. VoT [50] introduces textual imagery representations for dynamic reasoning. Benefiting from powerful generative models [5, 15, 43, 55, 59], approaches [6, 8, 16, 23] attempt to leverage explicit visual imagination to assist in reasoning. These approaches show that generating intermediate visuals enhances reasoning and interpretability, offering valuable insights for affordance reasoning which demands complex spatial and interaction understanding. Unlike existing end-to-end methods for affordance prediction, our agentic framework decouples reasoning from grounding, allowing for the seamless integration of these multimodal reasoning techniques.

3. Motivation

Affordance prediction is a task that fundamentally requires two complementary capabilities: *high-level reasoning* for interpreting instructions with object parts, and *low-level grounding* for precisely localizing. As illustrated in Fig. 2 (a), while specialized vision foundation models excel at fine-grained localization, they lack the semantic understanding required to interpret complex task instructions. Conversely, as illustrated in Fig. 2 (b), while recent MLLMs demonstrate impressive reasoning capabilities, they often produce coarse or inaccurate spatial predictions, rendering them insufficient for precise affordance prediction.

Existing paradigm attempts to solve this dichotomy

through monolithic end-to-end models. These approaches try to enhance reasoning models’ grounding abilities [28, 29, 45] through training MLLMs on visual grounding data (e.g., bounding boxes, key points, masks). However, as Fig. 2 (c) shows, this tightly-coupled paradigm is less than ideal. They still introduces fundamental limitations: ❶ **Limited generalization**: training on limited datasets cannot cover the diversity of real-world scenarios, leading to brittleness on novel objects and environments; ❷ **Capability trade-offs**: optimizing for both reasoning and grounding simultaneously forces the model to balance different objectives, where improvements in one capability may degrade the other; ❸ **Poor flexibility**: the monolithic design prevents independent upgrades when more powerful foundation models emerge, requiring costly retraining of the entire system; and ❹ **Gap to closed-source models**: as these pipelines are restricted to open-source checkpoints, they cannot leverage the most capable closed-source models, thereby limiting the ceiling of reasoning ability.

Therefore, our work aims at exploring a fundamentally different approach: *decouple reasoning and grounding into specialized, coordinated agents*. We argue that affordance prediction is inherently multi-stage. Rather than forcing a single model to master both capabilities, we design each component independently using state-of-the-art foundation models and orchestrate them through an agentic framework at test time.

This paradigm shift can offer compelling advantages: (I) **Training-free generalization**, by leveraging pre-trained models’ broad knowledge, the system generalizes to diverse scenarios without task-specific fine-tuning or expen-

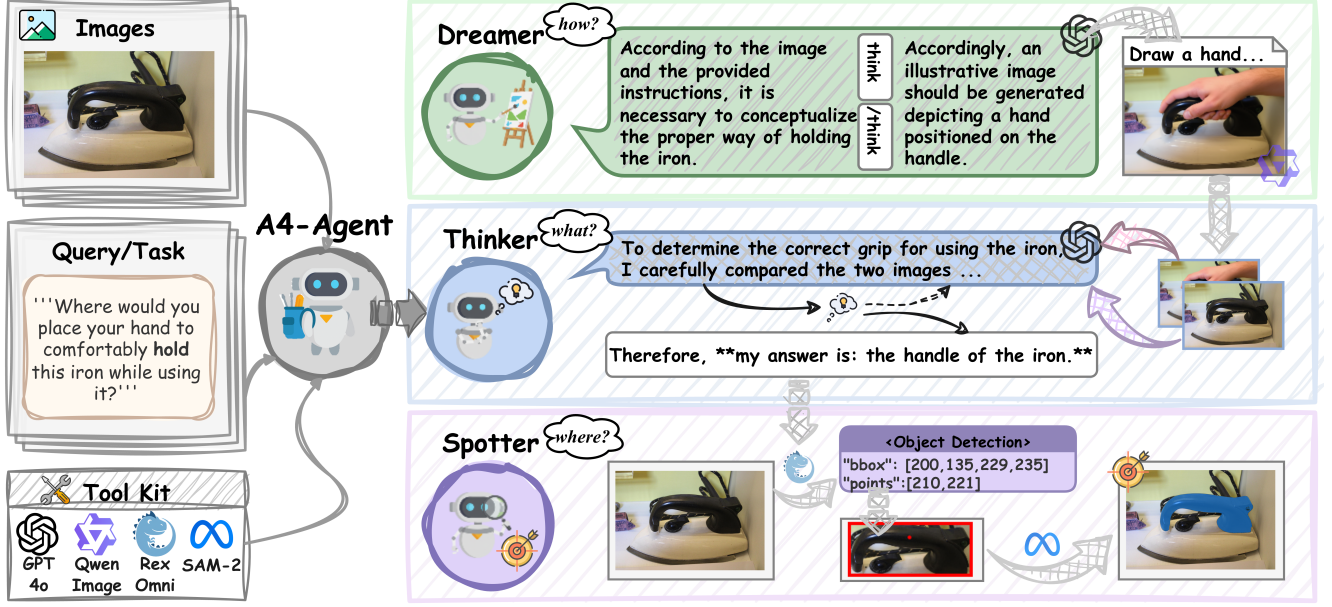


Figure 3. The pipeline of our A4-Agent framework, which decouples affordance prediction into three stages. (1) **Dreamer**: Imagines the interaction by generating a simulated image. (2) **Thinker**: Reasons over the original and simulated images to produce a textual description of the actionable object part. (3) **Spotter**: Takes this description to locate the part with bounding boxes and keypoints, then refines them into a precise segmentation mask.

sive data collection; (II) **Modular specialization**, each component exploits the complementary strengths of different models and can be independently upgraded as better models become available; and (III) **Interpretable reasoning**, explicit intermediate steps make the decision-making process transparent and debuggable, facilitating error diagnosis and system refinement.

4. A4-Agent: Agentic Affordance Reasoning

4.1. Problem Definition

We formulate affordance prediction as a visual grounding problem conditioned on natural language instructions. Given an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ along with a task description \mathbf{T} (e.g., “open the refrigerator”), the objective is to identify the affordance region \mathcal{A}_{ff} that enables the specified interaction:

$$\mathcal{A}_{\text{ff}} = \mathcal{F}(\mathbf{I}, \mathbf{T}), \quad (1)$$

where \mathcal{A}_{ff} denotes the spatial region(s) where task-relevant interactions occur. Depending on downstream applications, this region can be represented as bounding boxes $\{\mathbf{B}_i\}_{i=1}^N$, key points $\{\mathbf{P}_i\}_{i=1}^N$, or segmentation masks $\{\mathbf{M}_i\}_{i=1}^N$. Following recent work [45, 49], we adopt segmentation masks as the primary representation for their pixel-level precision.

4.2. Framework Overview

Building on the motivation outlined in Sec. 3, we introduce A4-Agent, a training-free, agentic framework for zero-

shot affordance prediction that implements the decoupling principle. Unlike end-to-end models that directly regress (\mathbf{B}, \mathbf{M}) from (\mathbf{I}, \mathbf{T}) , A4-Agent first infers which object part requires interaction (reasoning) and then determines its location (grounding):

$$\mathcal{A}_{\text{ff}} = \text{Ground}(\text{Reason}(\mathbf{I}, \mathbf{T})). \quad (2)$$

Specifically, the reasoning process follows a two-step pipeline: a Dreamer, which imagines *how* the operation can be (Sec. 4.3), and a Thinker, which decides *what* part to the operation (Sec. 4.4). The grounding process is then handled by the Spotter to locate *where* to operate using a coarse-to-fine approach (Sec. 4.5): it initially identifies broad regions via bounding boxes and key points, which are then refined by a segmentation model to produce pixel-accurate masks. Overall framework is shown in Fig. 3.

4.3. Dreamer: Imagine *how* to Operate

When humans reason about the affordances of a tool, they often begin by mentally simulating how the hand would interact with the tool and envisioning the broader usage scenario. Inspired by this process, we designed our Dreamer: rather than relying solely on text-based reasoning for affordance prediction, we first prompt the agent to use an image-generation module to visualize a plausible interaction state (e.g., a hand grasping a handle, a door being opened) based on the observation \mathbf{I} and task \mathbf{T} .

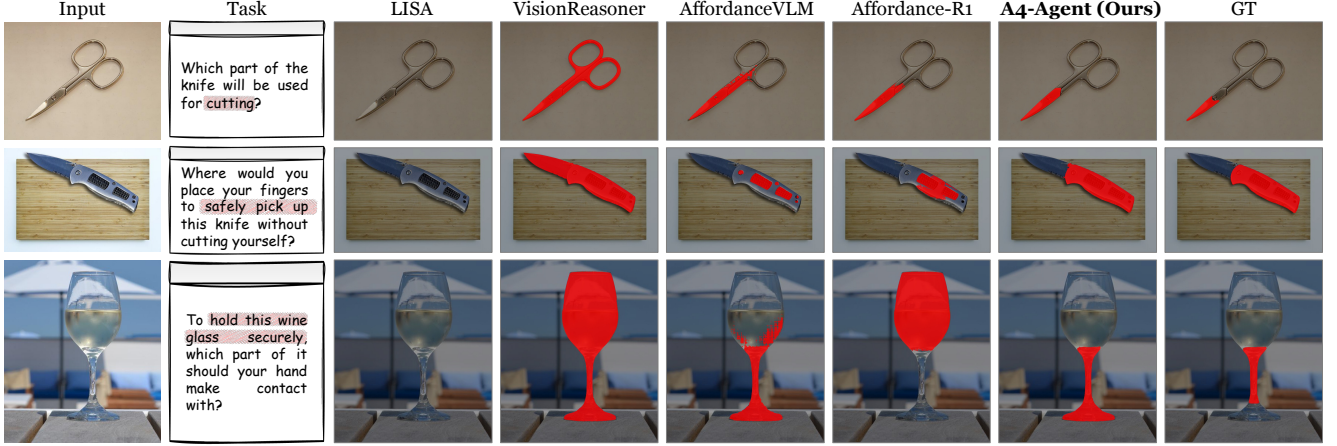


Figure 4. Qualitative comparison on ReasonAff dataset. Our method continuously predicts appropriate components according to task requirements, achieving results most consistent with ground truth and even surpassing Affordance-R1 specifically trained on this dataset.

To construct the image editing prompt that drives this imagination step, we query a VLM with an instruction template applied to the pair (\mathbf{I}, \mathbf{T}) . Formally,

$$\mathbf{T}_{\text{sim}} = \Phi_{\text{VLM}}(\mathbf{I}, \mathbf{T}; \tau), \quad (3)$$

where Φ_{VLM} denotes the VLM and τ is our instruction template, which is detailed in the Appendix. This template asks the model to output a short, visually actionable description that: (i) names the target object and the functional part visible in \mathbf{I} ; (ii) specifies the minimal interaction and contact configuration (e.g., “a right hand grasping the vertical refrigerator handle”); and (iii) avoids attributes not supported by the image. This yields concise prompts suitable for image editing and robust across varied scenes. Following, we employ a generative model [48] to synthesize interaction scenarios. Given the original image \mathbf{I} and a simulation prompt \mathbf{T}_{sim} derived from the task instruction (e.g., “a hand grasping the refrigerator handle”), the model produces an edited image \mathbf{I}_{sim} :

$$\mathbf{I}_{\text{sim}} = \mathcal{G}(\mathbf{I}, \mathbf{T}_{\text{sim}}), \quad (4)$$

where \mathcal{G} denotes the image generation model. The imagined image \mathbf{I}_{sim} explicitly highlights where interaction should occur by depicting plausible contact and motion cues and can further guides the agent in evaluating whether the action pattern is reasonable, thereby improving both the success rate and interpretability of affordance reasoning. This process can fully leverage the priors of the generative model, utilizing its understanding of interaction states to aid the reasoning model’s reasoning process. Such seamless integration is made possible by our agentic framework.

4.4. Thinker: Decide *what* to Operate

The next step is to reason through the appropriate interactive areas in textual form. Given the original image \mathbf{I} , the

imagined interaction image \mathbf{I}_{sim} , and the task \mathbf{T} , we prompt VLM with a preset template (see Appendix for the exact prompt) to perform three steps: (1) perceive key components and candidate interaction points in \mathbf{I} ; (2) consult \mathbf{I}_{sim} to infer contact/motion cues consistent with the affordance; (3) ground the actionable part back in \mathbf{I} and return a compact, machine-readable specification.

The VLM returns two sections—*Thinking* (free-form rationale) and *Output* (a machine-readable JSON). We ignore the *Thinking* section and parse only the *Output* JSON with three fields: “task”, “object_name”, and “object_part”. The object_part is phrased as “the [object part] of the [object name]” (e.g., “the blade of the shears”). This yields a concise textual affordance description \mathbf{D} specifying *what* to interact with, without any spatial coordinates. This design reduces variance via explicit instruction-following, keeps the reasoning trace interpretable, and preserves modularity—stronger VLMs can be swapped in without retraining.

4.5. Spotter: Locate *where* to Operate

The Spotter translates semantic affordance descriptions from the reasoning process into precise pixel-level localizations. Given a textual description \mathbf{D} (e.g., “handle on the right refrigerator door”), we employ two complementary vision foundation models to achieve coarse-to-fine spatial grounding: an open-vocabulary detector for initial region identification, followed by a segmentation model for pixel-accurate mask refinement. This two-stage approach is motivated by the complementary strengths of existing foundation models: while segmentation models excel at producing precise boundaries, they require reliable visual prompts (e.g., boxes or points) rather than text, necessitating an initial detection step to bridge the semantic-geometric gap.

Table 1. Quantitative results on ReasonAff. **A4-Agent achieves SOTA performance in zero-shot manner without any training.**

Model	LLM	Reasoning	gIoU \uparrow	cIoU \uparrow	$P_{50}\uparrow$	$P_{50-95}\uparrow$
VLPart [42]	✗	✗	4.21	3.88	1.31	0.85
OVSeg [25]	✗	✗	16.52	10.59	9.89	4.12
SAN [51]	✗	✗	10.21	13.45	7.18	3.17
LISA-7B [22]	✓	✗	38.17	40.58	33.62	19.69
SAM4MLLM [7]	✓	✗	45.51	33.64	43.48	22.79
AffordanceLLM [36]	✓	✗	48.49	38.61	42.11	20.19
InternVL3-8B [58]	✓	✓	31.79	24.68	35.41	21.93
Qwen2.5VL-7B [3]	✓	✓	25.18	20.54	26.00	15.82
AffordanceVLM [49]	✓	✓	30.50	25.54	30.29	18.31
Seg-Zero [28]	✓	✓	59.26	48.03	61.33	45.87
Vision Reasoner [29]	✓	✓	63.04	52.70	67.33	47.23
Affordance-R1 [45]	✓	✓	67.41	62.72	74.50	55.22
A4-Agent (Ours)	✓	✓	70.52	64.62	75.24	55.22

Open-Vocabulary Detection. We begin by using Rex-Omni [20], a state-of-the-art open-vocabulary object detector, to perform initial spatial localization from textual descriptions. Given textual description \mathbf{D} provided by Thinker, Rex-Omni outputs: **Bounding Boxes** $\{\mathbf{B}_i\}_{i=1}^N$: Rectangular regions that coarsely enclose the affordance parts. **Key Points** $\{\mathbf{P}_i\}_{i=1}^N$: Representative spatial anchors within each affordance region (*e.g.*, the center of a handle).

Fine-Grained Segmentation with SAM. We then pass the bounding boxes \mathbf{B}_i and key points \mathbf{P}_i predicted by Rex-Omni as prompts to SAM, which generates detailed segmentation masks $\{\mathbf{M}_i\}_{i=1}^N$ that delineate the precise boundaries of the affordance regions. This prompt-based approach requires no additional training, directly leveraging SAM’s powerful generalization capabilities developed through large-scale pretraining. The final affordance prediction aggregates multi-granular spatial information:

$$\mathcal{A}_{\text{ff}} = \{(\mathbf{B}_i, \mathbf{P}_i, \mathbf{M}_i)\}_{i=1}^N, \quad (5)$$

providing comprehensive spatial representations suitable for a variety of downstream applications—coarse bounding boxes for rapid scene understanding, key points for interaction targeting, and fine segmentation masks for precise manipulation planning.

In our Spotter, each model capitalizes on its strengths, and both can be independently upgraded as improved models emerge, without the need for end-to-end retraining.

5. Experiment

5.1. Experimental Settings

Implementation Details A4-Agent is a training-free framework coordinating pre-trained foundation models. In our complete agent, the VLM we used is GPT-4o [19], the generative model we used is Qwen-Image-Editing [48]. For the open-vocabulary object detection, we use Rex-Omni [20]; and SAM2-Large [38] for the segmentation.

Table 2. Quantitative results on RAGNet-3DOI and RAGNet-HANDAL. **A4-Agent achieves SOTA performance in zero-shot manner without any training.**

Model	Zero-shot	3DOI		HANDAL-easy		HANDAL-hard	
		gIoU \uparrow	cIoU \uparrow	gIoU \uparrow	cIoU \uparrow	gIoU \uparrow	cIoU \uparrow
G-DINO [27]	✓	4.1	3.9	3.6	3.0	3.4	3.1
LISA [22]	✓	12.3	8.1	15.5	11.9	12.3	8.1
GLaMM [37]	✓	4.4	2.9	4.7	3.5	5.0	3.5
Vision-Reasoner [29]	✓	39.6	30.3	29.6	19.8	27.7	16.7
Affordance-R1 [45]	✓	39.0	33.4	43.1	38.7	40.7	37.9
AffordanceVLM [49]	✗	38.1	39.4	58.3	58.1	58.2	57.8
A4-Agent (Ours)	✓	63.9	58.3	61.1	61.7	61.0	59.6

Datasets We evaluate A4-Agent on three quantitative benchmarks and a set of open-world images to assess both reasoning-aware affordance prediction and generalization to diverse scenarios. **Crucially, our framework is completely zero-shot—it has never been trained or fine-tuned on any of these datasets.**

1) ReasonAff [45]: A reasoning-oriented dataset built upon Instruct-Part [44] with complex instructions requiring deep semantic understanding. We use the test split containing 600 image-task pairs.

2) RAGNet [49]: A large-scale reasoning-based affordance segmentation dataset. We evaluate on two subsets: RAGNET-3DOI and RAGNET-HANDAL, containing 3,018 image-task pairs in total.

3) UMD Part Affordance [33]: A standard affordance dataset covering 17 object categories with 7 affordance types. Following prior work [45], we sample one-tenth of the frames, yielding 1,922 test images.

4) Open-World Images: To evaluate generalization beyond standard benchmarks (which focus mainly on kitchen and household scenes), we collect diverse images from Phys-ToolBench [56] and web sources for qualitative evaluation.

5.2. Quantitative Results

Results on ReasonAff Dataset. Table 1 and Fig. 4 presents results on ReasonAff, which demands deep reasoning over implicit contextual instructions. A4-Agent achieves state-of-the-art performance across all metrics without any training. Compared to supervised methods like AffordanceLLM (48.49 gIoU) and reasoning-enhanced approaches like Vision Reasoner (63.04 gIoU) and Affordance-R1 (67.41 gIoU), A4-Agent reaches 71.83 gIoU, demonstrating superior reasoning ability and generalization.

This performance stems from three design principles. First, decoupling reasoning from grounding leverages complementary strengths—VLMs excel at semantic interpretation while specialized vision models provide precise localization. Second, the “think-with-imagination” mechanism grounds abstract instructions in synthesized visual representations, enhancing affordance understanding in complex scenarios. Third, unlike end-to-end models constrained by

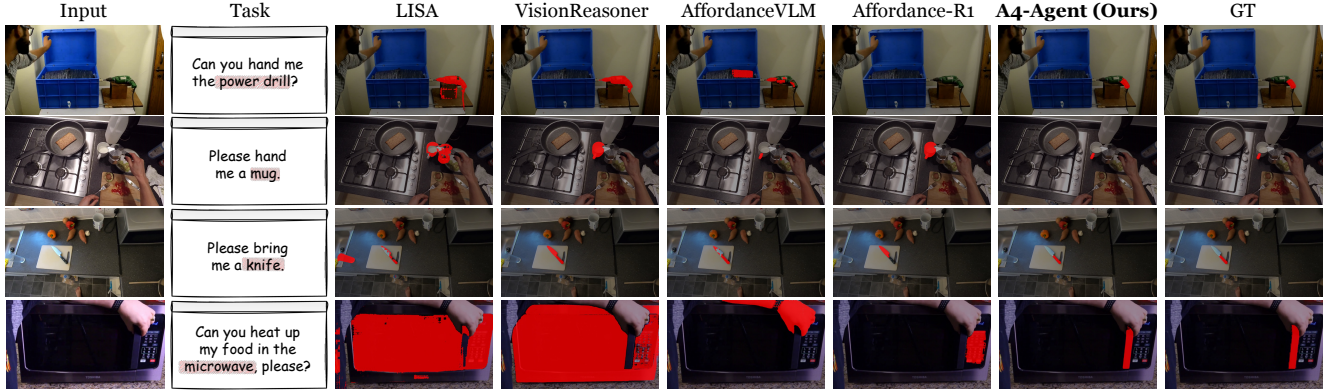


Figure 5. Qualitative comparison on RAGNet dataset. Our zero-shot method effectively reasons over task instructions to identify correct regions and precisely localize them with masks, closely matching ground truth. This outperforms baseline methods including AffordanceVLM trained on this dataset.

training data, our zero-shot approach naturally generalizes to ReasonAff’s diverse instructions.

Results on RAGNet Dataset. Table 2 shows results on RAGNet, which focuses on reasoning-based affordance segmentation. Our framework demonstrates exceptional zero-shot performance, significantly outperforming all baselines. On 3DOI, A4-Agent achieves 63.9 gIoU, surpassing Vision-Reasoner by over 24 points. This extends to HANDAL-hard and HANDAL-hard, where our A4-Agent also achieves highest score. Qualitative comparison is shown in Fig. 5.

Critically, A4-Agent even surpasses the supervised AffordanceVLM trained on this dataset, validating that *agentic coordination of foundation models outperforms task-specific fine-tuning* for complex reasoning tasks. This stems from our ability to decompose abstract instructions into actionable steps and accurately ground them visually—a core benefit of our decoupled architecture.

Results on UMD Dataset. Besides the task of complex reasoning-intensive affordance prediction, A4-Agent also excels at the more traditional tasks of predicting affordances from action concepts. As shown in Table 3, A4-Agent still achieves state-of-the-art performance, significantly outperforming the baselines by 15.53 gIoU, demonstrating a deep understanding of the many possible uses of different parts of common objects. This result is predictable, as this fundamental capability is arguably straightforward for powerful pre-trained models. This further supports our motivation: training-free coordination of specialized foundation models can exhibit strong generalization, as these models already possess sufficiently rich general knowledge.

5.3. Qualitative Results on Open-World Images

To further validate the performance of A4-Agent on open-world scenarios, we perform a qualitative experiments of open-world images. Fig. 6 shows A4-Agent’s strong per-

Table 3. Zero-shot results on UMD dataset. A4-Agent outperforms fine-tuned methods without any training.

Model	gIoU↑	cIoU↑	P_{50} ↑	P_{50-95} ↑
LISA-7B [22]	41.90	41.23	39.65	19.33
SAM4MLLM [7]	12.40	8.41	4.12	0.05
AffordanceLLM [36]	43.11	38.97	41.56	22.36
Qwen2.5VL-7B [3]	33.21	29.83	25.17	10.45
InternVL3-7B [58]	30.46	28.73	18.67	9.94
AffordanceVLM [49]	25.41	17.96	9.37	25.10
Seg-Zero [28]	44.26	39.30	39.93	16.53
Vision Reasoner [29]	44.00	39.71	39.04	16.10
Affordance-R1 [45]	49.85	42.24	53.35	34.08
A4-Agent (Ours)	65.38	59.81	77.31	43.78

formance across challenging scenarios: (1) **Novel objects:** Successfully identifying actionable regions on objects absent from standard benchmarks (e.g., digital equipment); (2) **Complex scenes:** Accurately identifying the most suitable part of a tool in complex environment (e.g., the tip of a screwdriver); (3) **Deep reasoning:** Using strong reasoning abilities to logically deduce the appropriate tools (e.g., a slotted spoon can be used to drain water, a rock can serve as a substitute for a hammer to drive nails).

Unlike baselines, which often fail on out-of-distribution objects, A4-Agent maintains consistent performance by leveraging broad knowledge from web-scale pre-trained models. This confirms that training-free coordination has great potential for real-world application.

5.4. Ablation Study

Importance of Imagination in Affordance Reasoning. Table 4 evaluates visual imagination’s contribution. The imagination mechanism provides consistent improvements across all metrics for all base models. Notably, open-source Qwen-2.5-VL (7B) with imagination even outperforms closed-source GPT-4o using text-only reasoning.

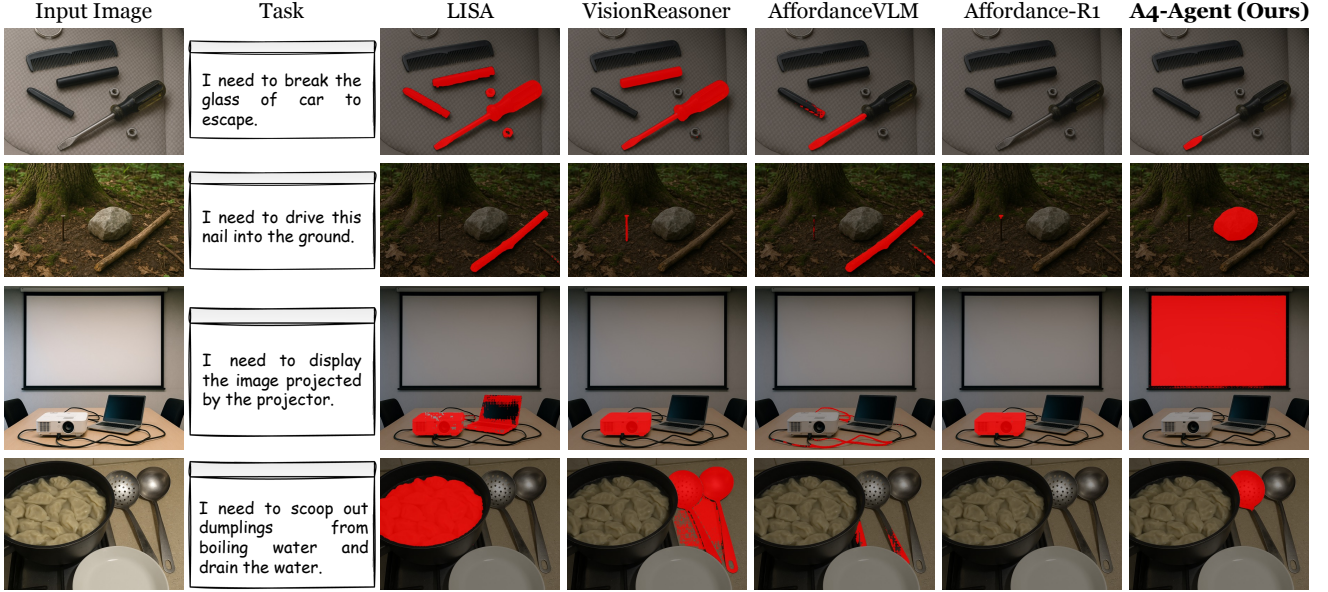


Figure 6. Qualitative results on open-world images. A4-Agent demonstrates robust affordance reasoning across diverse scenarios, consistently produces reasonable regions based on complex instructions.

Table 4. Ablation on Imagination on RAGNet-3DOI Dataset. *Affordance-R1 was fine-tuned from Qwen-2.5-VL-7B. T-w-I refers to think-with-imagination, which is the *dreamer*.

Method	Reasoning backbone	gIoU \uparrow	cIoU \uparrow
AffordanceVLM	LISA	38.10	39.40
Affordance-R1	Qwen-2.5-VL-7B*	39.04	33.39
Ours w/o T-w-I	Qwen-2.5-VL-7B	58.48	49.26
Ours w T-w-I	Qwen-2.5-VL-7B	63.02 (\uparrow 4.54)	49.87 (\uparrow 0.61)
Ours w/o T-w-I	GPT-4o	62.30	54.43
Ours w T-w-I	GPT-4o	63.94 (\uparrow 1.64)	58.30 (\uparrow 3.87)

This validates that grounding reasoning in synthesized visual representations enhances affordance understanding, especially when textual descriptions alone are insufficient. The visual imagination mechanism serves as a bridge, allowing the reasoning model to effectively tap into and leverage the vast prior knowledge about interaction encapsulated within the generative model.

Robustness to Different Components. We analyze A4-Agent’s robustness to different component choices.

Reasoning Backbone. Table 5 shows that replacing Qwen-2.5-VL with the more powerful GPT-4o significantly improves performance. This demonstrates A4-Agent’s flexibility to seamlessly incorporate stronger foundation models as they become available.

Segmentation Backbone. Replacing SAM2-Large with smaller variants (SAM2-Base-Plus/Tiny) causes slight performance drops, but the framework remains highly effective and significantly outperforms baselines. The performance

Table 5. Ablation on different components on RAGNet-3DOI Dataset. *AffordanceVLM is finetuned from LISA. SAM2-L,B,T denotes SAM2-Large, Base-plus, Tiny.

Method	Reasoning	Segmentation	gIoU \uparrow	cIoU \uparrow
AffordanceVLM	LISA*	LISA*	38.10	39.40
Affordance-R1	Qwen-2.5-VL-7B*	SAM2-L	39.04	33.39
	Qwen-2.5-VL-7B*	SAM2-T	36.13 (\downarrow 2.91)	30.76 (\downarrow 2.63)
Ours	GPT-4o	SAM2-L	62.30 (\uparrow 4.54)	54.43 (\uparrow 5.17)
	Qwen-2.5-VL-7B	SAM2-L	58.48	49.26
	Qwen-2.5-VL-7B	SAM2-B	56.84 (\downarrow 1.64)	48.87 (\downarrow 0.39)
	Qwen-2.5-VL-7B	SAM2-T	56.32 (\downarrow 2.16)	47.18 (\downarrow 2.08)

drop is also smaller than baseline method Affordance-R1. This underscores the robustness of our approach even with weaker grounding components.

6. Conclusion

In this paper, we present A4-Agent, a novel training-free framework for affordance prediction. Our key contribution is decoupling the task into high-level reasoning and low-level grounding, enabling the use of vision-language models for semantic interpretation and vision foundation models for localization. We also introduce an imagination mechanism in reasoning, where a generative model visualizes potential interactions to improve the process. Extensive experiments show that this zero-shot approach outperforms supervised methods on challenging benchmarks and generalizes well to open-world scenarios. The success of A4-Agent highlights the potential of agentic coordination of foundation models for complex affordance prediction.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2
- [2] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. *arXiv preprint arXiv:2207.09450*, 2022. 2
- [3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6, 7, 1
- [4] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE international conference on computer vision*, pages 1017–1025, 2015. 2
- [5] Harold Haodong Chen, Haojian Huang, Qifeng Chen, Harry Yang, and Ser-Nam Lim. Hierarchical fine-grained preference optimization for physically plausible video generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 3
- [6] Harold Haodong Chen, Disen Lan, Wen-Jie Shu, Qingyang Liu, Zihan Wang, Sirui Chen, Wenkai Cheng, Kanghao Chen, Hongfei Zhang, Zixin Zhang, et al. Tivibench: Benchmarking think-in-video reasoning for video generative models. *arXiv preprint arXiv:2511.13704*, 2025. 3
- [7] Yi-Chia Chen, Wei-Hua Li, Cheng Sun, Yu-Chiang Frank Wang, and Chu-Song Chen. Sam4mllm: Enhance multimodal large language model for referring expression segmentation. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. 6, 7, 1
- [8] Ethan Chern, Zhulin Hu, Steffi Chern, Siqi Kou, Jiadi Su, Yan Ma, Zhijie Deng, and Pengfei Liu. Thinking with Generated Images, 2025. 3
- [9] Hengshuo Chu, Xiang Deng, Xiaoyang Chen, Yinchuan Li, Jianye Hao, and Liqiang Nie. 3d-affordancellm: Harnessing large language models for open-vocabulary affordance detection in 3d worlds. *arXiv preprint arXiv:2502.20041*, 2025. 2
- [10] Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 3d affordancenet: A benchmark for visual object affordance understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1778–1787, 2021. 2
- [11] Xianqiang Gao, Pingrui Zhang, Delin Qu, Dong Wang, Zhigang Wang, Yan Ding, Bin Zhao, and Xuelong Li. Learning 2d invariant affordance knowledge for 3d affordance grounding. *arXiv preprint arXiv:2408.13024*, 2024. 2
- [12] James J Gibson. The theory of affordances. *Hilldale, USA*, 1(2):67–82, 1977. 2
- [13] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8359–8367, 2018. 2
- [14] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2
- [15] Litao Guo, Xinli Xu, Luozhou Wang, Jiantao Lin, Jinsong Zhou, Zixin Zhang, Bolan Su, and Ying-Cong Chen. Comfyind: Toward general-purpose generation via tree-based planning and reactive feedback. *arXiv preprint arXiv:2505.17908*, 2025. 3
- [16] Ziyu Guo, Xinyan Chen, Renrui Zhang, Ruichuan An, Yu Qi, Dongzhi Jiang, Xiangtai Li, Manyuan Zhang, Hongsheng Li, and Pheng-Ann Heng. Are video models ready as zero-shot reasoners? an empirical study with the mme-cof benchmark. *arXiv preprint arXiv:2510.26802*, 2025. 3
- [17] Cheng-Chun Hsu, Zhenyu Jiang, and Yuke Zhu. Ditto in the house: Building articulation models of indoor scenes through interactive perception. *arXiv preprint arXiv:2302.01295*, 2023. 2
- [18] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaoshen Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *ArXiv*, abs/2503.06749, 2025. 2
- [19] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 6
- [20] Qing Jiang, Junan Huo, Xingyu Chen, Yuda Xiong, Zhaoyang Zeng, Yihao Chen, Tianhe Ren, Junzhi Yu, and Lei Zhang. Detect anything via next point prediction, 2025. 6
- [21] Ashish Kumar, Saurabh Gupta, David Fouhey, Sergey Levine, and Jitendra Malik. Visual memory for robust path following. *Advances in neural information processing systems*, 31, 2018. 2
- [22] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 2, 6, 7, 1
- [23] Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025. 3
- [24] Gen Li, Varun Jampani, Deqing Sun, and Laura Sevilla-Lara. Locate: Localize and transfer object parts for weakly supervised affordance grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10922–10931, 2023. 2
- [25] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7061–7070, 2023. 6, 1
- [26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems 2023, NeurIPS 2023*, 2023. 2

- [27] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 6, 1
- [28] Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025. 2, 3, 6, 7, 1
- [29] Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified visual perception and reasoning via reinforcement learning. *arXiv preprint arXiv:2505.12081*, 2025. 3, 6, 7, 1
- [30] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2252–2261, 2022. 2
- [31] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning visual affordance grounding from demonstration videos. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2
- [32] Teli Ma, Jia Zheng, Zifan Wang, Ziyao Gao, Jiaming Zhou, and Junwei Liang. Glover++: Unleashing the potential of affordance learning from human behaviors for robotic manipulation. *arXiv preprint arXiv:2505.11865*, 2025. 2
- [33] Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 1374–1381. IEEE, 2015. 6, 2
- [34] Toan Nguyen, Minh Nhat Vu, An Vuong, Dzong Nguyen, Thieu Vo, Ngan Le, and Anh Nguyen. Open-vocabulary affordance detection in 3d point clouds. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5692–5698. IEEE, 2023. 2
- [35] OpenAI. OpenAI o1. <https://openai.com/o1/>, 2024. 2
- [36] Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. Affordancellm: Grounding affordance from vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7587–7597, 2024. 2, 6, 7, 1
- [37] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024. 6, 1
- [38] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 6, 1
- [39] Yawen Shao, Wei Zhai, Yuhang Yang, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. Great: Geometry-intention collaborative inference for open-vocabulary 3d object affordance grounding. *arXiv preprint arXiv:2411.19626*, 2024. 2
- [40] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 2
- [41] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. Vlm-r1: A stable and generalizable r1-style large vision-language model. *ArXiv*, abs/2504.07615, 2025. 2
- [42] Peize Sun, Shoufa Chen, Chenchen Zhu, Fanyi Xiao, Ping Luo, Saining Xie, and Zhicheng Yan. Going denser with open-vocabulary part segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15453–15465, 2023. 6, 1
- [43] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwei Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3
- [44] Zifu Wan, Yaqi Xie, Ce Zhang, Zhiqiu Lin, Zihan Wang, Simon Stepputtis, Deva Ramanan, and Katia Sycara. Instruct-part: Task-oriented part segmentation with instruction reasoning. In *The 63rd Annual Meeting of the Association for Computational Linguistics*, 2025. 6, 1
- [45] Hanqing Wang, Shaoyang Wang, Yiming Zhong, Zemin Yang, Jiamin Wang, Zhiqing Cui, Jiahao Yuan, Yifan Han, Mingyu Liu, and Yuexin Ma. Affordance-r1: Reinforcement learning for generalizable affordance reasoning in multimodal large language model. *arXiv preprint arXiv:2508.06206*, 2025. 2, 3, 4, 6, 7, 1
- [46] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, pages 24824–24837. Curran Associates, Inc., 2022. 2
- [47] Zeming wei, Junyi Lin, Yang Liu, Weixing Chen, Jingzhou Luo, Guanbin Li, and Liang Lin. 3daffordsplat: Efficient affordance reasoning with 3d gaussians, 2025. 2
- [48] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. 5, 6
- [49] Dongming Wu, Yanping Fu, Saikie Huang, Yingfei Liu, Fan Jia, Nian Liu, Feng Dai, Tiancai Wang, Rao Muhammad Anwer, Fahad Shahbaz Khan, et al. Ragnet: Large-scale reasoning-based affordance segmentation benchmark towards general grasping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11980–11990, 2025. 2, 4, 6, 7, 1

- [50] Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Mind’s eye of LLMs: Visualization-of-thought elicits spatial reasoning in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3
- [51] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2945–2954, 2023. 6, 1
- [52] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 2
- [53] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Grounding 3d object affordance from 2d interactions in images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10905–10915, 2023. 2
- [54] Chunlin Yu, Hanqing Wang, Ye Shi, Haoyang Luo, Sibe Yang, Jingyi Yu, and Jingya Wang. Seqafford: Sequential 3d affordance reasoning via multimodal large language model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1691–1701, 2025. 2
- [55] Hongfei Zhang, Kanghao Chen, Zixin Zhang, Harold Haodong Chen, Yuanhuiyi Lyu, Yuqi Zhang, Shuai Yang, Kun Zhou, and Yingcong Chen. Dualcam-ctrl: Dual-branch diffusion model for geometry-aware camera-controlled video generation. *arXiv preprint arXiv:2511.23127*, 2025. 3
- [56] Zixin Zhang, Kanghao Chen, Xingwang Lin, Lutao Jiang, Xu Zheng, Yuanhuiyi Lyu, Litao Guo, Yinchuan Li, and Ying-Cong Chen. Phystoolbench: Benchmarking physical tool understanding for mllms. *arXiv preprint arXiv:2510.09507*, 2025. 6
- [57] Ding Zhong, Xu Zheng, Chenfei Liao, Yuanhuiyi Lyu, Jiale Chen, Shengyang Wu, Linfeng Zhang, and Xuming Hu. Omnisam: Omnidirectional segment anything model for uda in panoramic semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23892–23901, 2025. 1
- [58] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, et al. Internv13: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. 6, 7, 1
- [59] Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, Chi Zhang, Yang You, Zhaoxiang Zhang, Dawei Zhao, Liang Xiao, Jian Zhao, Jiwen Lu, and Guan Huang. Is sora a world simulator? a comprehensive survey on general world models and beyond. *arXiv preprint arXiv:2405.03520*, 2024. 3

A4-Agent: An Agentic Framework for Zero-Shot Affordance Reasoning

Supplementary Material

7. More Implementation Detail

7.1. Details of Baseline Methods

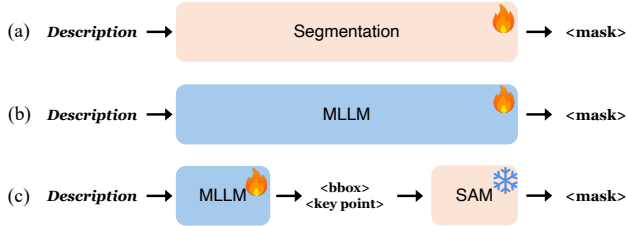


Figure 7. Illustration of different categories of baseline methods.

We comprehensively selected methods suitable for the affordance prediction task as our baselines, which fall into the following major categories:

a) Open-Vocabulary Segmentation. These methods take open-vocabulary textual prompts as input and, given a description of an object, output a corresponding segmentation mask. Representative models include VL-Part [42], OVSeg [25], SAN [51], and Grounding-DINO(G-DINO) [27].

b) MLLM-Enhanced End-to-End Segmentation. These approaches fine-tune MLLMs to directly produce mask tokens and decode them into masks. Given a textual description, the models can generate segmentation masks in an end-to-end manner. This category includes AffordanceLLM [36], AffordanceVLM [49], LISA [22], SAM4MLLM [7], and GLaMM [37].

c) MLLM for Grounding + SAM for Segmentation. These methods follow a two-stage paradigm: the MLLM first performs task-aware grounding by predicting bounding boxes and keypoints for the target objects, and then a segmentation model (e.g., SAM2 [38]) takes these as input to produce the final masks. Representative methods include Seg-Zero [28], Vision Reasoner [29], and Affordance-R1 [45], all of which are fine-tuned from open-source MLLMs. We also include open-source MLLMs with strong grounding ability such as Qwen-2.5-VL [3] and InternVL-3 [58] for comparison.

7.2. Evaluation Metrics

Following the standard evaluation protocol in affordance prediction [44, 45] and semantic segmentation [38, 57], we adopt four complementary metrics to comprehensively assess prediction quality including gIoU, cIoU, P@50, P@50:95.

gIoU (Generalized IoU): The average Intersection-over-Union across all images, measuring overall segmentation quality.

cIoU (Cumulative IoU): The cumulative intersection over cumulative union, providing a dataset-level quality measure.

P@50 (Precision at IoU=0.5): The percentage of predictions with IoU exceeding 0.5, evaluating high-quality predictions.

P@50:95: Average precision across IoU thresholds from 0.5 to 0.95 with 0.05 increments, providing a strict assessment of segmentation accuracy.

7.3. System prompt of our Agent

Prompt for Dreamer

```
You are an "Imagination-driven Image-Editing Prompt Writer".
Input: (a) an image, (b) a TASK description.
Task: Based on the input image and TASK, imagine a person or another object interacting with a target object within the scene to do the task. Then, produce ONE concise, photorealistic image-editing prompt to be used by a downstream model to edit the image, depicting this interaction.

Requirements:
- The prompt must clearly describe the interaction, including the action, the state of the target object, and any necessary manipulators (e.g., a person's hand, a tool).
- Refer to the existing object and scene; do not replace them.
- Preserve the identity (shape, texture, color) of existing objects and the background. The camera viewpoint should remain unchanged.
- If introducing a person, describe the pose and action of the relevant body parts (e.g., a hand gripping a handle) realistically.
- Enforce physical plausibility: the scale, perspective, lighting, and shadows of any new elements must seamlessly match the original image.
- Ensure all occlusions are logical.

Output format:
- Output ONLY the editing prompt text (no JSON, no lists, no quotes, no explanations).
- Begin with 'Edit the input image to...' and keep it to 1-3 sentences plus a short style clause (e.g., 'photorealistic, seamless inpainting').
- End with "keep others unchanged".

The given TASK is:
```

Prompt for Thinker

Given the image of an object, the task is to decide which object to use and predict the part of the object that matches the provided task. The task instruction is "TASK".
The first image is the original image of the object. The second image is the image of the object interacting with a person or another object in relation to the given affordance type for your reference.

****Follow these reasoning steps**:**

1. Identify the key components of the object in the first image (e.g., shape, features, possible points of interaction).
2. Analyze the second image to understand how the object is interacting with a person or another object in relation to the given affordance type.
3. Go back to the first image and ground the part of the object in the image and output the result in a structured JSON format.

****Output format**:**

```
### Thinking
thinking process
### Output
{
  "task": "the task instruction",
  "object_name": "the name of the object",
  "object_part": "the [object part] of the [
    object name] (e.g. the blade of the
    shears)"
}
```

8. More Exploratory Experiments

Here we present additional exploratory experiments. For example, although Rex-Omni is used here as our object detection model, its backbone is an MLLM, which gives it much stronger language understanding capabilities than other traditional detection models. Motivated by this, we constructed, in an exploratory manner, a framework that uses only Rex-Omni and SAM. In terms of the A4-Agent, this corresponds to using only the Spotter module, without the preceding Dreamer and Thinker. We evaluate this variant on RAGNet-3DOI, and the results are shown below in the last line in Tab. 6.

Table 6. More Exploratory results on RAGNet-3DOI.

Dreamer	Thinker	Spotter	gIoU↑	cloU↑
✓	✓	✓	63.94	58.30
✗	✓	✓	62.30	54.43
✗	✗	✓	45.91	39.82

Although Spotter itself has some reasoning ability, its backbone is only a small MLLM with limited reasoning capacity, which leads to suboptimal performance. This further validates our motivation: *by decoupling the reasoning and grounding processes, we can fully exploit their*

respective strengths and easily scale up the system to improve its overall performance.

9. More Intermediate Results

We here show more intermediate results of our A4-Agent in Figure 8 to 13, where Fig. 8 and Fig. 9 are sampled results on ReasonAff dataset [45], Fig. 10 and Fig. 11 are sampled results on UMD dataset [33], and Fig. 12 and Fig. 13 are sampled results on RAGNet Dataset [49].

Input Image & Task Description



To hold this bottle securely, which areas on its body should your hand grip?

Dreamer

Edit the input image to include a hand gripping the bottle firmly around its midsection, just below the label, ensuring the fingers wrap securely around the glass while the thumb supports the back. The hand should appear natural and realistic, with visible details like skin texture and light reflections, seamlessly integrated into the scene. Keep others unchanged.



Thinker

Thinking

First identify which object to use from the scene. The scene contains a single object, which is clearly a bottle. The task instruction focuses on holding a bottle securely, so the relevant object is the bottle in the image.

Identify the key components of this object in the image (e.g., shape, features, possible points of interaction). The bottle has a cylindrical body, a neck, and a cap. The main components for interaction are the bottle's body, which provides a firm area for gripping.

Analyze the object and the task instruction to determine the part of the object that matches the task instruction. To hold a bottle securely, the body of the bottle is the area that should be gripped. This part is wide enough to provide a firm hold and is where most people naturally grip a bottle.

Ground the part of the object in the image and output the result in a structured JSON format. The bottle's body is slightly below the neck and extends to the lower part of the bottle. This is the ideal area for gripping.

Output

```
{ "object_part": "the body of the bottle",  
  "object_name": "bottle",  
  "task": "To hold this bottle securely, which areas on its body should your hand grip?"  
}
```



Spotter

Rex-Omni



the body of the bottle



SAM-2



Figure 8. Full Demonstration of Intermediate Results. Sampled from results on the ReasonAff dataset.

Input Image & Task Description



What part of the faucet would you need to press down to turn it on?

Dreamer

Edit the input image to show a hand gently pressing down on the lever of the faucet, causing water to flow from the spout. The hand should be positioned naturally, with fingers wrapped around the lever, and the water stream should be visible. Keep others unchanged. Photorealistic, seamless inpainting.



Thinker

Thinking

Identify the key components of the object in the first image: The object in the first image is a faucet with two main components: the spout and two handles. The handles are typically used to control the flow of water and can be pressed down or lifted to turn the faucet on or off.

Analyze the second image: The second image shows a person interacting with the faucet, specifically pressing down on one of the handles to turn it on. This indicates that the handles are the interactive components in relation to the given task of turning the faucet on.

Ground the part of the object in the first image: Based on the analysis of both images, the part of the faucet that needs to be pressed down to turn it on is one of the handles. In the first image, the handle that should be pressed is the one on the right.

Output

```
{
  "task": "What part of the faucet would you need to press down to turn it on?",
  "object_name": "faucet",
  "object_part": "the handle of the faucet"
}
```



Spotter

Rex-Omni



the handle of the faucet

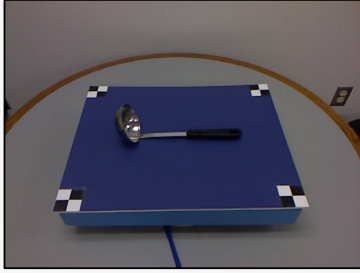


SAM-2



Figure 9. Full Demonstration of Intermediate Results. Sampled from results on the ReasonAff dataset.

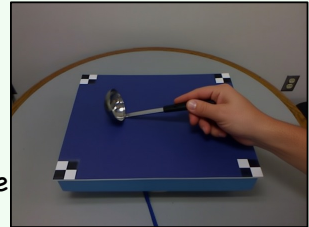
Input Image & Task Description



Grasp

Dreamer

Edit the input image to show the existing ladle being grasped by a hand; position the hand around the handle, ensuring the thumb and fingers wrap naturally; maintain the ladle's orientation and placement on the blue surface; ensure lighting, shadows, and perspective match the original; preserve the ladle's identity and background; photorealistic, seamless inpainting.



Thinker

Thinking

In the first image, the ladle is resting on a flat surface, and there are no visible signs of interaction with a person or another object.

In the second image, a hand is shown grasping the handle of the ladle, indicating the part of the ladle that can be interacted with using the "grasp" affordance.

The handle of the ladle is the part that would typically be used to hold and manipulate the ladle, making it the most relevant part for the "grasp" affordance.

Output

```
{
  "object_part": "the handle of the ladle",
  "affordance_type": "grasp",
}
```



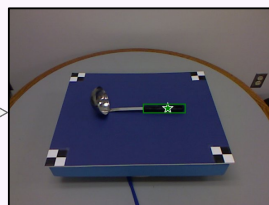
GPT-4o

Spotter

Rex-Omni



the handle of the ladle

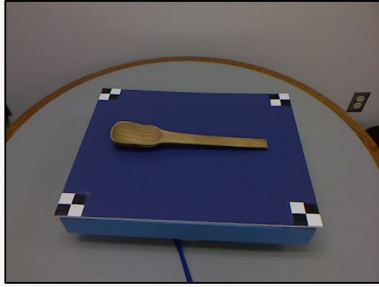


SAM-2



Figure 10. **Full Demonstration of Intermediate Results.** Sampled from results on the UMD dataset.

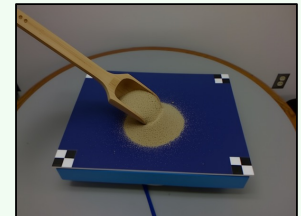
Input Image & Task Description



Scoop

Dreamer

Edit the input image to show the existing wooden spoon scooping up a small amount of sand or similar granular material from a surface beneath it; ensure the spoon is positioned at an angle with its bowl facing upward, capturing the material visibly lifting off the surface; preserve the object's identity and background; photorealistic, seamless inpainting.



Thinker

Thinking

In the first image, the spoon is resting on a flat surface without any interaction.

In the second image, the spoon is being used to scoop a granular substance, which is consistent with the affordance type "scoop".

The part of the spoon that interacts with the granular substance is the bowl of the spoon, which is designed to hold and transfer the substance.

Output

```
{
  "object_part": "bowl of the spoon",
  "affordance_type": "scoop"
}
```

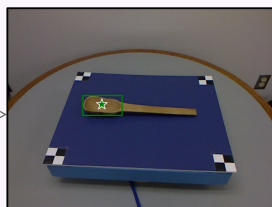


GPT-4o

Spotter

bowl of the spoon

Rex-Omni



SAM-2

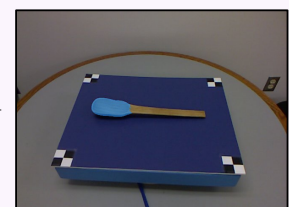


Figure 11. **Full Demonstration of Intermediate Results.** Sampled from results on the UMD dataset.

Input Image & Task Description



Can you please open the door for me?

Dreamer

Edit the input image to show a hand gently pushing the door open from the inside, revealing more of the room behind it, while the child looks curiously towards the viewer. Ensure the hand is positioned naturally near the door handle, and maintain the lighting and shadows consistent with the existing scene. Photorealistic, seamless inpainting. Keep others unchanged.



Thinker

###Thinking

Identify the key components of the object in the first image: The object is a door with a round doorknob. The key component of interaction for this task is the doorknob, which is used to open the door.

Analyze the second image: The second image shows a hand interacting with the doorknob, indicating that the doorknob is the component used to perform the task of opening the door.

Ground the part of the object in the image: Based on the task of opening the door, the relevant part of the object is the doorknob.

###Output

```
{
  "task": "Can you please open the door for me?",
  "object_name": "door",
  "object_part": "the doorknob of the door",
}
```

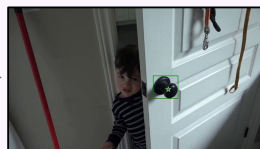


Spotter

Rex-Omni



the doorknob of the door



SAM-2

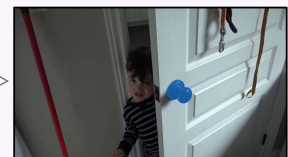


Figure 12. **Full Demonstration of Intermediate Results.** Sampled from results on the RAGNet dataset.

Input Image & Task Description



Can you please preheat
the oven for me?

Dreamer

Edit the input image to show a hand turning the oven dial to the right, indicating the preheating process has started. The hand should be positioned near the oven control panel, with fingers gripping the dial firmly. Ensure the hand is realistic in size and detail, casting a subtle shadow on the oven surface. Keep others unchanged. Photorealistic, seamless inpainting.



Thinker

Thinking

Identify the key components of the object in the first image: The first image shows a kitchen setting with a dishwasher and an oven. The oven is located on the right side of the image, below the stove. The oven has a control panel with knobs and a digital display. The knobs are likely used to control the temperature and settings of the oven.

Analyze the second image: In the second image, a person is interacting with the oven. The person's hand is holding one of the knobs on the control panel of the oven. This indicates that the knob is the part of the oven used to preheat it, as per the task instruction.

Ground the part of the object in the first image:

Based on the interaction observed in the second image, the relevant part for preheating the oven is the knob on the control panel.

Output {

"task": "Can you please preheat the oven for me?",

"object_name": "oven",

"object_part": "the knob of the oven"

}



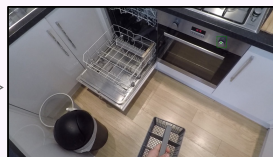
GPT-4o

Spotter

Rex-Omni



the knob of the oven



SAM-2



Figure 13. **Full Demonstration of Intermediate Results.** Sampled from results on the RAGNet dataset.