

Free-form language-based robotic reasoning and grasping

Runyu Jiao^{1,2,†}, Alice Fasoli^{1,†}, Francesco Giuliani¹, Matteo Bortolon^{1,2,3}, Sergio Povoli¹, Guofeng Mei¹, Yiming Wang¹, Fabio Poiesi¹

¹Fondazione Bruno Kessler, ²University of Trento, ³Istituto Italiano di Tecnologia

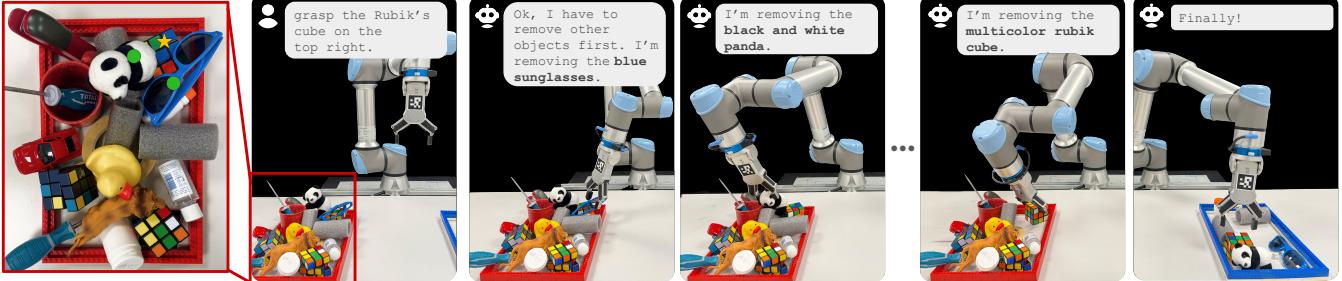


Fig. 1: To enable a human to command a robot using free-form language instructions, our method leverages the world knowledge of Vision-Language Models to interpret instructions and reason about object spatial relationships. This is important when the target object (★) is not directly graspable, requiring the robot to first identify and remove obstructing objects (●). By optimizing the sequence of actions, our approach ensures efficient task completion.

Abstract—Performing robotic grasping from a cluttered bin based on human instructions is a challenging task, as it requires understanding both the nuances of free-form language and the spatial relationships between objects. Vision-Language Models (VLMs) trained on web-scale data, such as GPT-4o, have demonstrated remarkable reasoning capabilities across both text and images. *But can they truly be used for this task in a zero-shot setting? And what are their limitations?* In this paper, we explore these research questions via the free-form language-based robotic grasping task and propose a novel method, FreeGrasp, leveraging the pre-trained VLMs’ world knowledge to reason about human instructions and object spatial arrangements. Our method detects all objects as keypoints and uses these keypoints to annotate marks on images, aiming to facilitate GPT-4o’s zero-shot spatial reasoning. This allows our method to determine whether a requested object is directly graspable or if other objects must be grasped and removed first. Since no existing dataset is specifically designed for this task, we introduce a synthetic dataset FreeGraspData by extending the MetaGraspNetV2 dataset with human-annotated instructions and ground-truth grasping sequences. We conduct extensive analyses with both FreeGraspData and real-world validation with a gripper-equipped robotic arm, demonstrating state-of-the-art performance in grasp reasoning and execution. Project website: <https://tev-fbk.github.io/FreeGrasp/>.

I. INTRODUCTION

Vision-Language Models (VLMs) encode vast semantic knowledge about the world, which is essential for interpreting nuanced human free-form language instructions [1]. This valuable source of information can enable robots to operate effectively in diverse, unseen, and unstructured environments [2], [3]. Interpreting human instructions is only one part of the challenge; grounding the retrieved information in the

real world is equally critical. This process requires spatial reasoning [4], [5], *i.e.* understanding and acting upon spatial relationships between objects [6], [7]. By achieving these capabilities together, we can tackle novel problems and enable embodied agents to perform object manipulation and rearrangement in less-controlled environments [8].

Recently, several studies have explored the use of pre-trained large-scale models for robotic control [9]–[11]. Building on this direction, we investigate how to leverage VLMs’ knowledge to enable a robot to interpret human commands. However, unlike previous works, our focus is specifically on evaluating the robustness of VLMs to free-form language instructions combined with spatial reasoning capabilities for robotic grasping. Our setup, illustrated in Fig. 1, consists of a robotic arm with a two-finger gripper, an exocentric RGB-D camera, a red bin containing various objects in clutter, and an empty blue bin where the grasped objects are put. The objects in the red bin are arranged haphazardly, and multiple instances of the same object may be present, e.g., Rubik’s cubes in this case. Our robotic system is commanded by users through instructions, *e.g.* grasp the Rubik’s cube on the top right, with the goal of identifying which objects must be removed first in order to grasp the user-specified target. The system should minimize the number of grasping actions to complete the task efficiently.

To this end, we introduce a free-form language-based robotic grasping approach, *FreeGrasp*, which leverages VLMs’ world knowledge for reasoning about human instructions and object spatial arrangement. Specifically, all visible objects within a container (or bin) are detected as keypoints, and the input image is augmented using mark-based visual prompting [10]. The augmented image is then analyzed by GPT-4o [1], which interprets the free-form language instruction and reasons spatially about which object to interact

[†]Equal contribution. This project was supported by Fondazione VRT under the project Make Grasping Easy, PNRR ICSC National Research Centre for HPC, Big Data and Quantum Computing (CN00000013), and FAIR - Future AI Research (PE00000013), funded by NextGeneration EU.

with. GPT-4o's output is used to segment the input image and identify the correct object instance, particularly when multiple instances of the same object class are present. Depth information and the segmented object instance are then used to estimate the grasp pose [12]. Note that, our study aims to *investigate* the reasoning capabilities of pre-trained models, hence, we do not train any model of our pipeline on task-specific data. We validate our approach through both synthetic data and real-world robotic experiments. Since no specific dataset exists to evaluate this task setup, we extended an existing bin-picking dataset, MetaGraspNetV2 [13], creating a new evaluation dataset named FreeGraspData. Specifically, we involved ten human participants to annotate free-form language instructions. FreeGraspData consists of scenarios with different task difficulties (Easy, Medium, Hard) based on obstruction conditions, each contains cases with and without object ambiguity, *i.e.* the presence of multiple instances of the same class. In total, FreeGraspData features 300 scenarios, each is associated with three annotated instructions from different annotators. We compare the performance of FreeGrasp against the recent approach ThinkGrasp [11], which also features grasp reasoning with GPT-4o. Experiments show that FreeGrasp can robustly interpret human instructions and infer the actions to grasp a requested object more effectively than ThinkGrasp. In summary, our contributions are:

- We introduce a novel robotic grasping setup where reasoning is key to interpreting human free-form language instructions and understanding spatial arrangement.
- We propose a novel method that leverages VLMs' world knowledge to address this task setup without additional training on task-specific data.
- We construct a new evaluation dataset that extends MetaGraspNetV2 [13] to validate the effectiveness of FreeGrasp in this novel setup.
- Our robotic experiments confirm the advantages of FreeGrasp in reasoning and grasping under real-world conditions with clutter and object ambiguity.

II. RELATED WORKS

VLM-driven robotic grasping. Multimodal learning has significantly enhanced robots' ability to interpret language instructions and perform grasp reasoning tasks. Research has focused on task-oriented grasping and spatial affordance prediction through VLMs. For example, RoboPoint [14] integrates VLMs to predict spatial affordances, improving object interaction understanding. SpatialVLA [15] enhances Vision-Language-Action (VLA) models with spatial representations, improving task planning and execution. GLOVER [16] advances open-vocabulary reasoning for task-oriented grasping, enabling robots to perform diverse tasks without task-specific training. To improve spatial reasoning, SpatialCoT [17] combines coordinate alignment with chain-of-thought reasoning for improved embodied task planning. SpatialPIN [18] refines this by integrating 3D priors and advanced prompting techniques, enhancing object recognition, scene understanding, and navigation. ThinkGrasp [11], somewhat related to our work, combines visual recognition and language-based reasoning

to improve part grasping in cluttered environments. Unlike ThinkGrasp, our method, FreeGrasp, can handle free-form language instructions perform spatial reasoning more effectively. Moreover, FreeGrasp can interpret complex language descriptions to resolve object ambiguities, ensuring accurate identification of target objects even in cluttered or ambiguous environments.

Learning-based and traditional robotic grasping. Traditional methods, which often rely on heuristic or geometric approaches, may struggle in complex scenarios. To address these limitations, researchers have increasingly emphasized relationship reasoning to enhance robotic systems' robustness and adaptability, enabling them to better understand object interactions and perform tasks in cluttered environments. A key milestone is VMRN [19], which introduced a framework and dataset for predicting object manipulation relationships, highlighting the role of relationship reasoning in robotics. HSRN [20] advanced this by proposing a hierarchical stacking relationship prediction method, optimizing manipulation in cluttered, occluded settings. Large-scale datasets like REGRAD [21] and MetaGraspNetV2 [13] have been pivotal. REGRAD [21] focuses on safe, object-specific grasping in clutter, while MetaGraspNetV2 [13] offers a comprehensive dataset for bin picking, integrating relationship reasoning and dexterous grasping. These datasets have enabled more sophisticated grasp planning strategies, improving performance in complex scenarios. Recent works [22], [23] integrate visual relationship reasoning and semantic scene understanding into grasp planning. D3GD [22] proposes a scalable method for real-time object relationship reasoning. RelationGrasp [24] introduces prompt learning for grasp detection and relationship prediction, showing promise in unstructured environments. GOAL [25] addresses occlusion by combining occlusion-aware perception with relationship reasoning, significantly improving grasp success in highly cluttered scenes.

III. OUR METHOD

A. Overview

We address the robotic reasoning and grasping task and propose an effective and modular pipeline by leveraging the world knowledge of pre-trained VLMs without additional training. Our setup consists of a bin containing objects organized haphazardly, an exocentric RGB-D camera with top-down view, a robotic manipulator, and a user who provides a free-form language instruction to request the robot to grasp a target object, *e.g.* "grasp the red toy car at the bottom." Fig. 2 shows the diagram of FreeGrasp.

An episode starts upon receiving the user's instruction. At each step, the pipeline starts with the proposed *vision-language grasp reasoning* module. With the RGB observation, the module first localizes all the objects within the container, forming a holistic understanding of the scene [26]. To facilitate VLMs in visual spatial reasoning, we augment the visual prompt by annotating identity (ID) marks for each localized object on the input image. Then, we feed the mark-based visual prompt to GPT-4o [1] to reason about object

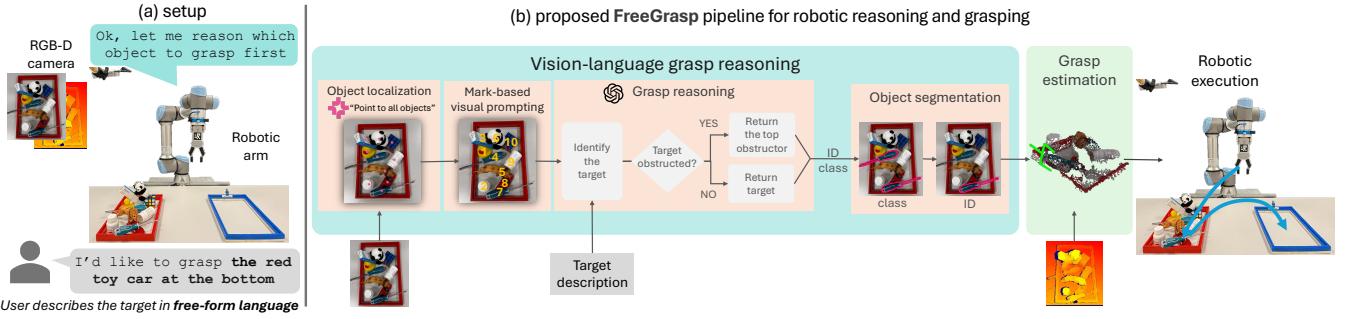


Fig. 2: FreeGrasp pipeline. (a) The setup considered for the robotic reasoning and grasping task and (b) the proposed pipeline that leverages pre-trained VLMs in a zero-shot manner without additional training.

spatial relationships. GPT-4o determines whether the target object is free of obstruction, thus directly graspable, or is obstructed by other objects that must be removed first. GPT-4o’s output consists of the ID and class name of the next object to be grasped, along with information indicating whether it is the target object. With the class name and ID, the object segmentation model further segments the output object [27]. The *grasp estimation* module determines the most suitable grasp pose for picking the segmented object [12]. Lastly, the robotic arm plans and executes its trajectory to grasp the object and place it in a predefined location, *e.g.*, another bin. After each step, the positions of objects may change. The pipeline thus repeats with a new RGB-D acquisition. The episode for the robotic reasoning and grasping task terminates until the user-specified target object is being grasped or certain criteria is met.

B. Vision-language grasp reasoning

Given the RGB image capturing the scene and the user’s free-form language instruction, this module leverages state-of-the-art VLMs to visually ground objects and reason about their relationship to determine the object to grasp, without additional model training.

Object localization. There exist several VLM-based open-vocabulary object localization methods like [28]. We mainly explored two options, and empirically selected the best-performing one for our task setup. One option is to first prompt a VLM (*e.g.* GPT-4o [1]) for obtaining a list of object names in the bin and then use open-vocabulary segmentation models (*e.g.* LangSAM [27]) for object localization, similar to [29]. Alternatively, we can directly prompt a VLM with visual grounding capability (*e.g.* Molmo [26]) to point at all objects in the bin. The output is a list of 2D coordinates, which are typically around the centers of the corresponding objects (Fig. 2). We evaluated the object localization performance with MetaGraspNetv2 [13] and report the Average Precision (AP), Average Recall (AR), and F1 score in Fig. 3. Directly leveraging Molmo leads to the better localization performance than the alternative with GPT-4o and LangSAM. The latter is prone to produce duplicate or fragmented masks among nearby objects, which is non-trivial to address with simple post-processing, *e.g.* removing excessively small or large masks. Therefore, we selected Molmo for pointing.



Fig. 3: Object localization performance with different VLM-based method on MetaGraspNetv2.

Mark-based visual prompting. VLMs can reason more effectively when inputs are presented in a multiple-choice format [10], [30]. Hence, we assign a unique number to each object, and augment the input RGB image by annotating the numbered markers at the 2D object coordinates provided by the previous step (Fig. 2). Then, we input the mark-based visual prompt into the reasoning VLM for grasp reasoning.

Grasp reasoning. The VLM must determine which object to grasp in each episode, which is the most essential yet challenging aspect of our proposed pipeline. Effective grasp reasoning requires the VLM to understand both: i) free-form language instructions and ii) spatial relationships, particularly in terms of obstructions. We leverage GPT-4o, as it demonstrated the strongest reasoning capabilities among the VLMs we tested.

Free-form language instructions from different users can vary dramatically when referring to the same object, due to different life backgrounds of users. For example, in our experiments, the target object in Fig. 5 (bottom left) is referred by different users as juice box, juice, or refill pouch. Such ambiguous (even wrong) instructions pose significant challenges to the VLM to reason on the correct target object. To resolve ambiguities, the capability in extracting relevant contextual information is critical, including spatial references (*e.g.* top left corner) or part names (*e.g.* pouch). In the case of user instruction the juice box on the top left corner, GPT-4o responds as I don’t see a juice box in the top left corner. If you mean the object labeled “3”, it is free of obstacles.

For obstruction reasoning, the VLM needs to understand spatial relationships among the target object and its nearby objects. This is particularly challenging to infer from a single image when objects are layered on top of each other. Despite GPT-4o being the best VLM in terms of reasoning, its performance in this regard remains fairly poor. To improve

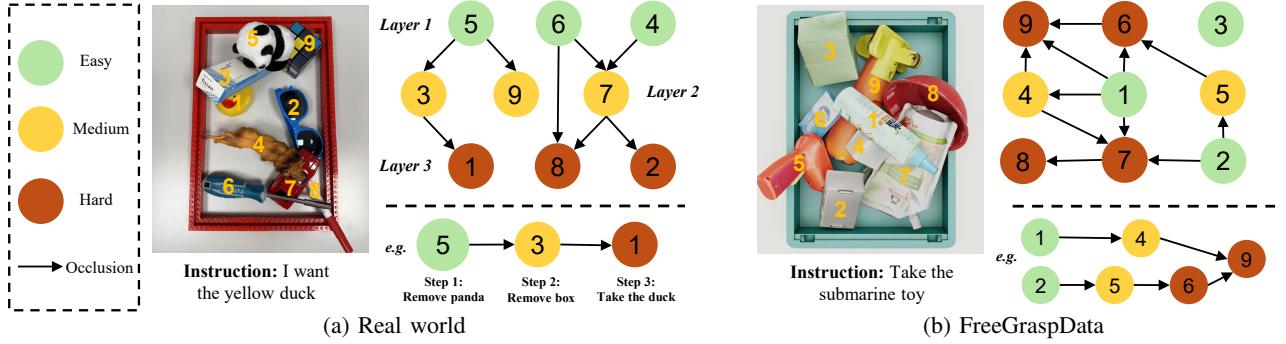


Fig. 4: Occlusion graphs in representative scenarios. (a) Real-world scene and its corresponding occlusion graph. (b) Synthetic scene from FreeGraspData and its occlusion graph. Objects are colored by difficulty, and directed edges indicate occlusion dependencies. Example task instructions and step-by-step action plans are shown beneath each graph.

occlusion reasoning, we initially explored recent advances in Chain-of-Thought (CoT) reasoning [31] by first generating a detailed description of all objects and then summarizing it into a scene graph [32] that captures object adjacency. Surprisingly, this approach did not enhance performance. We observed that the model frequently hallucinated spatial relationships and produced inconsistent solutions for the same prompt, even when the VLM’s temperature was set to zero. Instead, we found that an effective strategy is to *contextualize the task* within the prompt by explicitly defining key aspects: the setup (*e.g.* a robotic arm with a parallel gripper for bin picking), the objective (*e.g.* grasp the target without obstruction), the reasoning logic (*e.g.* determine actions based on obstructions), and the possible actions (*e.g.* return the target or identify the top obstructor). With this structured prompt, the VLM reliably returns the ID and class name of the object to grasp.

Object segmentation. We segment the object instance identified in the image. For instance segmentation, simply using the 2D coordinate given by the object ID as visual prompt to class-agnostic segmentation models, *e.g.* SAM [33], results in inaccurate instance masks with little semantic control. We thus first perform semantic segmentation using LangSAM [27] (Fig. 2) with the class name. As there might be multiple instances with the same class, we then use the object ID to filter the instance mask of interest.

C. Grasp estimation

Lastly, we estimate the object’s grasp pose with respect to the robotic arm. Once the object instance is segmented, we use GraspNet [12] to regress the grasp pose. GraspNet operates on a colored point cloud, therefore we first lift the RGB and depth image into 3D using the camera’s intrinsic parameters. Then, through the instance mask, we crop the point cloud to retain only the region corresponding to the object of interest and select the grasp pose estimated by GraspNet with the highest confidence.

IV. FREE-FORM LANGUAGE GRASPING DATASET

We introduce the free-from language grasping dataset (FreeGraspData), a novel dataset built upon MetaGraspNetv2 [13] to evaluate the robotic grasping task with free-form language instructions. MetaGraspNetv2 is a large-scale simulated dataset featuring challenging aspects of

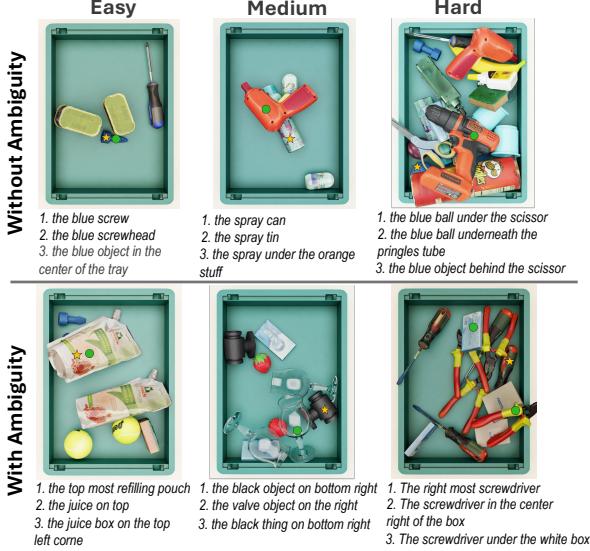


Fig. 5: Examples of FreeGraspData at different task difficulties with three user-provided instructions. \star indicates the target object, and \bullet indicates the ground-truth objects to pick.

robot vision in the bin-picking setting, including multi-view RGB-D images and metadata, *e.g.* object categories, amodal segmentation masks, and occlusion graphs indicating occlusion relationships between objects from each viewpoint. To build FreeGraspData, we selected scenes containing at least four objects to ensure sufficient scene clutter. FreeGraspData extends MetaGraspNetV2 in three aspects: i) we derive the ground-truth grasp sequence until reaching the target object from the occlusion graphs, ii) we categorize the task difficulty based on the obstruction level and instance ambiguity, and iii) we provide free-form language instructions collected from human annotators. Examples of occlusion graphs for real-world and FreeGraspData scenes are illustrated in Fig. 4.

A. Ground-truth grasp sequence

We obtain the ground-truth grasp sequence based on the object occlusion graphs provided in MetaGraspNetV2. As the visual occlusion does not necessarily indicate obstruction, we thus first prune the edges in the provided occlusion graph that are less likely to form obstruction. Following the heuristic that less occlusion indicates less chance of obstruction, we remove

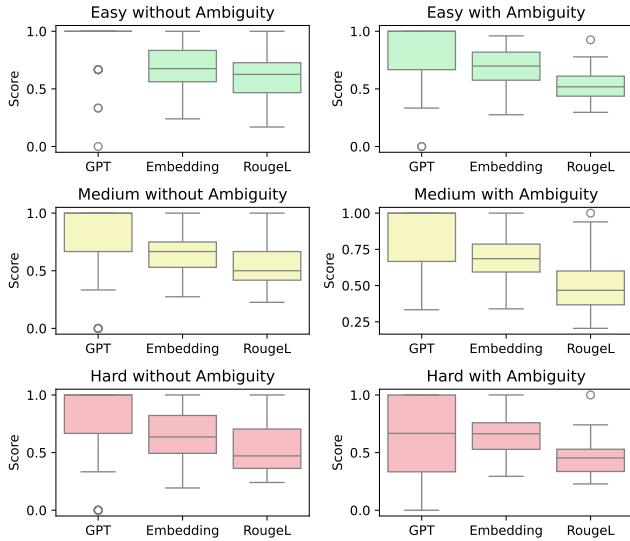


Fig. 6: Similarity distribution among the three user-defined instructions used in the FreeGraspData scenarios.

the edges where the percentage of the occlusion area of the occluded object is below 1%. From the node representing the target object, we can then traverse the pruned graph to locate the leaf node, which is the ground-truth object to grasp first. The sequence from the leaf node to the target node forms the correct sequence for the robotic grasping task.

B. Grasp difficulty categorization

We leverage the pruned occlusion graph to classify the grasping difficulty of target object into three levels: “*Easy*” refers to unobstructed target objects, *i.e.* leaf nodes in the pruned occlusion graph; “*Medium*” refers to objects obstructed by at least one object, *i.e.* the maximum hop distance to the leaf nodes is 1; and “*Hard*” refers to objects obstructed by a chain of other objects, *i.e.* the maximum hop distance to the leaf nodes is more than 1. We label objects as “*Ambiguous*” if multiple instances of the same class exist in the scene. Based on these criteria, we obtain six robotic grasping difficulty categories, also shown in Fig. 5: “*Easy without Ambiguity*”, “*Medium without Ambiguity*”, “*Hard without Ambiguity*”, “*Easy with Ambiguity*”, “*Medium with Ambiguity*”, and “*Hard with Ambiguity*”.

C. Free-form language user instructions

For each of the six difficulty categories, we randomly select 50 objects, resulting in 300 robotic grasping scenarios. For each scenario, we provide multiple users with a top-down image of the bin and a visual indicator highlighting the target object. We randomly select three user instructions for each scenario, yielding a total of 900 evaluation scenarios. This results in diverse language instructions, as shown in Fig. 5.

Fig. 6 illustrates the similarity distribution among the three user-defined instructions in FreeGraspData, based on GPT-4o’s interpretability, semantic similarity, and sentence structure similarity. To assess GPT-4o’s interpretability, we introduce a novel metric, the GPT score, which measures GPT-4o’s coherence in responses. For each target, we provide GPT-

4o with an image containing overlaid object IDs and ask it to identify the object specified by each of the three instructions. The GPT score quantifies the fraction of correctly identified instructions, ranging from 0 (no correct identifications) to 1 (all three correct). We evaluate semantic similarity using the embedding score, defined as the average SBERT [34] similarity across all pairs of user-defined instructions. We assess structural similarity using the Rouge-L score, computed as the average Rouge-L [35] score across all instruction pairs. Results indicate that instructions referring to the same target vary significantly in sentence structure (low Rouge-L score), reflecting differences in word choice and composition, while showing moderate variation in semantics (medium embedding score). Interestingly, despite these variations, the consistently high GPT scores across all task difficulty levels suggest that GPT-4o is robust in identifying the correct target in the image, regardless of differences in instruction phrasing.

V. EXPERIMENTS

We conduct experiments using both our synthetic evaluation dataset, FreeGraspData, and a gripper-equipped robotic arm in the real world. We compare FreeGrasp against the state-of-the-art method, ThinkGrasp [11], which also uses GPT-4o for reasoning. To comprehensively evaluate robotic grasping, we introduce novel metrics that assess performance based on both intermediate steps and the final robotic grasp. Using FreeGraspData, we also perform extensive ablation studies on the key components of FreeGrasp to validate our design choices. Lastly, through real-world experiments, we demonstrate the effectiveness of FreeGrasp in handling practical challenges such as imperfect depth measurements and robotic grasp execution.

A. Experiments on FreeGraspData

Experimental setup. We compare FreeGrasp and ThinkGrasp on FreeGraspData, which includes 900 evaluation scenarios across the six task difficulty levels outlined in Sec. IV. Both methods receive a top-down RGB image as input, along with user-defined task instructions. Since each object in the evaluation set is associated with three different user instructions, we report the final performance metrics as the mean and standard deviation across these three instructions.

Performance metrics. Since FreeGraspData is a static dataset, we mainly evaluate grasp reasoning and object segmentation. We do not assess grasp estimation with FreeGraspData, as this aspect is best evaluated through robotic execution (Sec. V-B). For grasp reasoning, we report the Reasoning Success Rate (RSR), *i.e.* the proportion of successful episodes in which the predicted object ID matches one of the ground-truth IDs. For object segmentation, we report the Segmentation Success Rate (SSR), *i.e.* the proportion of successful episodes in which the output mask achieves an Intersection over Union of at least 0.5 with the ground-truth mask. Since ThinkGrasp does not produce object IDs, we can only report its SSR. Moreover, we evaluate a variant of FreeGrasp to assess the impact of object localization on reasoning, as measured by RSR. This variant uses ground-truth (GT) localization instead of Molmbo.

TABLE I: Experiments on FreeGraspData. Higher metric values (SSR and RSR) indicate better performance. Best performance under each setting is in *italic*.

Method	Reas.	Segm.	Metric	Easy		Medium		Hard	
				w/o Amb.	w Amb.	w/o Amb.	w Amb.	w/o Amb.	w Amb.
ThinkGrasp [11]	✓	✓	SSR	0.63±0.02	0.46±0.02	0.13±0.03	0.16±0.02	0.05±0.02	0.15±0.02
FreeGrasp	✓	✓	SSR	0.64±0.03	0.64±0.04	0.40±0.04	0.35±0.02	0.13±0.01	0.13±0.02
FreeGrasp	✓(GT)		RSR	0.83±0.02	0.77±0.02	0.46±0.03	0.31±0.06	0.21±0.01	0.16±0.04
FreeGrasp	✓(Molmo)		RSR	0.83±0.06	0.85±0.07	0.46±0.04	0.33±0.04	0.22±0.04	0.15±0.04

Discussion. Tab. I presents the results for SSR and RSR using FreeGraspData. FreeGrasp significantly outperforms ThinkGrasp across almost all difficulty levels, except in “*Easy without Ambiguity*” scenario where both methods perform similarly. This shows that FreeGrasp effectively handles object ambiguity, thanks to its object localization design and mark-based visual prompting. As scene clutter increases (Medium & Hard scenarios), we observe that ThinkGrasp performs better in ambiguous cases than in non-ambiguous ones. A qualitative analysis of the results suggests that this is due to ThinkGrasp’s tendency to segment all instances of the user-specified object. In fact, the larger the number of ambiguities (*i.e.* multiple instances of the target class), the higher the likelihood that the ground-truth obstructor belongs to the same class as the user-specified object. Lastly, when comparing Molmo-based object localization with ground-truth (GT) detections, we find that Molmo achieves a higher RSR than GT localization. This is because Molmo tends to miss highly occluded objects, increasing the likelihood that the reasoning module will select objects on the surface.

B. Experiments in the real world

Experimental setup. We evaluate FreeGrasp and ThinkGrasp [11] in a real-world scenario with a UR5e robotic arm equipped with an OnRobot RG2 parallel gripper. The scene setup is similar as in prior work ThinkGrasp [11] for fair comparison. The scene is captured by a RealSense D415 RGB-D camera with a top-down view featuring the cluttered red bin, at a resolution of 1280×720. We implement the motion planning and control via the MoveIt motion planning framework with ROS. Adhering to the definition of six task difficulties in Sec. IV, we compose ten evaluation scenarios per difficulty level (Fig. 7), each featuring a unique arrangement of objects

Evaluation protocol. At each evaluation episode, the robotic arm is tasked to grasp and place the target object from its cluttered bin (red) to the empty bin (blue), upon receiving the user instruction describing the target in free-form language.

The failure of the pipeline at each time step may occur at different modules, *i.e.* vision-language grasp reasoning, grasp estimation, or robotic execution. To comprehensively evaluate the pipeline’s performance, we identify the failures at: i) Segmentation (S): when the predicted segmentation mask does not correspond to one of the GT unobstructed objects; ii) Pose (P): when the predicted pose does not correspond to one of the GT unobstructed objects; and iii) Motion (M): when robotic arm fails to retain one of the GT unobstructed objects in the gripper throughout the motion. Interestingly,

we find that a segmentation failure does not necessarily lead to a pose failure. As shown in Fig. 8, as long as the wrong segmentation mask includes a GT unobstructed object, the resulted pose can remain correct.

To study the impact of localized failures, we further devise three operational settings by defining the stopping criteria based on localized failures: i) *Setting (S, P, M)* is the most stringent setting, where an evaluation episode will be terminated if any of the localized failures occur throughout the robotic reasoning and grasping task; ii) *Setting (P, M)* relaxes the segmentation failure, where an evaluation episode will be terminated only if a pose failure or motion failure occurs; and iii) *Setting (P)* relaxes the motion failure, where an evaluation episode will only be terminated if a pose failure occurs.

Performance metrics. We use three metrics, inspired by navigation tasks [36], to measure the effectiveness and efficiency of ThinkGrasp and FreeGrasp. We define the *success* of an episode when the correct target object is grasped and placed to its pre-defined destination. The *success rate (SR)* reports the ratio of successful episodes among all evaluation episodes, *i.e.* $SR = \frac{N_s}{N}$, where N is the number of total evaluation episodes and N_s is the number of successful episodes. The *path efficiency (SE)* reports the normalized inverse step counts, *i.e.* $PE = \frac{1}{N_s} \sum_{i=1}^{N_s} \frac{l_i}{p_i}$, where l_i is the number of minimum steps to solve the episode as defined by an oracle (*i.e.* the operator), and p_i is the actual number of steps the robot actually had to take to solve the episode. Finally, the *Success Weighted (normalized inverse) path length (SPL)*, calculated as $SPL = \frac{1}{N} \sum_{i=1}^N S_i \frac{l_i}{p_i}$, where S_i is 1 if episode i is successful, 0 otherwise.

Discussion. Tab. II reports the results of FreeGrasp and ThinkGrasp, in real-world experiments across three operational settings. FreeGrasp outperforms ThinkGrasp in nearly all evaluated settings and task difficulties, except in the least difficult scenario, *i.e.* *Easy without Ambiguity*. ThinkGrasp struggles to identify and segment the correct object to grasp as scene clutter and object ambiguity increase. Under the most stringent *Setting (S, P, M)*, ThinkGrasp fails all evaluation episodes in the *Medium* and *Hard* scenarios. In contrast, FreeGrasp performs better in these challenging scenarios, thanks to the *vision-language grasp reasoning* module, which helps identify the correct object to pick.

In *Setting (P, M)*, we sometimes observe slightly better performance than in *Setting (S, P, M)*. This occurs when the segmentation mask includes multiple objects (which counts as a segmentation failure), and one of them is a ground-truth obstructed object. In such cases, the estimated grasp pose is

TABLE II: Results of real-world experiments. Higher metric values (SR, PE and SPL) indicate better performance. Best performance under each setting is in *italic*.

Method	Easy												Medium												Hard											
	Stop criteria			w/o Amb.			w Amb.			w/o Amb.			w Amb.			w/o Amb.			w Amb.			w/o Amb.			w Amb.											
	S	P	M	SR	PE	SPL	SR	PE	SPL	SR	PE	SPL	SR	PE	SPL	SR	PE	SPL	SR	PE	SPL	SR	PE	SPL	SR	PE	SPL									
ThinkGrasp [11]	✓	✓	✓	0.60	1.0	0.60	0.40	0.71	0.28	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0							
FreeGrasp	✓	✓	✓	0.50	1.0	0.50	0.80	0.85	0.68	0.20	1.0	0.20	0.20	1.0	0.20	0.0	0.0	0.0	0.0	0.0	0.0	0.10	1.0	0.10	0.10	1.0	0.10									
ThinkGrasp [11]	✓	✓	✓	0.70	1.0	0.70	0.40	0.71	0.28	0.10	1.0	0.10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0								
FreeGrasp	✓	✓	✓	0.70	1.0	0.70	0.80	0.85	0.68	0.20	1.0	0.20	0.20	1.0	0.20	0.10	1.0	0.10	0.10	0.10	0.10	0.10	1.0	0.10	0.10	1.0	0.10									
ThinkGrasp [11]	✓			1.0	0.95	0.95	0.70	0.74	0.52	0.40	0.92	0.37	0.10	0.67	0.07	0.10	1.0	0.10	0.10	0.10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0								
FreeGrasp	✓			0.90	0.94	0.85	0.80	0.85	0.68	0.60	0.92	0.55	0.40	0.90	0.36	0.20	1.0	0.20	0.10	0.10	0.10	0.0	0.0	0.0	0.0	0.0	0.0	0.0								

more likely to be correct, leading to more successful episodes. We qualitatively demonstrate such a case in Fig. 8, where the second row shows a case with incorrect segmentation but a correct estimated pose. In contrast, the first row shows a case where both segmentation and estimated pose are correct.

Lastly, as expected, when robotic execution is not considered (*i.e.* under *Setting (P)*), we observe the best performance. Note that, the lower performance in the *Medium* and *Hard* scenarios highlights the difficulty of this robotic reasoning and grasping task, which we deem far from being solved. Future work in visual spatial reasoning is necessary to better address challenges like scene clutter and ambiguity, which are critical in real-world applications.

C. Computational analysis

FreeGrasp ran on a workstation equipped with a 24 GB NVIDIA RTX 4500 GPU. The computational analysis of FreeGrasp was conducted using 60 episodes, covering all task difficulties in our real-world experiments. The mean total execution time is 15.39 seconds, with the breakdown of time for each main component as follows: *object localization* with Molmo (9.12s), *grasp reasoning* with GPT-4o (5.46s), *object segmentation* with LangSAM (0.71s), and *grasp estimation* with GraspNet (0.10s). Since the camera is externally mounted, VLM-based reasoning and pose estimation can be performed in parallel with robotic manipulation after the first step of the episode, while the robot is in motion.

VI. CONCLUSIONS

We introduced FreeGrasp, a novel approach that leverages pre-trained VLMs for robotic grasping by interpreting free-form instructions and reasoning about spatial relationships. Our investigation showed that while VLMs, *e.g.* GPT-4o, are known for strong general reasoning capabilities, they struggle with visual spatial reasoning, highlighting an important gap for both visual grounding and spatial awareness. FreeGrasp was designed to mitigate such gap with mark-based visual prompting and contextualized reasoning, outperforming the state-of-the-art, ThinkGrasp, in both synthetic and real-world robotic validations. Overall, FreeGrasp demonstrates the potential of combining VLMs with modular reasoning to tackle robotic grasping challenges in real-world scenarios. *Limitations and future works.* A key limitation we observed of FreeGrasp is GPT-4o’s limited visual-spatial capability, particularly in understanding object occlusion. While we tested specialized spatial VLMs like SpaceLLaVA [6], they



Fig. 7: Samples from real-world experiments for different task difficulties. ★ indicates the user-described target object, and ● are the GT objects to pick.

underperformed compared to GPT-4o. Another limitation is the lack of a mechanism to adapt the initial human instructions as scenes change with object removal. Future work could explore memory mechanisms to track scene changes or adaptive instruction updates using VLMs to dynamically adjust references as objects are removed.

REFERENCES

- [1] OpenAI, “GPT-4 Technical Report,” *arXiv:2303.08774*, 2024.
- [2] M. Ahn *et al.*, “Do as i can, not as i say: Grounding language in robotic affordances,” in *CoRL*, 2022.
- [3] B. Chen, F. Xia, B. Ichter, K. Rao, K. Gopalakrishnan, M. S. Ryoo, A. Stone, and D. Kappler, “Open-vocabulary Queryable Scene Representations for Real World Planning,” *RAL*, 2023.
- [4] K. Xu, S. Zhao, Z. Zhou, Z. Li, H. Pi, Y. Wang, and R. Xiong, “A Joint Modeling of Vision-Language-Action for Target-oriented Grasping in Clutter,” in *ICRA*, 2023.
- [5] C. Pek, G. F. Schuppe, F. Esposito, J. Tumova, and D. Kräigic, “SpaTiaL: monitoring and planning of robotic tasks using spatio-temporal logic specifications,” *Autonomous Robots*, 2023.
- [6] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Driess, P. Florence, D. Sadigh, L. Guibas, and F. Xia, “SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities,” in *CVPR*, 2024.
- [7] P. Rabino and T. Tommasi, “A Modern Take on Visual Relationship Reasoning for Grasp Planning,” *RAL*, 2025.
- [8] W. Du, S. Li, G. Xiang, F. Gao, and F. Shuang, “Multi-Teachers Distillation Strategy for Target-Oriented Collision-Free Grasping in Clutter,” *RAL*, 2024.
- [9] W. Huang, W. Huang, C. Wang, Y. Li, R. Zhang, and L. Fei-Fei, “ReKep: Spatio-Temporal Reasoning of Relational Keypoint Constraints for Robotic Manipulation,” in *CoRL*, 2024.

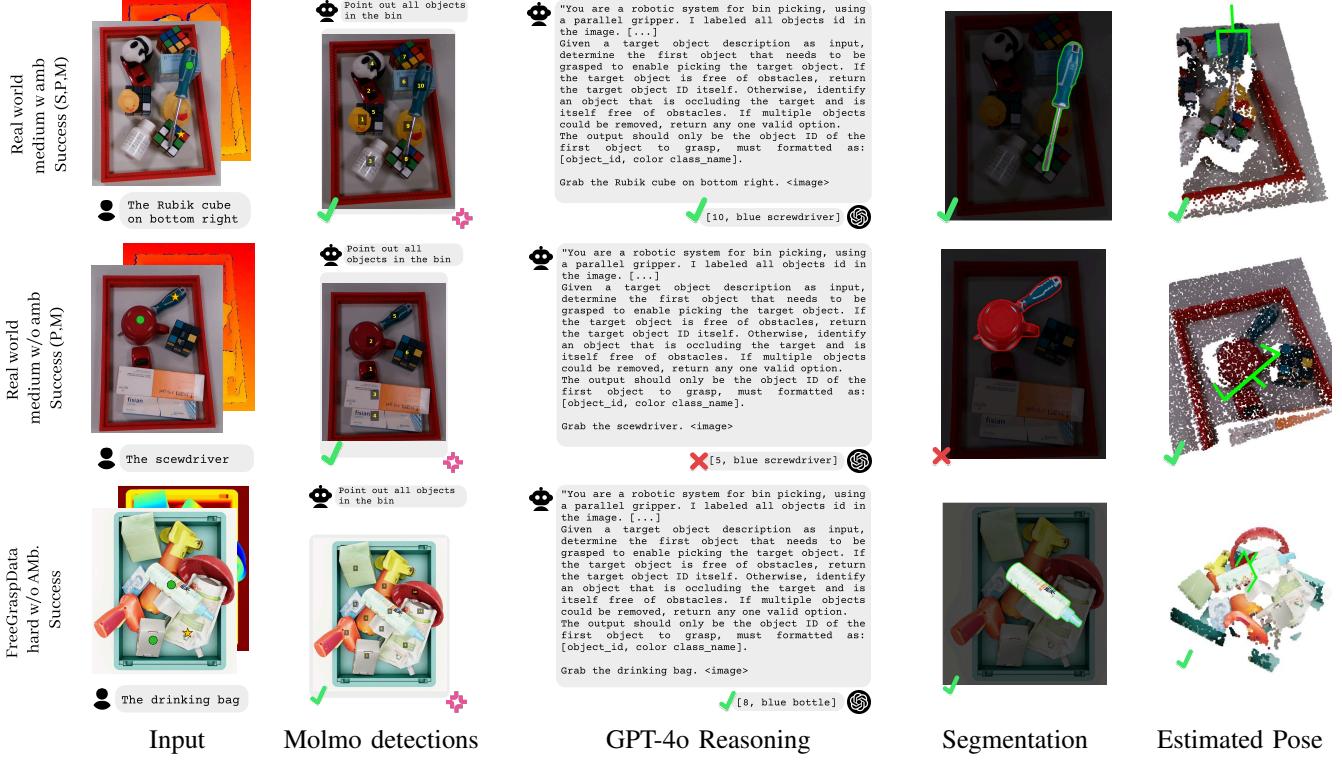


Fig. 8: Examples of three successful scenarios with FreeGrasp: At the top, we show a real-world case where segmentation, pose, and motion were all correctly identified. In the middle, we present another real-world scenario where segmentation was incorrect, but GraspNet was still able to identify the correct pose to remove the obstacle. At the bottom, we display a successful example from a FreeGraspData scenario, featuring a highly cluttered environment.

- [10] F. Liu, K. Fang, P. Abbeel, and S. Levine, “MOKA: Open-World Robotic Manipulation through Mark-Based Visual Prompting,” in *RSS*, 2024.
- [11] Y. Qian, X. Zhu, O. Biza, S. Jiang, L. Zhao, H. Huang, Y. Qi, and R. Platt, “ThinkGrasp: A Vision-Language System for Strategic Part Grasping in Clutter,” in *CoRL*, 2024.
- [12] H.-S. Fang, C. Wang, M. Gou, and C. Lu, “Graspnet-1billion: A large-scale benchmark for general object grasping,” in *CVPR*, 2020.
- [13] M. Gilles, Y. Chen, E. Z. Zeng, Y. Wu, K. Furmans, and A. Wong, “MetaGraspNetV2: All-in-One Dataset Enabling Fast and Reliable Robotic Bin Picking via Object Relationship Reasoning and Dexterous Grasping,” in *TASE*, 2024.
- [14] W. Yuan, J. Duan, V. Blukis, W. Pumacay, R. Krishna, A. Murali, A. Mousavian, and D. Fox, “Robopoint: A vision-language model for spatial affordance prediction in robotics,” in *CoRL*, 2024.
- [15] D. Qu, H. Song, Q. Chen, Y. Yao, X. Ye, Y. Ding, Z. Wang, J. Gu, B. Zhao, D. Wang, *et al.*, “Spatialvla: Exploring spatial representations for visual-language-action model,” *arXiv*, 2025.
- [16] T. Ma, Z. Wang, J. Zhou, M. Wang, and J. Liang, “GLOVER: Generalizable open-vocabulary affordance reasoning for task-oriented grasping,” *arXiv:2411.12286*, 2024.
- [17] Y. Liu *et al.*, “SpatialCoT: Advancing Spatial Reasoning through Coordinate Alignment and Chain-of-Thought for Embodied Task Planning,” *arXiv:2501.10074*, 2025.
- [18] C. Ma, K. Lu, T.-Y. Cheng, N. Trigoni, and A. Markham, “SpatialPIN: Enhancing Spatial Reasoning Capabilities of Vision-Language Models through Prompting and Interacting 3D Priors,” in *NeurIPS*, 2024.
- [19] H. Zhang, X. Lan, X. Zhou, Z. Tian, Y. Zhang, and N. Zheng, “Visual manipulation relationship network for autonomous robotics,” in *Humanoids*, 2018.
- [20] Z. Wu, J. Tang, X. Chen, C. Ma, X. Lan, and N. Zheng, “Prioritized planning for target-oriented manipulation via hierarchical stacking relationship prediction,” in *IROS*, 2023.
- [21] H. Zhang, D. Yang, H. Wang, B. Zhao, X. Lan, J. Ding, and N. Zheng, “Regrad: A large-scale relational grasp dataset for safe and object-specific robotic grasping in clutter,” *RAL*, 2022.
- [22] P. Rabino and T. Tommasi, “A modern take on visual relationship reasoning for grasp planning,” *RAL*, 2025.
- [23] C. Tang, J. Yu, W. Chen, B. Xia, and H. Zhang, “Relationship oriented semantic scene understanding for daily manipulation tasks,” in *IROS*, 2022, pp. 9926–9933.
- [24] S. Liu, T. J. Teo, Z. Lin, and H. Zhu, “Relationgrasp: Object-oriented prompt learning for simultaneously grasp detection and manipulation relationship in open vocabulary,” in *IROS*, 2024, pp. 10 890–10 896.
- [25] L. Li, A. Cherouat, H. Snoussi, and T. Wang, “Grasping with occlusion-aware ally method in complex scenes,” *IEEE Transactions on Automation Science and Engineering*, 2024.
- [26] M. Deitke *et al.*, “Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Multimodal Models,” *arXiv:2409.17146*, 2024.
- [27] L. Medeiros, “Language Segment-Anything,” 2025. [Online]. Available: <https://github.com/lucamedeiros/lang-segment-anything>.
- [28] S. Liu *et al.*, “Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection,” in *ECCV*, 2024.
- [29] G. Mei, L. Riz, Y. Wang, and F. Poiesi, “Vocabulary-free 3d instance segmentation with vision and language assistant,” in *3DV*, 2025.
- [30] J. Yang, H. Zhang, F. Li, X. Zou, C. Li, and J. Gao, “Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V,” *arXiv:2310.11441*, 2023.
- [31] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” in *NeurIPS*, 2023.
- [32] J. Yang, X. Chen, N. Madaan, M. Iyengar, S. Qian, D. F. Fouhey, and J. Chai, “3D-GRAND: A Million-Scale Dataset for 3D-LLMs with Better Grounding and Less Hallucination,” *arXiv:2406.05132*, 2024.
- [33] A. Kirillov *et al.*, “Segment Anything,” in *ICCV*, 2023.
- [34] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” in *EMNLP*, 2019.
- [35] C.-Y. Lin and F. J. Och, “Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics,” in *ACL*, 2004.
- [36] P. Anderson *et al.*, “On evaluation of embodied navigation agents,” *arXiv:1807.06757*, 2018.