Obstruction reasoning for robotic grasping

Runyu Jiao[1,2] Matteo Bortolon[1] Francesco Giuliari[1] Alice Fasoli[1]
Sergio Povoli[1] Guofeng Mei[1] Yiming Wang[1] Fabio Poiesi[1]

[1]Fondazione Bruno Kessler [2]University of Trento

{rjiao, mbortolon, fgiuliari, alfasoli, spovoli, gmei, ywang, poiesi}@fbk.eu

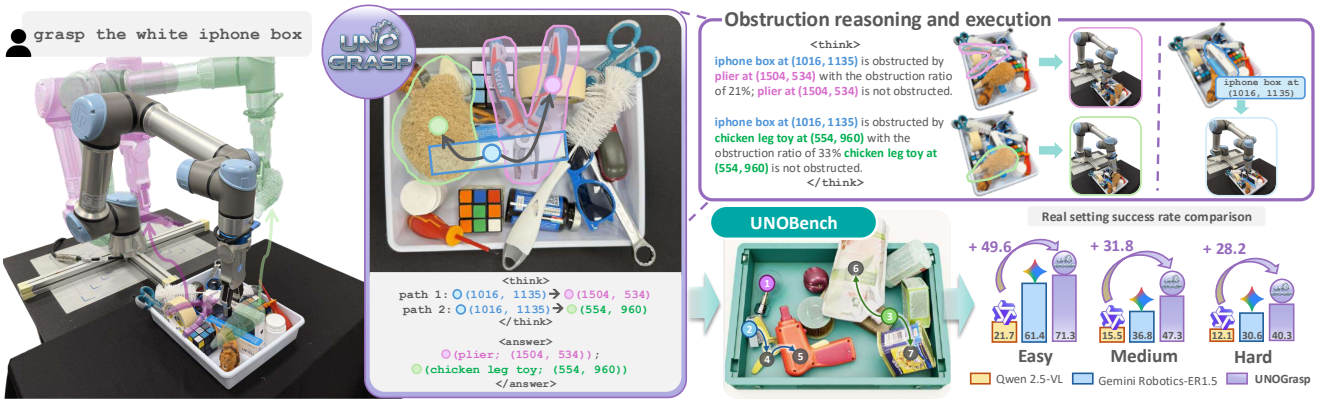

Figure 1. UNOGrasp performs multi-step obstruction reasoning for robotic grasping in cluttered scenes. Given an RGB-D image and a natural-language goal (e.g., `grasp the white iphone box`), UNOGrasp reasons and grounds spatial information to infer the sequence of steps to unobstruct a requested object. We also introduce UNOBench to comprehensively benchmark obstruction reasoning.

Abstract

Successful robotic grasping in cluttered environments not only requires a model to visually ground a target object but also to reason about obstructions that must be cleared beforehand. While current vision-language embodied reasoning models show emergent spatial understanding, they remain limited in terms of obstruction reasoning and accessibility planning. To bridge this gap, we present UNOGrasp, a learning-based vision-language model capable of performing visually-grounded obstruction reasoning to infer the sequence of actions needed to unobstruct the path and grasp the target object. We devise a novel multi-step reasoning process based on obstruction paths originated by the target object. We anchor each reasoning step with obstruction-aware visual cues to incentivize reasoning capability. UNOGrasp combines supervised and reinforcement finetuning through verifiable reasoning rewards. Moreover, we construct UNOBench, a large-scale dataset for both training and benchmarking, based on MetaGraspNetV2, with over 100k obstruction paths annotated by humans with obstruction ratios, contact points, and natural-language instructions. Extensive experiments and real-robot evaluations show that UNOGrasp significantly improves obstruction reasoning and grasp success across both synthetic and real-world environments, outperforming generalist and proprietary alternatives. Project website: https://tev-fbk.github.io/UnoGrasp/.

1. Introduction

Making robots interact with highly cluttered and unstructured 3D environments, such as bin-picking or object assembly following natural-language instructions, is an important skill for robotic manipulation [28]. Successful grasping of a target object that is requested in natural language, demands Vision-Language Models (VLMs) [4] to not only visually ground and differentiate the target object, but also understand inter-object physical dependencies. When objects impose on one another, the resulting physical obstruction can make manipulators fail in cluttered settings, as it prevents the robot's end-effector from successfully accessing the target object [12]. While detection-based approaches can estimate obstruction relationships [18], their design does not extend to broader embodied reasoning or multi-step action planning required by VLMs. Although VLM-based spatial reasoning is crucial for robotic manipulation, emerging benchmarks [10, 15, 24, 29] reveal that existing VLMs are generally limited in spatial reasoning necessary for physical interaction in the embodied context. The challenge inherent to dense, cluttered scenes, where objects physically obstruct one another, still remains largely underexplored.

Preliminary research by Jiao *et al*. [12] explores VLMs' zero-shot ability, exploiting Molmo [9] for visual grounding and prompting GPT-4o [2] to reason whether to clear obstructing objects first. The recent release of (*proprietary*) Gemini Robotics ER [1] features a generalist model that exhibits interesting spatial understanding and grounding abil-

ities. While promising, current research [1, 12] remains shallow in the task formalization and lacks in-depth investigation on how to evaluate and promote obstruction reasoning.

In addition, while classical grasp planning methods typically address cluttered manipulation through grasp affordance modeling and action selection [5, 8, 26, 35], we instead study obstruction understanding as a spatial perception and reasoning problem, abstracted from low-level control.

We advance embodied spatial reasoning for robotic grasping in clutter, primarily focusing on *obstruction reasoning* with the objective of identifying obstruction paths directed from user-requested target objects. We aim to benchmark and enhance existing VLMs on obstruction reasoning capability, in order to unobstruct the paths and promote successful grasp of the target object. To this end, we introduce UNOBench, a dataset for both training and benchmarking VLMs' obstruction reasoning. UNOBench is based on MetaGraspNetV2 [11], featuring diverse daily objects in both synthetic and real scenes. We associate each object with a human-annotated natural-language description to uniquely identify the instance in clutter. UNOBench provides 100k+ obstruction paths with rich metadata, *e.g.* obstruction ratios, contact points, natural-language descriptions. We also propose a set of evaluation metrics to quantify models' reasoning performance at both object and obstruction-path levels.

Moreover, we introduce UNOGrasp, a VLM equipped with novel visually-grounded obstruction reasoning ability for inferring the sequence of accessible obstructing objects that must be removed. UNOGrasp addresses obstruction reasoning via formulating a directed graph with objects as nodes and obstruction relations as edges, allowing it to effectively infer accessible obstructors. UNOGrasp is trained on UNOBench using a two-stage approach: supervised fine-tuning (SFT) to initialize its reasoning capability, then reinforcement fine-tuning (RFT) based on verifiable rewards and obstruction-aware visual cues to boost the model's reasoning. We benchmark UNOGrasp against Gemini Robotics-ER 1.5 [1] and Qwen2.5-VL [4] baselines. UNOGrasp outperforms these baselines in both synthetic and real scenes of UNOBench. Notably, we also conduct real-world robotic experiments in a laboratory environment, confirming UNOGrasp's advantage over Gemini Robotics-ER 1.5 in terms of obstruction reasoning. *Unlike proprietary models, we will release data, model and code publicly.*

In summary, our main contributions are:

- We pioneer the deep study of spatial obstruction reasoning for robotic grasping in challenging cluttered scenes.
- We introduce UNOBench, the first large-scale benchmark for training and testing obstruction reasoning, with evaluation protocols and metrics to quantify reasoning accuracy.
- We propose UNOGrasp, a VLM trained with a novel graph-based recipe that encourages obstruction reasoning with obstruction-aware visual cues, like occlusion ratio.
- UNOBench confirms that obstruction reasoning remains an open challenge in embodied spatial reasoning, while UNOGrasp achieves state-of-the-art performance.

2. Related work

Spatial reasoning with VLMs. VLMs are limited in 3D spatial reasoning despite high VQA performance [15, 29]. Research addresses this by fine-tuning models with explicit 3D knowledge, such as metric distances (SpatialVLM [7]) or depth inputs (SpatialBot [6]). Further progress utilizes reinforcement learning for robotic manipulation tasks [25, 38], exploits intermediate representations [34] or descriptive scene graphs [16], and employs visual prompting techniques (*e.g.*, Set-of-Mark [31, 32]) to enhance reasoning. Affordance understanding is also being integrated (RoboPoint [33], A_0 [30]), and VISO-Grasp [23] tackles visibility constraints. Yet, most works do not consider scenarios where target objects are obstructed, thus hindering successful manipulation, while UNOGrasp addresses this exact challenge.

Obstruction reasoning in robotic grasping. Classical cluttered grasping methods formulate the problem as grasp affordance prediction and sequential manipulation planning based on physical feasibility [5, 8, 26, 35]. Grasping under obstruction requires inferring object geometry and feasible grasps from partial observations [14, 27]. An equally important challenge is inferring the sequence of actions needed to clear complex arrangements (*e.g.*, stacked objects) for access [12, 18, 36]. Recent VLM approaches for cluttered grasping include dedicated planners like RelationGrasp [13], and reasoning models like ThinkGrasp [17] and FreeGrasp [12], which use LLMs for object removal planning and visual prompting, respectively. Unlike prior methods that construct scene-level graphs over all objects [11, 18], UNOGrasp builds a compact obstruction graph centered on the target object, focusing on its accessibility.

Datasets and benchmarks on obstruction reasoning. Progress in obstruction reasoning include early works like VMRN [36] and REGRAD [37] that use object relationships to formulate obstruction graphs, while MetaGraspNetV2 [11] and amodal segmentation datasets like UOAIS-SIM [3] focus on occlusion-related challenges. However, these datasets are not aligned with complex, reasoning-oriented manipulation tasks that require understanding multi-object obstruction chains. VLM benchmarks (*e.g.*, EmbSpatial-Bench [10], Spatial457 [29], CAPTURe [15]) primarily test static perception rather than action-centric obstruction reasoning (i.e., planning clearance actions). However, most datasets lack language annotations, limiting their utility for VLMs and preventing linguistically-grounded obstruction reasoning. To bridge this gap, we introduce UNOBench, the first benchmark with annotated free-form language object descriptions to enable VLMs to jointly reason about occlusions and corresponding unobstructing actions.

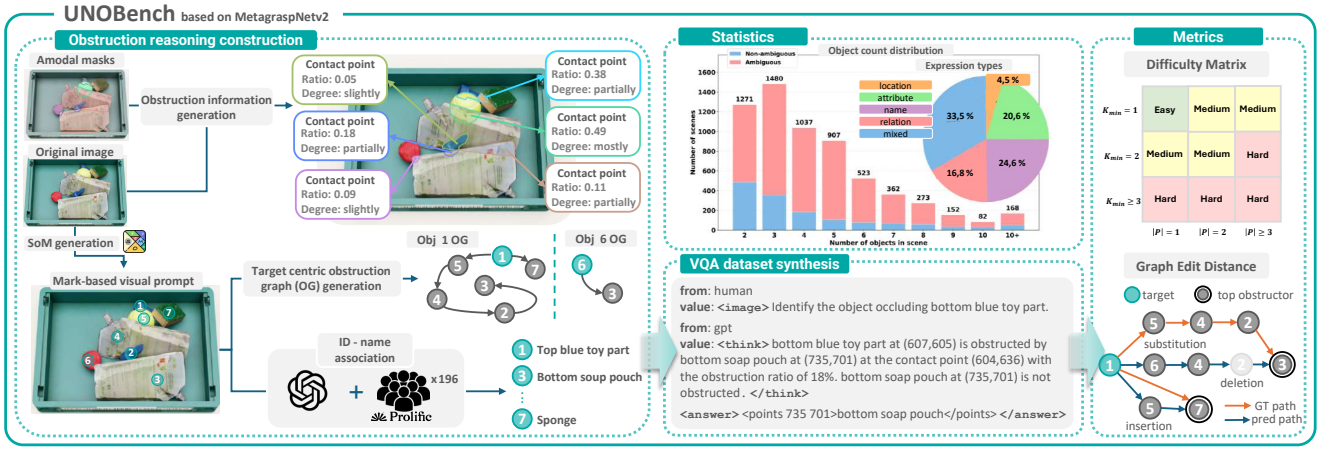

Figure 2. UNOBench features two unique characteristics: (i) human-annotated free-form language instructions about objects in cluttered bins, and (ii) per-bin obstruction graphs for grounded spatial reasoning. Human annotators through the Prolific platform were involved to refine the initial GPT-4o generated annotations. UNOBench features three levels of difficulty and introduces novel evaluation metrics.

3. UNOBench

UNOBench is built upon MetaGraspNetV2 [11] dataset, and is both for training and benchmarking (Fig. 2). MetaGraspNetV2 provides amodal segmentation and object geometry, but it lacks explicit supervision for high-level reasoning and language grounding. UNOBench introduces two unique characteristics: (i) human-annotated free-form language instructions of objects in cluttered bins, (ii) per-bin obstruction graphs for grounded spatial reasoning. UNOBench enables obstruction-aware and language-guided grasping through, *structured obstruction reasoning construction* to enrich each scene with physical obstruction information, object-centric graphs, and semantic knowledge, and *obstruction-aware VQA synthesis* to transform these structured annotations into a VQA dataset suitable for model training and evaluation.

Structured obstruction reasoning construction. We semi-automatically construct a symbolic representation encoding visual obstruction structures in four steps: (a) *Set-of-Marks (SoM) preparation:* we overlay unique numeric marks [31] on each object instance in the ground-truth masks, assigning an ID and centroid (x, y); (b) *Obstruction information:* from amodal masks, we compute contact points, obstruction ratios, and obstruction degrees (slightly, partially, mostly, heavily obstructed); (c) *Object-centric obstruction graph:* for each target, we build a directed graph where nodes are SoM IDs and edges represent "the obstructed $\rightarrow$ the obstructing" relations with associated obstruction attributes; (d) *ID-name-coordinate association:* GPT-4o processes mark-based prompts to generate names for all IDs, forming $(id, name, (x, y))$ triplets. We rely on human annotators to refine 5,400 challenging images, where 196 native speakers on Prolific reviewed 41,193 object names (80 minutes per person), followed by expert rechecking of all scenes, resulting in 4,678 corrected images and 17,261 revised object names, ensuring linguistic accuracy and visual consistency.

Obstruction-aware VQA synthesis. We generate two complementary datasets following the structured `<think>` and `<answer>` format. These two datasets form a unified benchmark: the *Oracle with Set-of-Mark (SoM)* dataset assesses structured reasoning with explicit grounding, while the *Natural Language Prompting* dataset evaluates obstruction reasoning and grounding based on free-form instructions. Specifically: (i) Oracle (SoM): Starting from the object-centric obstruction graph (OG), we use predefined templates to generate questions and reasoning traces. All instances are represented by numeric IDs only (no names or coordinates). This setting solely evaluates the model's reasoning capability, as all object instances are unambiguously identified via SoM. (ii) Natural Language Prompting: Building upon the Oracle formulation, this setting better reflects real-world robot usage, where users' prompts are questions given in free-form language without explicit IDs or coordinates. The model-generated `<think>` and `<answer>` traces include both object names and coordinates at each reasoning step, reflecting realistic human–robot interaction. This dataset measures a model's ability on both obstruction reasoning and spatial grounding with linguistic instructions.

Dataset statistics. UNOBench comprises synthetic and real scenes. The former includes 6,255 scenes with 25,020 view images, 97,066 object instances annotated with names, and 108,174 reasoning paths. The latter includes 520 scenes with one view per scene, 2,232 object instances annotated with names, and 2,552 obstruction paths.

3.1. Evaluation protocol

We introduce our metrics and benchmark split below.

Outcome-level metrics. For each target object, we quantify the correctness of the final answer (the predicted top-level obstruction set, $\mathcal{F}_{\text{pred}}(o_t)$) against the ground truth ($\mathcal{F}_{\text{GT}}(o_t)$).

We report Success Rate Precision (SR-P), Recall (SR-R), and the F1-score, which collectively measure the accuracy of the model's final action output in the <answer> section.

Reasoning-level metrics. We assess the model's reasoning ability in <think> at two levels. *Object-level reasoning* is computed using Object Triplet Precision (OP), Recall (OR), and $F1_{rel}$, as in [18]. For each pair of objects, we deem a true positive when both objects and their obstruction relationship are correctly identified. *Path-level reasoning* is computed using our new metric Multi-Path Normalized Edit Distance (MP_NED). MP_NED measures the structural alignment between predicted ($\mathcal{P}$=$\{p_i\}_{i=1}^{m}$) and ground-truth ($\mathcal{G}$=$\{g_j\}_{j=1}^{n}$) reasoning paths. Formally, $\text{NED}(p_i, g_j) = \frac{\text{EditDist}(p_i, g_j)}{\max(|p_i|, |g_j|)}$, where EditDist is the Levenshtein distance. We then find the minimal-cost assignment via the Hungarian algorithm using $C_{ij} = \text{NED}(p_i, g_j)$ as the cost matrix. MP_NED is the mean cost over matched pairs: $\text{MP\_NED} = \max(m, n)^{-1} \sum_{(i,j)\in\text{match}} C_{ij}$. A lower MP_NED indicates closer structural alignment of the reasoning paths. More details are referred to *Supp. Mat.*

Difficulty-based evaluation split. We categorize the difficulty of each target object based on its obstruction graph depth ($K_{\min}$) and the number of distinct reasoning paths ($|P|$). We divide the benchmark into four difficulty levels (*No-Occ*, *Easy*, *Medium*, and *Hard*), as summarized below:

Level	Condition	Interpretation
No-Occ	$K_{\min} = 0$	No obstruction
Easy	$K_{\min} = 1$, $\lvert P\rvert = 1$	Single-path reasoning
Medium	($K_{\min} = 1$, $\lvert P\rvert > 1$) or ($K_{\min} = 2$, $\lvert P\rvert \leq 2$)	Multi-path or shallow depth
Hard	$K_{\min} \geq 3$ or ($K_{\min} = 2$, $\lvert P\rvert > 2$)	Deep or complex reasoning

4. UNOGrasp

Problem formulation. Let us define a cluttered workspace with a set of N visible objects as $\mathcal{O} = \{o_1, o_2, \ldots, o_N\}$ (Fig. 2). Given an RGB-D observation $I = (I_{rgb}, I_d)$ and a free-form textual instruction $q \in \mathcal{L}$ (e.g., `grasp the white box on the leftmost`) that uniquely refers to a target object $o_t \in \mathcal{O}$, we aim to produce an unobstruction plan to grasp o_t. I_{rgb} is used for reasoning and action sequence planning. I_d is used to estimate 3D grasping points. If o_t is unobstructed, UNOGrasp instructs a direct grasp. Otherwise, UNOGrasp identifies the minimal sequence of actions to access o_t. This sequence begins by identifying all top-level obstructing objects, those that obstruct o_t but are themselves accessible, *i.e.* free of obstruction. Each action corresponds to an object removal. Note that our action plan is purely based on the obstruction's existence. Various obstructions may impede grasping differently, thus quantifying obstruction severity is highly challenging and application-dependent, out of the scope of this paper.

UNOGrasp formulates obstruction reasoning as a directed graph. We train it on a portion of UNOBench's synthetic set using a two-stage approach: supervised fine-tuning (SFT) to initialize its reasoning capability, then reinforcement fine-tuning (RFT) based on obstruction-aware visual cues to boost model's reasoning. Fig. 3 illustrates UNOGrasp stages.

4.1. Target-centric obstruction graph

Given the instruction q in free-form language, the model performs spatial reasoning to ground the linguistic reference in I_{rgb} and identify o_t among all visible objects. If o_t is not initially visible (*e.g.* deeply beneath), the model uses contextual cues to infer the most plausible candidate. Once o_t is localized, we model its visual obstruction. Instead of reasoning over all possible pairwise obstruction relations [18], we construct a target-centric obstruction graph that exclusively captures the objects relevant to the accessibility of o_t.

The sequence of actions required to unobstruct the target object o_t can be modeled as a directed graph, $G_t = (\mathcal{V}_t, E_t)$. The node set $\mathcal{V}_t$ includes the target object o_t and all objects that directly or indirectly obstruct it. A directed edge $(o_i, o_j) \in E_t$ indicates that object o_i is obstructed by object o_j when viewed from the camera viewpoint. Edges are directed from the obstructed object to the obstructors, forming one or more obstruction paths that originate at the target object o_t and terminate at the accessible top-level obstructors. Any object that appears along one of these obstruction paths can be considered an ancestor of o_t, and must be removed before o_t becomes fully accessible. Let $\mathcal{A}(o_t)$ be the set of ancestor objects of o_t that is defined as

$$\mathcal{A}(o_t) = \{\, o_i \in \mathcal{O} \mid \exists \text{ a directed path } [o_t, \cdots, o_i] \in G_t \,\}. \quad (1)$$

Objects that lie on top of the clutter are defined as

$$\mathcal{F}(o_t) = \{\, o_i \in \mathcal{A}(o_t) \mid \nexists\, o_j \text{ s.t. } (o_i, o_j) \in E_t \,\}. \quad (2)$$

Each $o_i \in \mathcal{F}(o_t)$ is a visible and graspable object; removing any of these will reduce the obstruction of o_t.

Lastly, we express the reasoning objective as:

$$f_\Theta(I, q) = \begin{cases} \mathcal{A}(o_t) \to \mathcal{F}(o_t), & \text{if } o_t \text{ is obstructed,} \\ o_t, & \text{otherwise.} \end{cases} \quad (3)$$

f_Θ is parametric, with Θ being its parameters. We aim to train f_Θ to output $\mathcal{F}(o_t)$ through obstruction reasoning $\mathcal{A}(o_t)$. When multiple top-level obstructors exist, $\mathcal{F}(o_t)$ provides a set of next-step candidates, from which the robot can select based on constraints like graspability, or reachability.

Fig. 3 shows three examples of target-centric obstruction graphs: G_1, G_2, and G_3. For the light bulb (o_1), which is unobstructed, its graph G_1 contains only the object itself. G_2 represents the obstruction graph of the stapler (o_2). The stapler is blocked by the yellow propeller (o_4), which is

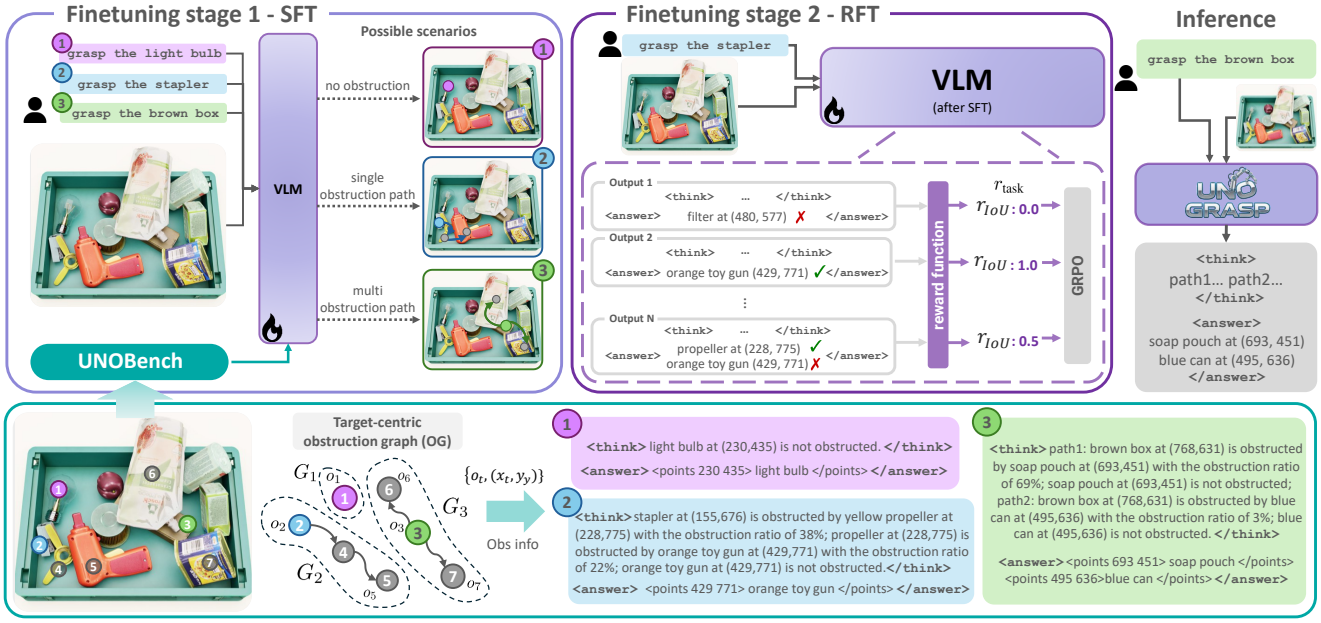

Figure 3. UNOGrasp is a VLM trained through supervised fine (SFT) on UNOBench to learn structured obstruction-path reasoning, and through GRPO-based reinforcement finetuning (RFT) to further boost its reasoning ability using outcome-driven IoU and format rewards. During inference, given an RGB image and a target object as language instruction, UNOGrasp reasons over multiple obstruction paths (<think> traces) and directly outputs the sequence of actions (<answer>) required to remove obstructions and grasp the target.

further blocked by the orange toy gun (o_5). Thus, both objects are ancestors, $\mathcal{A}(o_2) = \{o_4, o_5\}$. However, only the orange toy gun is itself unobstructed, making it the top-level obstructing set: $\mathcal{F}(o_2) = \{o_5\}$. In G_3, both the soap pouch (o_6) and blue can (o_7) obstruct the brown box (o_3) and are themselves unobstructed, meaning $\mathcal{F}(o_3) = \mathcal{A}(o_3) = \{o_6, o_7\}$. Ultimately, the model's objective is to accurately infer these minimal obstruction sets, $\mathcal{F}(o_t)$, which guide the robot in deciding the next necessary action. We next describe how f_Θ is trained to achieve this reasoning capability.

4.2. Training pipeline

We develop a two-stage training pipeline to train f_Θ (Eq. 3) with visually grounded, obstruction-aware reasoning: i) warm-start supervised finetuning (SFT) that encourages the model to output visually grounded reasoning chains aligned with the obstruction graph of the referred target; ii) reinforcement finetuning (RFT) that optimizes task-relevant behaviors using our novel obstruction-aware rewards.

SFT with visually-grounded chains. We finetune f_Θ on UNOBench (§3) to interpret free-form language instructions (q) and associate them with a unique target object (o_t) in the visual scene. This grounding is supervised using two methods for explicit reference: the object's name and its image coordinates $\{o_t, (x_t, y_t)\}$, and a SoM visual prompt, where objects are assigned unique IDs. These explicit cues are essential for disambiguating multiple instances based on spatial or relational cues in q. The fine-tuning instructs f_Θ to identify if (i) o_t is unobstructed, (ii) o_t has a single obstruction path, and (iii) o_t has multiple obstruction paths. The model generates a step-by-step reasoning chain where every step is anchored to a physically adjacent (contacting) neighbor, ensuring the chain traverses valid obstructions. We encourage f_Θ to quantify obstruction levels as auxiliary signals within the chain, aligning with findings that spatial grounding encourages visual reasoning [20]. This process strengthens the model's ability to reconstruct complete obstruction paths and identify the top-level obstructors $\mathcal{F}(o_t)$ that constitute the next-step action set.

RFT with obstruction-aware rewards. Starting from the SFT-bootstrapped model f_Θ, we perform RFT to enhance its grounded reasoning ability. The model produces a natural language description and corresponding image coordinates for each object, enabling f_Θ to progressively refine its attention using task-relevant visual information [20]. We adopt Group Relative Policy Optimization (GRPO) [21] to average rewards across multiple sampled outputs, using a standard formulation similar to related works [20, 38]. Importantly, we define a novel, task-specific reward r as a weighted combination of a format reward r_{fmt} and a task reward r_{task}:

$$r = \lambda_{\text{fmt}}\, r_{\text{fmt}} + \lambda_{\text{task}}\, r_{\text{task}}. \tag{4}$$

The format reward, r_{fmt}, is binary (1 or 0), promoting structural validity by checking for the correct presence and closure of the reasoning <think> and action <answer> contexts. The task reward, r_{task}, supervises the grounded output $\mathcal{F}(o_t)$ (the content of the action context) using a set-level

Table 1. Path-level reasoning results on UNOBench synthetic test set. ICL: In-Context Learning. SFT: Supervised finetuning. SR: Success Rate; MP_NED: Multi-Path Normalized Edit Distance; P: Precision; R: Recall; F1: F1-Score; Best results **bold**; second best underlined.

Method	No obstructions		Easy				Medium				Hard			
	SR (%)↑	MP_NED↓	SR-P (%)↑	SR-R (%)↑	SR-F1 (%)↑	MP_NED↓	SR-P (%)↑	SR-R (%)↑	SR-F1 (%)↑	MP_NED↓	SR-P (%)↑	SR-R (%)↑	SR-F1 (%)↑	MP_NED↓
Oracle (with SoM)														
Gemini Robotics-ER 1.5 [1]	68.7	0.17	57.7	62.8	59.3	0.25	33.0	30.2	29.8	0.56	5.9	5.7	5.4	0.74
Gemini Robotics-ER 1.5 [1] (ICL)	54.6	0.25	67.1	73.3	69.1	0.20	41.8	39.7	38.7	0.50	14.9	14.8	13.8	0.68
Qwen2.5-VL [4] (ICL)	9.8	0.64	24.7	36.7	27.2	0.55	20.7	27.2	20.3	0.73	9.7	17.2	10.2	0.79
Qwen2.5-VL [4] (SFT)	88.7	-	69.6	70.3	69.8	-	64.2	53.1	56.5	-	38.1	33.3	34.3	-
UNOGrasp	**94.8**	**0.03**	**82.8**	**84.4**	**83.3**	**0.11**	**74.8**	**67.2**	**69.1**	**0.37**	**56.8**	**55.3**	**54.5**	**0.51**
Natural Language Prompting														
Gemini Robotics-ER 1.5 [1]	50.2	0.88	51.8	52.8	52.1	0.84	36.9	30.9	32.5	0.87	11.7	9.4	10.1	0.91
Gemini Robotics-ER 1.5 [1] (ICL)	45.3	0.83	60.6	61.8	61.0	0.80	45.5	37.3	39.5	0.85	17.2	13.4	14.6	0.89
Qwen2.5-VL [4] (ICL)	11.2	0.84	11.8	13.1	12.2	0.86	11.8	10.8	10.6	0.88	9.7	9.4	8.9	0.89
Qwen2.5-VL [4] (SFT)	91.4	-	65.2	65.5	65.3	-	59.6	47.9	51.5	-	33.9	31.5	31.9	-
UNOGrasp	**92.5**	**0.06**	**74.8**	**75.1**	**74.9**	**0.20**	**68.0**	**55.8**	**59.7**	**0.53**	**42.2**	**35.4**	**37.2**	**0.67**

Table 2. Object-level reasoning results on UNOBench synthetic test set. ICL: In-Context Learning; SFT: Supervised fine tuning; OP: Object triplet Precision; OR: Object triplet Recall; $F1_{rel}$: Object triplet F1-Score; Best results **bold**; second best underlined.

Method	Easy			Medium			Hard			Overall		
	OP↑	OR↑	$F1_{rel}$↑	OP↑	OR↑	$F1_{rel}$↑	OP↑	OR↑	$F1_{rel}$↑	OP↑	OR↑	$F1_{rel}$↑
Oracle (with SoM)												
Gemini Robotics-ER 1.5 [1]	56.0	62.3	57.9	57.0	30.6	37.4	46.1	14.4	20.8	56.0	51.2	50.6
Gemini Robotics-ER 1.5 [1] (ICL)	67.8	76.5	70.4	69.1	39.8	47.3	**67.6**	24.3	33.0	68.2	63.8	62.3
Qwen2.5-VL [4] (ICL)	19.6	33.3	22.8	21.4	19.7	18.5	13.9	10.8	10.6	20.0	28.5	21.1
UNOGrasp (ours)	**81.3**	**85.9**	**82.6**	**77.8**	**57.3**	**62.0**	63.0	**43.6**	**48.7**	**79.7**	**76.0**	**75.3**
Natural Language Prompting												
Gemini Robotics-ER 1.5 [1]	2.5	2.7	2.6	4.6	2.1	2.8	7.8	2.0	3.1	3.3	2.5	2.6
Gemini Robotics-ER 1.5 [1] (ICL)	3.4	3.7	3.5	6.2	2.6	3.5	6.1	1.7	2.6	4.3	3.3	3.5
Qwen2.5-VL [4] (ICL)	2.1	3.6	2.5	4.0	2.9	3.1	3.7	2.2	2.7	2.7	3.3	2.7
UNOGrasp (ours)	**65.6**	**67.4**	**66.1**	**57.4**	**33.0**	**39.5**	**44.6**	**20.1**	**25.7**	**62.6**	**56.0**	**57.2**

(Intersection over Union) IoU metric:

$$r_{\text{task}} = \frac{|\mathcal{F}_{\text{pred}}(o_t) \cap \mathcal{F}_{\text{gt}}(o_t)|}{|\mathcal{F}_{\text{pred}}(o_t) \cup \mathcal{F}_{\text{gt}}(o_t)|}. \quad (5)$$

This IoU provides a smoother optimization signal than binary correctness, rewarding partially correct predictions for more stable learning. Although this reward only supervises the final prediction $\mathcal{F}(o_t)$, experiments in §5.2 show it also contributes to improving the quality of the internal obstruction reasoning path $\mathcal{A}(o_t)$. For completeness, a path-level fidelity metric r_{path} is also evaluated post hoc.

5. Experiments

We compare UNOGrasp against two VLMs. We use Gemini Robotics-ER 1.5 [1] as proprietary baseline in two variants: base model, provided with a prompt and real output examples, and In-Context Learning (ICL), prompted with three few-shot examples covering all obstruction types (no, single-path, and multi-path obstruction). Coordinate expressions are adapted to the model's native syntax for both. We use Qwen2.5-VL-3B [4] as open-source baseline in two variants: ICL, prompted as for Gemini ICL; SFT, finetuned on UNOBench using the same supervised setup as UNOGrasp but without the `<think>` reasoning part; UNOGrasp is built on Qwen2.5-VL-3B, and trained with SFT and RFT. We evaluate all methods on both the *oracle (SoM)* and *natural language prompting* splits of UNOBench. The synthetic scenes are split into training, validation, and testing sets with a 7:1:2 ratio. All the real scenes are exclusively used for testing. We follow the procedure detailed in §3.1.

Implementation details. We train f_Θ on 4 A100-SXM-64GB GPUs, using 2 epochs for SFT, and 1 epoch for RFT (with a generation group size of 4). For Gemini Robotics-ER 1.5, temperature is set to 0.1, and thinking budget to 2000.

5.1. Quantitative analysis

Tab. 1 and 2 report path-level and object-level obstruction reasoning results, respectively, on the UNOBench synthetic test set. Similarly, Tab. 3 and 4 report these results on the UNOBench real set. *UNOBench enhances both reasoning and decision accuracy.* Models finetuned on UNOBench (Qwen2.5-VL (SFT) and UNOGrasp) achieve important gains in recognizing top obstructors compared to Qwen2.5-VL (ICL). UNOGrasp improvement in reasoning quality is larger, and surpasses the proprietary Gemini Robotics-ER 1.5 across most settings, even in real subsets that are never seen during training. *Reasoning ability is crucial for complex scenes.* As the obstruction level increases, the advantage of reasoning supervision grows. On the synthetic hard split, UNOGrasp surpasses Qwen2.5-VL (SFT) by +20.2% SR-F1; on the real hard split, the margin widens to +38.0%, confirming that process-level supervision is important for multi-path reasoning. *Hallucination might occur with no obstructions.* In both synthetic and real *No obstructions* settings, Gemini Robotics-ER 1.5 attains low SR (68.7% and 36.0%), often hallucinating obstructions even when the target is fully accessible. Qwen2.5-VL (ICL) performs even worse under the same condition. *Reasoning quality aligns with final accuracy.* A clear correlation is observed between

Table 3. Path-level reasoning results on UNOBench real set. ICL: In-Context Learning. SFT: Supervised fine tuning. SR: Success Rate; MP_NED: Multi-Path Normalized Edit Distance; P: Precision; R: Recall; F1: F1-Score. Best result **bold**; second best underlined.

Method	No obstructions		Easy				Medium				Hard			
	SR (%)↑	MP_NED↓	SR-P (%)↑	SR-R (%)↑	SR-F1 (%)↑	MP_NED↓	SR-P (%)↑	SR-R (%)↑	SR-F1 (%)↑	MP_NED↓	SR-P (%)↑	SR-R (%)↑	SR-F1 (%)↑	MP_NED↓
Oracle (with SoM)														
Gemini Robotics-ER 1.5 [1]	36.0	0.85	47.7	48.3	47.9	0.80	40.8	28.2	32.1	0.90	31.8	27.3	28.8	0.91
Gemini Robotics-ER 1.5 [1] (ICL)	35.0	0.83	60.1	62.9	60.9	0.80	44.2	34.3	36.4	0.89	38.6	27.3	30.6	0.92
Qwen2.5-VL [4] (ICL)	1.3	0.56	42.4	47.2	43.5	0.37	38.3	28.2	30.0	0.68	24.4	26.1	22.7	0.79
Qwen2.5-VL [4] (SFT)	69.7	-	70.2	70.6	70.4	-	62.4	41.6	48.0	-	38.6	21.6	25.9	-
UNOGrasp	**72.5**	**0.16**	**76.2**	**79.0**	**77.2**	**0.15**	**76.6**	**59.6**	**64.4**	**0.40**	**79.5**	**59.1**	**63.9**	**0.55**
Natural Language Prompting														
Gemini Robotics-ER 1.5 [1]	37.7	0.85	50.3	51.0	50.5	0.79	47.3	32.7	37.1	0.89	38.9	33.3	35.2	0.89
Gemini Robotics-ER 1.5 [1] (ICL)	35.1	0.83	60.5	63.3	61.4	0.80	44.7	34.6	36.8	0.88	38.6	27.3	30.6	0.92
Qwen2.5-VL [4] (ICL)	10.0	0.83	21.3	22.6	21.7	0.79	19.6	13.8	15.5	0.86	11.4	13.6	12.1	0.90
Qwen2.5-VL [4] (SFT)	**70.0**	-	64.0	64.5	64.2	-	61.6	37.9	45.4	-	29.5	19.3	21.8	-
UNOGrasp	**70.0**	**0.23**	**71.1**	**71.8**	**71.3**	**0.26**	**62.9**	**40.1**	**47.3**	**0.63**	**54.5**	**35.2**	**40.3**	**0.76**

Table 4. Object-level reasoning results on UNOBench real set. ICL: In-Context Learning; SFT: Supervised fine tuning; OP: Object triplet Precision; OR: Object triplet Recall; $F1_{rel}$: Object triplet F1-Score; Best results **bold**; second best underlined.

Method	Easy			Medium			Hard			Overall		
	OP↑	OR↑	$F1_{rel}$↑	OP↑	OR↑	$F1_{rel}$↑	OP↑	OR↑	$F1_{rel}$↑	OP↑	OR↑	$F1_{rel}$↑
Oracle (with SoM)												
Gemini Robotics-ER 1.5 [1]	64.9	72.9	67.1	66.7	43.3	49.4	41.7	19.3	24.4	64.9	62.9	60.8
Gemini Robotics-ER 1.5 (ICL) [1]	74.6	81.8	76.7	**75.0**	46.6	54.3	**56.4**	18.0	26.4	**74.3**	69.9	68.8
Qwen2.5-VL (ICL) [4]	36.0	39.6	36.8	36.8	18.6	23.3	30.9	10.2	13.1	36.1	32.7	32.2
UNOGrasp	**75.7**	**84.7**	**78.1**	69.0	**57.2**	**58.2**	55.6	**43.9**	**44.7**	73.2	**75.5**	**71.4**
Natural Language Prompting												
Gemini Robotics-ER 1.5 [1]	3.7	4.0	3.8	2.9	1.4	1.8	10.2	2.6	3.8	3.6	3.2	3.2
Gemini Robotics-ER 1.5 [1] (ICL)	3.2	3.6	3.4	4.8	2.5	3.2	9.1	1.7	2.8	3.8	3.2	3.3
Qwen2.5-VL [4] (ICL)	5.9	6.9	6.2	5.1	2.7	3.3	11.4	3.2	4.9	5.8	5.6	5.3
UNOGrasp	**56.9**	**58.9**	**57.5**	**47.0**	**26.1**	**32.1**	**35.6**	**15.6**	**20.6**	**53.5**	**48.2**	**49.1**

Table 5. Ablation study on SFT on synthetic set (Overall).

Method	SR-F1 ↑	OR-F1 ↑	MP_NED ↓
Baseline	74.7	71.9	0.220
+ Contact point	75.3 (+0.6)	72.5 (+0.6)	0.216 (-0.004)
+ Degree word	75.1 (+0.4)	72.5 (+0.6)	0.217 (-0.003)
+ Occlusion ratio	**76.4** (+1.7)	**73.3** (+1.4)	**0.210** (-0.010)

MP_NED and SR-F1, where lower MP_NED consistently coincides with higher SR across all difficulty levels. *Limited effectiveness of ICL.* ICL generally enhances reasoning performance, but it tends to amplify hallucinations in non-obstruction cases. *Spatial grounding of multi-step reasoning can fail*. Baseline models often misalign reasoning steps with object coordinates, causing identity confusion and a high MP_NED (> 0.8). The low scores (mostly below 10) in Tab. 2 and 4 in the Natural Language Prompting setting indicate poor performance in spatial grounding.

5.2. Ablation studies

SFT with obstruction information. Tab. 5 shows results when adding obstruction cues during SFT instead of using only the obstruction graph. Contact point, Degree word, and Occlusion ratio cues improve success rates across all difficulty levels, with the largest gains on Hard cases. Ratio yields the most significant improvement ($+5.8\%$ SR-F1 on Hard), increasing precision and decreasing the reasoning error. The full table is provided in *Supp. Mat.*

RFT with obstruction-aware reward. Tab. 6 shows how results change when applying a set-level IoU reward to the predicted answers during RFT. This yields to consistent improvements across all metrics, with SR-F1 gains increasing with complexity: Easy ($+1.5\%$), Medium ($+2.0\%$), and Hard ($+4.4\%$). Greater improvements under severe obstruction indicate the IoU reward effectively encourages complete answer sets, proving beneficial over binary correctness when multiple ground-truth obstructors exist. Though applied only to `<answer>` output, reasoning metrics (OR-F1 and MP_NED) also improve, suggesting that promoting complete answers indirectly guides more faithful reasoning traces.

5.3. Qualitative analysis

Fig. 4 visualizes obstruction reasoning traces produced by UNOGrasp, Gemini Robotics-ER1.5 (ICL)[1], and Qwen2.5-VL (ICL)[4] in both synthetic and real settings. For each setting, we present examples across the three difficulty levels, along with two failure modes: incorrect reasoning but correct answers, and failures in both. We observe that Qwen2.5-VL (ICL) struggles with spatial grounding even on easy cases. Gemini tends to terminate prematurely during the multi-step reasoning in complex scenarios, missing top-level obstructors when multiple obstruction paths exist. In contrast, UNOGrasp handles multi-path reasoning effectively but can fail with visually similar objects or densely clustered groups.

5.4. Laboratory robotic experiments

We validate UNOGrasp on a UR5e robotic platform across 30 real-world scenarios with 25 distinct objects. Objects are placed in a bin and captured by a top-down ZED 2 camera. We compare against Qwen2.5-VL and Gemini Robotics-ER 1.5 (both ICL), using GroundedSAM [19] and GraspNet [11] for grasp pose prediction [12]. Methods predict and execute

Table 6. Ablation study on RFT (Synthetic set). The harder the scene is, the higher the contribution of RFT.

Variant	Easy			Medium			Hard			Overall		
	SR-F1↑	OR-F1↑	MP_NED↓	SR-F1↑	OR-F1↑	MP_NED↓	SR-F1↑	OR-F1↑	MP_NED↓	SR-F1↑	OR-F1↑	MP_NED↓
Baseline (SFT)	81.8	80.9	0.115	67.1	59.4	0.389	50.1	46.9	0.525	76.4	73.3	0.210
+ RFT on Answer	83.3 (+1.5)	82.6 (+1.7)	0.109 (-0.006)	69.1 (+2.0)	62.0 (+2.6)	0.370 (-0.019)	54.5 (+4.4)	48.7 (+1.8)	0.507 (-0.018)	78.2 (+1.8)	75.3 (+2.0)	0.201 (-0.009)

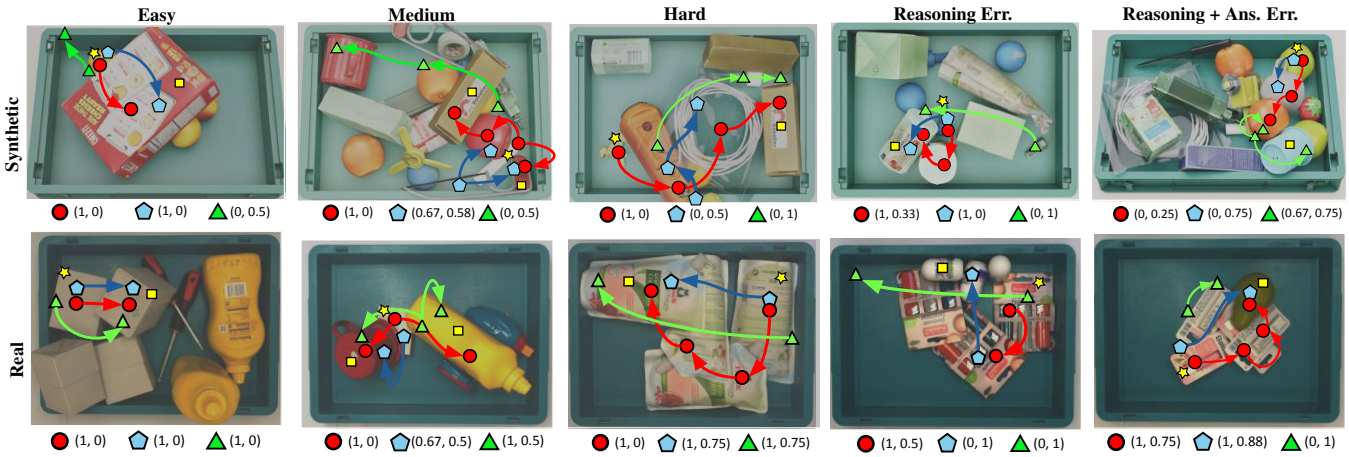

Figure 4. Qualitative results on UNOBench different splits, and in two types of failure. ☆ mark the target object, □ the top obstructor, ● UNOGrasp, ⬠ Gemini Robotics-ER 1.5, and △ Qwen2.5-VL (ICL) predictions with their reasoning traces. (SR-F1/MP_NED) scores are reported at the bottom of each image.

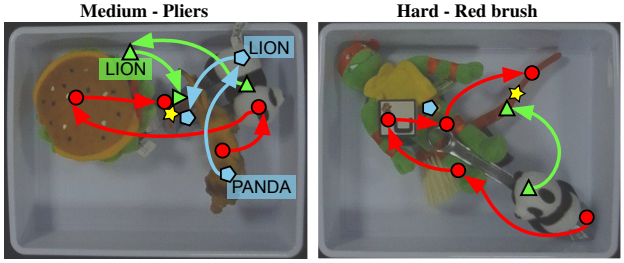

Figure 5. Qualitative results from laboratory robotics experiments. ☆ mark the target object, ● UNOGrasp, ⬠ Gemini Robotics-ER 1.5, and △ Qwen2.5-VL (ICL). Labels are shown for misaligned predictions (labels-spatial location disagreement). Difficulty level and target prompt are display at the top of the figure.

the next object removal. We measure the success ratio, deeming the action successful when all the objects are removed in the correct order to achieve the final grasping goal.

In Tab. 7, UNOGrasp matches Gemini Robotics-ER 1.5 performance on Easy and Medium, despite training on open-source synthetic data, and significantly outperforms it on Hard (+30%). Real-world conditions can feature high-contrast situations (Fig. 5), where black backgrounds and white bins create over-exposure conditions to which Gemini appears more robust. Both baselines exhibit similar failure patterns identified in §5.3, *i.e.*, regarding spatial grounding, which become more frequent in longer obstruction chains. Qwen2.5-VL demonstrates a tendency to over-reason, predicting the container itself as an obstruction even in trivial target-only scenarios.

Table 7. Real-world robotic experiment with the UR5e, success ratios across 30 scenarios with 25 objects.

Method	Easy	Medium	Hard	Average
Gemini Robotics-ER 1.5 [1]	**80%**	**30%**	10%	40%
Qwen2.5-VL [4]	10%	0%	0%	3%
UNOGrasp	**80%**	**30%**	**40%**	**50%**

6. Conclusions

We addressed the critical limitation of VLMs in obstruction reasoning, which can be used for robotic grasping. Our contributions are twofold: we introduced UNOBench, a novel dataset, based on MetaGraspNetV2 [11], featuring $100k+$ annotated obstruction paths for developing and benchmarking; and we proposed UNOGrasp, a VLM trained via sequential SFT and RFT with visually-grounded and obstruction-aware rewards. Experiments showed UNOGrasp significantly improved obstruction reasoning, achieving 78.2% precision on average on UNOBench, and 50% success rate on laboratory robotics experiments with a setup (*e.g.*, objects, camera) different from that of finetuning data. Importantly, real-robot evaluations confirm the advantage of UNOGrasp over existing generalist and proprietary model in difficult scenes, with good generalization capability despite being trained only with synthetic data. Future work can focus on scaling UNOGrasp and UNOBench, incorporating multi-view perception and further expanding the object diversity.

Acknowledgments

This work was supported by Fondazione VRT under the project Make Grasping Easy, by the Ministero delle Imprese e del Made in Italy (IPCEI Cloud DM 27 giugno 2022 – IPCEI-CL-0000007), and by FAIR (PE00000013), the European Union (Next Generation EU). We also acknowledge ISCRA for granting us access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking and hosted by CINECA (Italy).

References

[1] Abbas Abdolmaleki, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Ashwin Balakrishna, Nathan Batchelor, Alex Bewley, Jeff Bingham, Michael Bloesch, et al. Gemini robotics 1.5: Pushing the frontier of generalist robots with advanced embodied reasoning, thinking, and motion transfer. *arXiv:2510.03342*, 2025. 1, 2, 6, 7, 8

[2] 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:2303.08774*, 2023. 1

[3] Seunghyeok Back, Joosoon Lee, Taewon Kim, Sangjun Noh, Raeyoung Kang, Seongho Bak, and Kyoobin Lee. Unseen object amodal instance segmentation via hierarchical occlusion modeling. In *ICRA*, 2022. 2

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv:2502.13923*, 2025. 1, 2, 6, 7, 8, 4

[5] Abdeslam Boularias, James Bagnell, and Anthony Stentz. Learning to manipulate unknown objects in clutter by reinforcement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015. 2

[6] Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. In *ICRA*, 2025. 2

[7] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, 2024. 2

[8] Michael Danielczuk, Andrey Kurenkov, Ashwin Balakrishna, Matthew Matl, David Wang, Roberto Martín-Martín, Animesh Garg, Silvio Savarese, and Ken Goldberg. Mechanical search: Multi-step retrieval of a target object occluded by clutter. In *2019 international conference on robotics and automation (ICRA)*, pages 1614–1621. IEEE, 2019. 2

[9] Matt Deitke et al. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. In *CVPR*, 2025. 1

[10] Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. *arXiv:2406.05756*, 2024. 1, 2

[11] Maximilian Gilles, Yuhao Chen, Emily Zhixuan Zeng, Yifan Wu, Kai Furmans, Alexander Wong, and Rania Rayyes. MetaGraspNetV2: All-in-One Dataset Enabling Fast and Reliable Robotic Bin Picking via Object Relationship Reasoning and Dexterous Grasping. *IEEE TASE*, 2024. 2, 3, 7, 8, 12

[12] Runyu Jiao, Alice Fasoli, Francesco Giuliari, Matteo Bortolon, Sergio Povoli, Guofeng Mei, Yiming Wang, and Fabio Poiesi. Free-form language-based robotic reasoning and grasping. In *IROS*, 2025. 1, 2, 7, 5

[13] Songting Liu, Tat Joo Teo, Zhiping Lin, and Haiyue Zhu. Relationgrasp: Object-oriented prompt learning for simultaneously grasp detection and manipulation relationship in open vocabulary. In *IROS*, 2024. 2

[14] Weiheng Liu, Yuxuan Wan, Jilong Wang, Yuxuan Kuang, Xuesong Shi, Haoran Li, Dongbin Zhao, Zhizheng Zhang, and He Wang. Fetchbot: Learning generalizable object fetching in cluttered scenes via zero-shot sim2real. In *CoRL*, 2025. 2

[15] Atin Pothiraj, Elias Stengel-Eskin, Jaemin Cho, and Mohit Bansal. Capture: Evaluating spatial reasoning in vision language models via occluded object counting. In *ICCV*, 2025. 1, 2

[16] Zekun Qi, Wenyao Zhang, Yufei Ding, Runpei Dong, Xinqiang Yu, Jingwen Li, Lingyun Xu, Baoyu Li, Xialin He, Guofan Fan, et al. Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation. *arXiv:2502.13143*, 2025. 2

[17] Yaoyao Qian, Xupeng Zhu, Ondrej Biza, Shuo Jiang, Linfeng Zhao, Haojie Huang, Yu Qi, and Robert Platt. Thinkgrasp: A vision-language system for strategic part grasping in clutter. In *CoRL*, 2024. 2, 5

[18] Paolo Rabino and Tatiana Tommasi. A modern take on visual relationship reasoning for grasp planning. *IEEE RAL*, 2025. 1, 2, 4

[19] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 7, 12

[20] Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J. Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning. *arXiv:2505.23678*, 2025. 5

[21] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv:2402.03300*, 2024. 5

[22] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025. 4

[23] Yitian Shi, Di Wen, Guanqi Chen, Edgar Welte, Sheng Liu, Kunyu Peng, Rainer Stiefelhagen, and Rania Rayyes. Visograsp: Vision-language informed spatial object-centric 6-dof active view planning and grasping in clutter and invisibility. In *IROS*, 2025. 2

[24] Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *CVPR*, 2025. 1

[25] Zirui Song, Guangxian Ouyang, Mingzhe Li, Yuheng Ji, Chenxi Wang, Zixiang Xu, Zeyu Zhang, Xiaoqing Zhang, Qian Jiang, Zhenhao Chen, et al. Maniplvm-r1: Reinforcement learning for reasoning in embodied manipulation with large vision-language models. *arXiv:2505.16517*, 2025. 2

[26] Bingjie Tang and Gaurav S Sukhatme. Selective object rearrangement in clutter. In *Conference on Robot Learning*, pages 1001–1010. PMLR, 2023. 2

[27] Georgios Tziafas, Yucheng XU, Arushi Goel, Mohammadreza Kasaei, Zhibin Li, and Hamidreza Kasaei. Language-guided robot grasping: CLIP-based referring grasp synthesis in clutter. In *CoRL*, 2023. 2

[28] Che Wang, Jeroen van Baar, Chaitanya Mitash, Shuai Li, Dylan Randle, Weiyao Wang, Sumedh Sontakke, Kostas E Bekris, and Kapil Katyal. Demonstrating multi-suction item picking at scale via multi-modal learning of pick success. *arXiv:2506.10359*, 2025. 1

[29] Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M de Melo, Jieneng Chen, and Alan Yuille. Spatial457: A diagnostic benchmark for 6d spatial reasoning of large multimodal models. In *CVPR*, 2025. 1, 2

[30] Rongtao Xu, Jian Zhang, Minghao Guo, Youpeng Wen, Haoting Yang, Min Lin, Jianzheng Huang, Zhe Li, Kaidong Zhang, Liqiong Wang, et al. A0: An affordance-aware hierarchical model for general robotic manipulation. In *ICCV*, 2025. 2

[31] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4v. *arXiv:2310.11441*, 2023. 2, 3

[32] Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, et al. Magma: A foundation model for multimodal ai agents. In *CVPR*, 2025. 2

[33] Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction in robotics. In *CoRL*, 2024. 2

[34] Yifu Yuan, Haiqin Cui, Yibin Chen, Zibin Dong, Fei Ni, Longxin Kou, Jinyi Liu, Pengyi Li, Yan Zheng, and Jianye Hao. From seeing to doing: Bridging reasoning and decision for robotic manipulation. *arXiv:2505.08548*, 2025. 2

[35] Andy Zeng, Shuran Song, Kuan-Ting Yu, Elliott Donlon, Francois R Hogan, Maria Bauza, Daolin Ma, Orion Taylor, Melody Liu, Eudald Romo, et al. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. *The International Journal of Robotics Research*, 41(7):690–705, 2022. 2

[36] Hanbo Zhang, Xuguang Lan, Xinwen Zhou, Zhiqiang Tian, Yang Zhang, and Nanning Zheng. Visual manipulation relationship network for autonomous robotics. In *Humanoids*, 2018. 2

[37] Hanbo Zhang, Deyu Yang, Han Wang, Binglei Zhao, Xuguang Lan, Jishiyu Ding, and Nanning Zheng. REGRAD: A Large-Scale Relational Grasp Dataset for Safe and Object-Specific Robotic Grasping in Clutter. *IEEE RAL*, 2022. 2

[38] Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, and Shanghang Zhang. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. In *NeurIPS*, 2025. 2, 5