

outperforms state-of-the-art supervised methods across multiple benchmarks and demonstrates robust generalization to real-world settings.

Keywords: Affordance Prediction, Agentic Framework, Zero-Shot Generalization

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 fundamental to downstream applications including task planning [20], robotic grasping [2, 16], and human-robot collaboration [4, 12], 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 [23, 29, 30] 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 [35, 44, 48] 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 [11], 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 [10, 38, 52], human demonstration videos [31], and 3D perception through point clouds [8, 9, 33, 35, 53] or 3D Gaussian Splatting [46].

Recent advances have leveraged multimodal large language models (MLLMs) to enhance affordance understanding. For example, AffordanceLLM [35] and Seqafford [53] introduce special tokens into the vocabulary and map affordance regions to token embeddings for segmentation outputs. LISA [21] extends this paradigm by incorporating reasoning capabilities for language-driven segmentation tasks. More recently, Affordance-R1 [44] 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, 25, 51] have demonstrated remarkable capabilities in visual understanding, generation, and multimodal reasoning. Recent advances have significantly enhanced their reasoning abilities

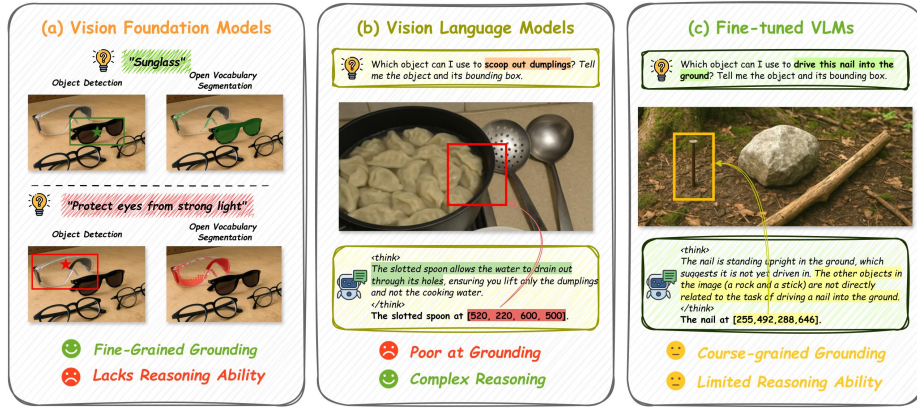


Fig. 2: Vision Foundation Models are good at fine-grained grounding, but are poor at reasoning. Vision Language Models are good at reasoning, but are poor at visual grounding. Some works finetuned VLMs for better grounding ability, but both abilities are underwhelming.

through inference-time scaling. OpenAI o1 [34] extends the Chain-of-Thought (CoT) [45] reasoning process to achieve superior performance, while DeepSeek-R1 [13] leverages reinforcement learning with GRPO [39] to further advance reasoning capabilities. Building on these successes, several works [17, 27, 40] have expanded these reasoning 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 [49] introduces textual imagery representations for dynamic reasoning. Benefiting from powerful generative models [14, 42, 54, 57], approaches [5, 7, 15, 22] 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 [27, 28, 44] 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 introduce fundamental limitations:

- ❶ **Limited generalization:** training on limited affordance datasets cannot cover the diversity of real-world scenarios, leading to brittleness on novel objects and environments. This is not merely a conceptual concern but is empirically measurable: as reported later in Tab. 1 and Tab. 2, Affordance-R1 excels on its in-domain ReasonAff (62.72 cIoU) yet drops to 33.4 cIoU on out-of-domain RAG-Net, while AffordanceVLM shows the exact opposite trend (39.4 cIoU on RAG-Net vs. 25.54 cIoU on ReasonAff). We attribute this to *specialization-induced forgetting*: fine-tuning a single model on limited affordance/grounding data improves dataset-specific accuracy but erodes the broad open-domain knowledge inherited from pre-training, whose data scale these specialized models can never match.
- ❷ **Capability trade-offs:** optimizing for both reasoning and grounding simultaneously forces the model to balance conflicting objectives, where improvements in one capability may degrade the other. Concretely, *no existing single model attains strong reasoning and grounding at the same time*: top-tier VLMs reason well but ground poorly (GPT-4o directly outputting boxes yields only 6.55 gIoU in Tab. 5), grounding-finetuned VLMs improve localization but remain suboptimal (35.07 gIoU), and specialized grounding models lack the reasoning to interpret complex instructions (45.91 gIoU).
- 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 expensive 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.

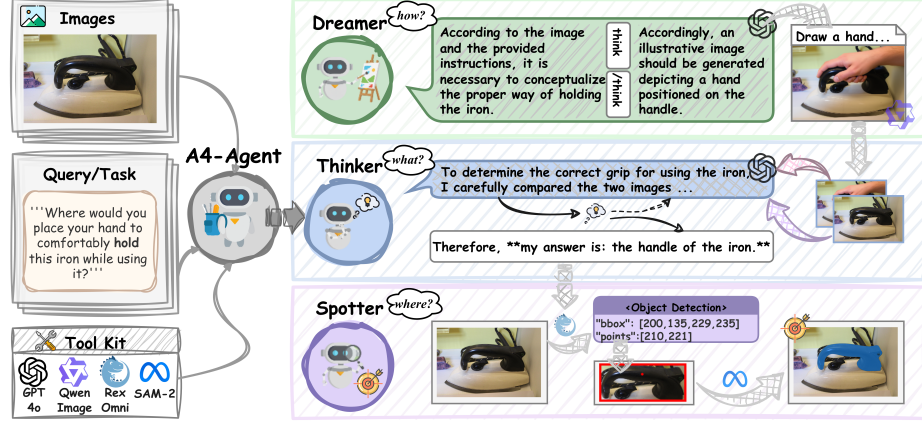


Fig. 3: The pipeline of our A4-Agent, which decouples affordance prediction into three stages. (1) **Dreamer**: Imagines the interaction by generating a simulated image. (2) **Thinker**: Reasons over the original and simulated images to produce a textual description of the actionable object part. (3)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

4 A4-Agent:Agentic Affordance Reasoning

4.1 Problem Definition

Following existing work [35, 44, 48], 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 $\mathcal{A}_{\text{ff}}$ that enables the specified interaction:

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

where $\mathcal{A}_{\text{ff}}$ 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 [44, 48], 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):

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

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 operate on (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 the interaction state required to accomplish the task and envisioning the broader usage scenario. We emphasize that this imagined interaction is ***not restricted to hand contact***: depending on the instruction and scene, it may involve a hand-object interaction (*e.g.*, a hand grasping a handle), an object-object interaction (*e.g.*, a kettle pouring water from its spout), or other physical interactions. Importantly, the functionally relevant region is therefore not always the grasped part—for “pour”, the target may be the spout/outlet rather than the handle. 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 based on the observation $\mathbf{I}$ and task $\mathbf{T}$.

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 [47] to synthesize interaction scenarios. Given the original image $\mathbf{I}$ and a simulation prompt $\mathbf{T}_{\text{sim}}$ derived from the task instruction (*e.g.*, “a hand grasping the refrigerator handle”), the model produces an edited image $\mathbf{I}_{\text{sim}}$:

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

where $\mathcal{G}$ denotes the image generation model. The imagined image $\mathbf{I}_{\text{sim}}$ explicitly highlights where interaction should occur by depicting plausible contact and motion cues and can further guide 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 $\mathbf{I}_{\text{sim}}$, 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 $\mathbf{I}_{\text{sim}}$ 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.

Open-Vocabulary Detection. We begin by using Rex-Omni [19], 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:

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

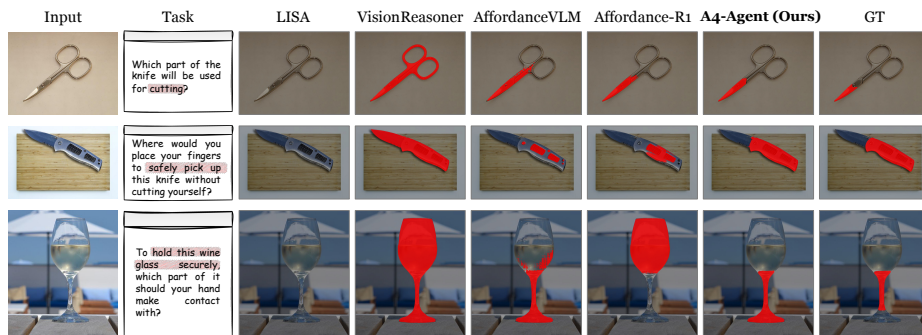


Fig. 4: Qualitative comparison on ReasonAff dataset. Our method continuously predicts appropriate components according to task requirements, achieving results most consistent with ground truth and even surpassing Affordance-R1 specifically trained on this dataset.

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

Model	LLM	Reasoning	gIoU $\uparrow$	cIoU $\uparrow$	$P_{50-95}\uparrow$	$P_{50}\uparrow$
VLPart [41]	✗	✗	4.21	3.88	1.31	0.85
OVSeg [24]	✗	✗	16.52	10.59	9.89	4.12
SAN [50]	✗	✗	10.21	13.45	7.18	3.17
LISA-7B [21]	✓	✗	38.17	40.58	33.62	19.69
SAM4MLLM [6]	✓	✗	45.51	33.64	43.48	22.79
AffordanceLLM [35]	✓	✗	48.49	38.61	42.11	20.19
InternVL3-8B [56]	✓	✓	31.79	24.68	35.41	21.93
Qwen2.5VL-7B [3]	✓	✓	25.18	20.54	26.00	15.82
AffordanceVLM [48]	✓	✓	30.50	25.54	30.29	18.31
Seg-Zero [27]	✓	✓	59.26	48.03	61.33	45.87
Vision Reasoner [28]	✓	✓	63.04	52.70	67.33	47.23
Affordance-R1 [44]	✓	✓	67.41	62.72	74.50	55.22
A4-Agent (Ours)	✓	✓	70.52	64.62	75.24	55.22

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 default setting, the VLM we used

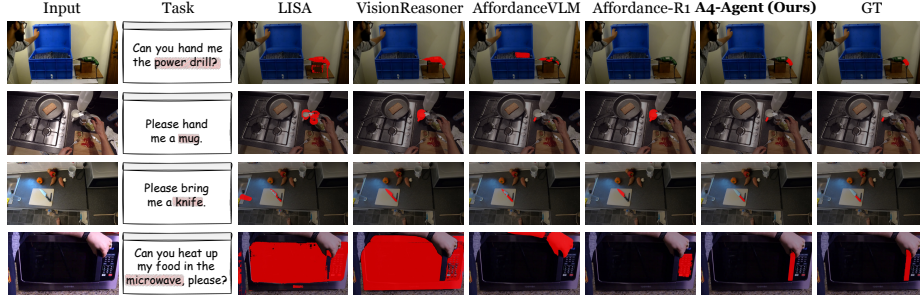


Fig. 5: Qualitative comparison on RAGNet dataset. Our zero-shot method effectively reasons over task instructions to identify correct regions and precisely localize them with masks, closely matching ground truth. This outperforms baseline methods including AffordanceVLM trained on this dataset.

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↑	cIoU↑	gIoU↑	cIoU↑	gIoU↑	cIoU↑
G-DINO [26]	✓	4.1	3.9	3.6	3.0	3.4	3.1
LISA [21]	✓	12.3	8.1	15.5	11.9	12.3	8.1
GLaMM [36]	✓	4.4	2.9	4.7	3.5	5.0	3.5
Vision-Reasoner [28]	✓	39.6	30.3	29.6	19.8	27.7	16.7
Affordance-R1 [44]	✓	39.0	33.4	43.1	38.7	40.7	37.9
AffordanceVLM [48]	✗	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

is GPT-4o [18], the backbone of the Dreamer is Qwen-Image-Editing [47]. For the open-vocabulary object detection and segmentation in Spotter, we use Rex-Omni [19] and SAM2-Large [37] respectively. To account for the stochasticity of the generative and language models, all reported results are averaged over three independent runs, and the standard deviation of every metric is consistently below 1.0, confirming the stability of our framework.

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 [44]: A reasoning-oriented dataset built upon Instruct-Part [43] with complex instructions requiring deep semantic understanding. We use the test split containing 600 image-task pairs.

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

Model	gIoU $\uparrow$	cIoU $\uparrow$	$P_{50}\uparrow$	$P_{50-95}\uparrow$
LISA-7B [21]	41.90	41.23	39.65	19.33
SAM4MLLM [6]	12.40	8.41	4.12	0.05
AffordanceLLM [35]	43.11	38.97	41.56	22.36
Qwen2.5VL-7B [3]	33.21	29.83	25.17	10.45
InternVL3-7B [56]	30.46	28.73	18.67	9.94
AffordanceVLM [48]	25.41	17.96	9.37	25.10
Seg-Zero [27]	44.26	39.30	39.93	16.53
Vision Reasoner [28]	44.00	39.71	39.04	16.10
Affordance-R1 [44]	49.85	42.24	53.35	34.08
A4-Agent (Ours)	65.38	59.81	77.31	43.78

2) RAGNet [48]: 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 [32]: A standard affordance dataset covering 17 object categories with 7 affordance types. Following prior work [44], 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 PhysToolBench [55] and web sources for qualitative evaluation.

5.2 Quantitative Results

Results on ReasonAff Dataset. Table 1 and Fig. 4 present 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 70.52 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 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 demon-

strates 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-easy 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.

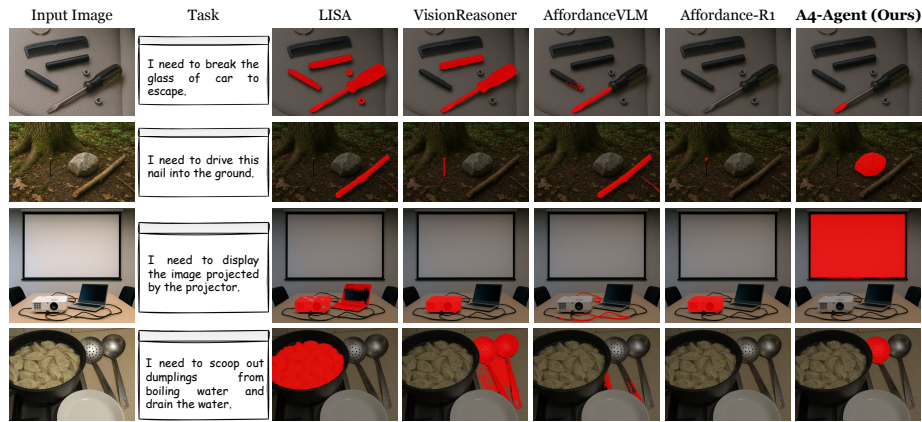


Fig. 6: Qualitative results on open-world images. A4-Agent demonstrates robust affordance reasoning across diverse scenarios, consistently produces reasonable regions based on complex instructions.

5.3 Qualitative Results on Open-World Images

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

Table 4: Ablations on component choices on RAGNet-3DOI Dataset.

#	Dreamer	Thinker	Spotter	gIoU↑	cIoU↑
0	<i>Best Baseline: Affordance-VLM</i>			38.10	39.40
1	Qwen-Image-Editing	GPT-4o	Rex-Omni + SAM2	63.94	58.30
2	<i>Nano-Banana-flash</i>	GPT-4o	Rex-Omni + SAM2	64.21	57.54
3	<i>FLUX.1 Kontext</i>	GPT-4o	Rex-Omni + SAM2	63.01	55.41
4	Qwen-Image-Editing	Qwen-2.5-VL-7B	Rex-Omni + SAM2	63.02	49.87
5	Qwen-Image-Editing	GPT-5.2	Rex-Omni + SAM2	64.65	59.21
6	Qwen-Image-Editing	GPT-4o	<i>DINO-X</i>	60.84	54.94
7	Qwen-Image-Editing	GPT-4o	<i>SAM3</i>	60.68	55.20

strong performance 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

Robustness to Different Component Choices. We analyze A4-Agent’s robustness to different component choices. As shown in Table 4, we establish the configuration in Line 1 (Qwen-Image-Editing as the Dreamer, GPT-4o as the Thinker, and Rex-Omni + SAM2 as the Spotter) as our baseline. We then conduct ablation studies by substituting the backbones of the Dreamer, Thinker, and Spotter: ❶ **Dreamer Backbone.** In Lines 1-3 of Tab. 4, switching the Dreamer from Qwen-Image-Editing to Nano-Banana-flash or FLUX.1 Kontext results in minimal performance fluctuation ($\sim 5\%$), still significantly outperforming the baselines. This indicates that *A4-Agent can consistently exploit the interaction priors of various visual generation models.* ❷ **Thinker Backbone.** Comparing Lines 1, 4, and 5, replacing the Thinker with a weaker open-source model (Qwen-2.5-VL-7B) results in a slight performance degradation, yet A4-Agent maintains a substantial lead over the baselines. Conversely, upgrading to the more powerful GPT-5.2 yields further performance gains. This demonstrates *A4-Agent’s flexibility to seamlessly incorporate stronger foundation models as they become available.* ❸ **Spotter Backbone.** Finally, comparing Lines 1, 6, and 7 reveals that switching the Spotter from Rex-Omni + SAM2 to DINO-X or SAM3 results in minimal performance variance ($\sim 5\%$). These results show that *A4-Agent exhibits strong robustness to the inherent biases of different detection and segmentation models.*

Table 5: Ablation on each component on the RAGNet-3DOI dataset.

#	Dreamer	Thinker	Spotter	gIoU↑	cIoU↑
1	✓	✓(GPT-4o)	✓	63.94	58.30
2	✗	✓(GPT-4o)	✓	62.30 (↓ 1.64)	54.43 (↓ 3.87)
3	✓	✓(Qwen-2.5-VL-7B)	✓	63.02	49.87
4	✗	✓(Qwen-2.5-VL-7B)	✓	58.48 (↓ 4.54)	49.26 (↓ 0.61)
5	✗	✗	✓	45.91 (↓ 18.0)	39.82 (↓ 18.5)
6	✓	✓(GPT-4o)	✗	6.55	3.89
7	✓	✓(Qwen-2.5-VL-7B)	✗	35.07	29.82

Effectiveness of Each Component. **① Dreamer:** Comparing Lines 1 and 3 with Lines 2 and 4 in Tab. 5, the Dreamer yields consistent improvements across all base models. While the performance gain for highly capable closed-source models like GPT-4o is relatively modest (+1.64 gIoU) —likely due to their already rich internal physical priors—the Dreamer provides a much more substantial performance leap (+4.54 gIoU) for the smaller open-source model (Qwen-2.5-VL-7B). Notably, the 7B model equipped with the Dreamer even outperforms the GPT-4o without Dreamer by 0.73 gIoU. This reveals that the Dreamer acts as an essential "visual physical simulator" that externalizes interaction priors, effectively compensating for the weaker internal knowledge of smaller models. Furthermore, additional representative cases provided in the supplementary material demonstrate that the Thinker effectively extracts valuable knowledge from the generative model’s priors during the reasoning process. **② Thinker:** Comparing Lines 1 and 5 in Tab. 5, we observe a drastic performance drop when using solely the Spotter without the Dreamer and Thinker, resulting in decreases of 18.0 in gIoU and 18.5 in cIoU. This decline occurs because, despite its strong open-vocabulary grounding capabilities, the Spotter lacks the requisite reasoning skills to independently perform affordance grounding. **③ Spotter:** Comparing Line 1 with Lines 6 and 7 in Tab. 5, we observe a significant performance degradation when the Spotter is removed and the Thinker is tasked with directly predicting bounding boxes for subsequent segmentation by SAM. This indicates that while the Thinker possesses strong reasoning abilities, its grounding capabilities are relatively weak. It can identify the general semantics of an object but struggles with precise localization, which would consequently hinder downstream robotic applications.

Overall, these results further demonstrate that the decoupled design of our A4-Agent allows each component to specialize in its respective function. This architecture not only circumvents the typical trade-off between reasoning and grounding capabilities but also leverages the generative model to assist in reasoning, ultimately achieving optimal overall performance.

5.5 Efficiency Analysis

The comparison of inference time, model size, GPU memory consumption between A4-Agent and baselines is provided in Tab. 6. It is worth noting that while

Table 6: Inference time using A40, model size, GPU memory, and metrics. All metrics are reported on the RAGNet-3DOI dataset.

Method	Inference Time	Model Size	GPU Memory	gIoU	cIoU
Affordance-R1	~7s	7B + 0.2B	~19 GB	39.04	33.39
Affordance-VLM	~4s	7B	~17 GB	38.10	39.40
A4-Agent (Lite)	~4s	3B + 3B + 0.2B	~16 GB	52.91	41.82
A4-Agent (Full)	~32s	20B + GPT-4o + 3B + 0.2B	~30 GB	63.94	58.30

the full pipeline maximizes performance, we also offer a Lite version (Qwen-2.5-VL-3B as Thinker, w/o Dreamer) for efficiency. The Lite version operates within less computational budget than baselines while delivering superior accuracy. This demonstrates that ***A4-Agent is highly adaptable, capable of supporting both high-performance configurations and lightweight, efficient robotic deployment.***

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. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 3
2. Bahl, S., Gupta, A., Pathak, D.: Human-to-robot imitation in the wild. arXiv preprint arXiv:2207.09450 (2022) 2
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 9, 11
4. Chao, Y.W., Wang, Z., He, Y., Wang, J., Deng, J.: Hico: A benchmark for recognizing human-object interactions in images. In: Proceedings of the IEEE international conference on computer vision. pp. 1017–1025 (2015) 2
5. Chen, H.H., Lan, D., Shu, W.J., Liu, Q., Wang, Z., Chen, S., Cheng, W., Chen, K., Zhang, H., Zhang, Z., et al.: Tivibench: Benchmarking think-in-video reasoning for video generative models. arXiv preprint arXiv:2511.13704 (2025) 4
6. Chen, Y.C., Li, W.H., Sun, C., Wang, Y.C.F., Chen, C.S.: Sam4mllm: Enhance multi-modal large language model for referring expression segmentation. In: European Conference on Computer Vision. pp. 323–340. Springer (2024) 9, 11

7. Chern, E., Hu, Z., Chern, S., Kou, S., Su, J., Ma, Y., Deng, Z., Liu, P.: Thinking with Generated Images (2025). <https://doi.org/10.48550/arXiv.2505.22525> 4
8. Chu, H., Deng, X., Chen, X., Li, Y., Hao, J., Nie, L.: 3d-affordancellm: Harnessing large language models for open-vocabulary affordance detection in 3d worlds. arXiv preprint arXiv:2502.20041 (2025) 3
9. Deng, S., Xu, X., Wu, C., Chen, K., Jia, K.: 3d affordancenet: A benchmark for visual object affordance understanding. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1778–1787 (2021) 3
10. Gao, X., Zhang, P., Qu, D., Wang, D., Wang, Z., Ding, Y., Zhao, B., Li, X.: Learning 2d invariant affordance knowledge for 3d affordance grounding. arXiv preprint arXiv:2408.13024 (2024) 3
11. Gibson, J.J.: The theory of affordances. *Hilldale, USA* **1**(2), 67–82 (1977) 3
12. Gkioxari, G., Girshick, R., Dollár, P., He, K.: Detecting and recognizing human-object interactions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8359–8367 (2018) 2
13. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025) 4
14. Guo, L., Xu, X., Wang, L., Lin, J., Zhou, J., Zhang, Z., Su, B., Chen, Y.C.: ComfyMind: Toward general-purpose generation via tree-based planning and reactive feedback. arXiv preprint arXiv:2505.17908 (2025) 4
15. Guo, Z., Chen, X., Zhang, R., An, R., Qi, Y., Jiang, D., Li, X., Zhang, M., Li, H., Heng, P.A.: Are video models ready as zero-shot reasoners? an empirical study with the mme-cof benchmark. arXiv preprint arXiv:2510.26802 (2025) 4
16. Hsu, C.C., Jiang, Z., Zhu, Y.: Ditto in the house: Building articulation models of indoor scenes through interactive perception. arXiv preprint arXiv:2302.01295 (2023) 2
17. Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Xu, Z., Hu, Y., Lin, S.: Vision-r1: Incentivizing reasoning capability in multimodal large language models. *ArXiv abs/2503.06749* (2025), <https://api.semanticscholar.org/CorpusID:276902576> 4
18. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024) 10
19. Jiang, Q., Huo, J., Chen, X., Xiong, Y., Zeng, Z., Chen, Y., Ren, T., Yu, J., Zhang, L.: Detect anything via next point prediction (2025), <https://arxiv.org/abs/2510.12798> 8, 10
20. Kumar, A., Gupta, S., Fouhey, D., Levine, S., Malik, J.: Visual memory for robust path following. *Advances in neural information processing systems* **31** (2018) 2
21. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9579–9589 (2024) 3, 9, 10, 11
22. Li, C., Wu, W., Zhang, H., Xia, Y., Mao, S., Dong, L., Vulić, I., Wei, F.: Imagine while reasoning in space: Multimodal visualization-of-thought. arXiv preprint arXiv:2501.07542 (2025) 4
23. Li, G., Jampani, V., Sun, D., Sevilla-Lara, L.: Locate: Localize and transfer object parts for weakly supervised affordance grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10922–10931 (2023) 2

24. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7061–7070 (2023) 9
25. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems 2023, NeurIPS 2023 (2023) 3
26. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023) 10
27. Liu, Y., Peng, B., Zhong, Z., Yue, Z., Lu, F., Yu, B., Jia, J.: Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. arXiv preprint arXiv:2503.06520 (2025) 4, 5, 9, 11
28. Liu, Y., Qu, T., Zhong, Z., Peng, B., Liu, S., Yu, B., Jia, J.: Visionreasoner: Unified visual perception and reasoning via reinforcement learning. arXiv preprint arXiv:2505.12081 (2025) 5, 9, 10, 11
29. Luo, H., Zhai, W., Zhang, J., Cao, Y., Tao, D.: Learning affordance grounding from exocentric images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2252–2261 (2022) 2
30. Luo, H., Zhai, W., Zhang, J., Cao, Y., Tao, D.: Learning visual affordance grounding from demonstration videos. IEEE Transactions on Neural Networks and Learning Systems (2023) 2
31. Ma, T., Zheng, J., Wang, Z., Gao, Z., Zhou, J., Liang, J.: Glover++: Unleashing the potential of affordance learning from human behaviors for robotic manipulation. arXiv preprint arXiv:2505.11865 (2025) 3
32. Myers, A., Teo, C.L., Fermüller, C., Aloimonos, Y.: Affordance detection of tool parts from geometric features. In: 2015 IEEE international conference on robotics and automation (ICRA). pp. 1374–1381. IEEE (2015) 11
33. Nguyen, T., Vu, M.N., Vuong, A., Nguyen, D., Vo, T., Le, N., Nguyen, A.: Open-vocabulary affordance detection in 3d point clouds. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5692–5698. IEEE (2023) 3
34. OpenAI: OpenAI o1. <https://openai.com/o1/> (2024) 4
35. Qian, S., Chen, W., Bai, M., Zhou, X., Tu, Z., Li, L.E.: Affordancellm: Grounding affordance from vision language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7587–7597 (2024) 2, 3, 6, 9, 11
36. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M.H., Khan, F.S.: Glamm: Pixel grounding large multimodal model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13009–13018 (2024) 10
37. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024) 10
38. Shao, Y., Zhai, W., Yang, Y., Luo, H., Cao, Y., Zha, Z.J.: Great: Geometry-intention collaborative inference for open-vocabulary 3d object affordance grounding. arXiv preprint arXiv:2411.19626 (2024) 3
39. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024) 4

40. Shen, H., Liu, P., Li, J., Fang, C., Ma, Y., Liao, J., Shen, Q., Zhang, Z., Zhao, K., Zhang, Q., Xu, R., Zhao, T.: Vlm-r1: A stable and generalizable r1-style large vision-language model. ArXiv **abs/2504.07615** (2025), <https://api.semanticscholar.org/CorpusID:277667819> 4
41. Sun, P., Chen, S., Zhu, C., Xiao, F., Luo, P., Xie, S., Yan, Z.: Going denser with open-vocabulary part segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15453–15465 (2023) 9
42. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025) 4
43. Wan, Z., Xie, Y., Zhang, C., Lin, Z., Wang, Z., Stepputtis, S., Ramanan, D., Sycara, K.: Instructpart: Task-oriented part segmentation with instruction reasoning. In: The 63rd Annual Meeting of the Association for Computational Linguistics (2025), <https://openreview.net/forum?id=IMEr4XgJSZ> 10
44. Wang, H., Wang, S., Zhong, Y., Yang, Z., Wang, J., Cui, Z., Yuan, J., Han, Y., Liu, M., Ma, Y.: Affordance-r1: Reinforcement learning for generalizable affordance reasoning in multimodal large language model. arXiv preprint arXiv:2508.06206 (2025) 2, 3, 5, 6, 9, 10, 11
45. Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, b., Xia, F., Chi, E., Le, Q.V., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 24824–24837. Curran Associates, Inc. (2022) 4
46. wei, Z., Lin, J., Liu, Y., Chen, W., Luo, J., Li, G., Lin, L.: 3daffordsplat: Efficient affordance reasoning with 3d gaussians (2025), <https://arxiv.org/abs/2504.11218> 3
47. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report (2025), <https://arxiv.org/abs/2508.02324> 7, 10
48. Wu, D., Fu, Y., Huang, S., Liu, Y., Jia, F., Liu, N., Dai, F., Wang, T., Anwer, R.M., Khan, F.S., et al.: Ragnet: Large-scale reasoning-based affordance segmentation benchmark towards general grasping. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11980–11990 (2025) 2, 6, 9, 10, 11
49. Wu, W., Mao, S., Zhang, Y., Xia, Y., Dong, L., Cui, L., Wei, F.: 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) 4
50. Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2945–2954 (2023) 9
51. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025) 3
52. Yang, Y., Zhai, W., Luo, H., Cao, Y., Luo, J., Zha, Z.J.: Grounding 3d object affordance from 2d interactions in images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10905–10915 (2023) 3

53. Yu, C., Wang, H., Shi, Y., Luo, H., Yang, S., Yu, J., Wang, J.: Seqafford: Sequential 3d affordance reasoning via multimodal large language model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1691–1701 (2025) 3
54. Zhang, H., Chen, K., Zhang, Z., Chen, H.H., Lyu, Y., Zhang, Y., Yang, S., Zhou, K., Chen, Y.: Dualcamctrl: Dual-branch diffusion model for geometry-aware camera-controlled video generation. arXiv preprint arXiv:2511.23127 (2025) 4
55. Zhang, Z., Chen, K., Lin, X., Jiang, L., Zheng, X., Lyu, Y., Guo, L., Li, Y., Chen, Y.C.: Phystoolbench: Benchmarking physical tool understanding for mllms. arXiv preprint arXiv:2510.09507 (2025) 11
56. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models (2025), <https://arxiv.org/abs/2504.10479> 9, 11
57. Zhu, Z., Wang, X., Zhao, W., Min, C., Deng, N., Dou, M., Wang, Y., Shi, B., Wang, K., Zhang, C., You, Y., Zhang, Z., Zhao, D., Xiao, L., Zhao, J., Lu, J., Huang, G.: Is sora a world simulator? a comprehensive survey on general world models and beyond. arXiv preprint arXiv:2405.03520 (2024) 4