Free-form language-based robotic reasoning and grasping

Runyu Jiao[1,2,†], Alice Fasoli[1,†], Francesco Giuliari[1], Matteo Bortolon[1,2,3], Sergio Povoli[1], Guofeng Mei[1], Yiming Wang[1], Fabio Poiesi[1]
[1]Fondazione Bruno Kessler, [2]University of Trento, [3]Istituto Italiano di Tecnologia

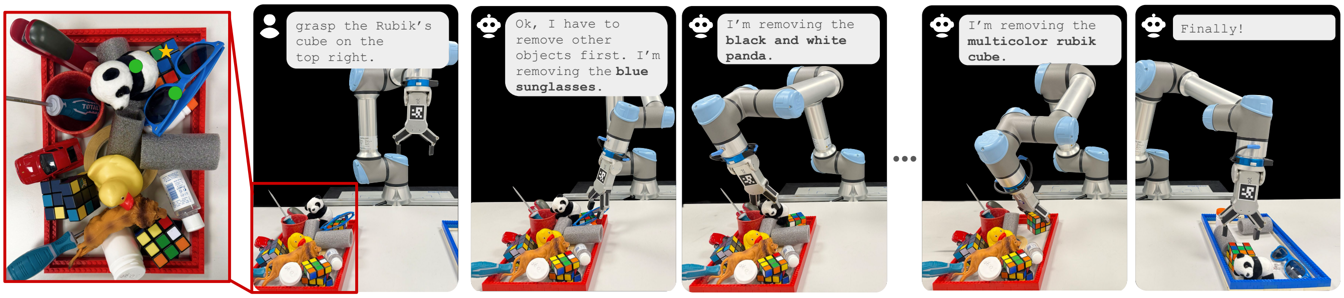

grasp the Rubik's cube on the top right.
Ok, I have to remove other objects first. I'm removing the **blue sunglasses**.
I'm removing the **black and white panda**.
I'm removing the **multicolor rubik cube**.
Finally!

Fig. 1: The task of robotic grasping from free-form language instructions in cluttered bins. The robot must interpret natural instructions, reason about spatial relationships, and handle obstructed targets (★) by first removing blocking objects (●).

***Abstract*— Performing robotic grasping from cluttered bins based on human instructions requires understanding both free-form language and spatial object relationships. We propose FreeGrasp, a novel method that leverages pre-trained vision–language models (VLMs) for zero-shot reasoning over human instructions and object arrangements. Our approach represents objects as keypoints, enabling the VLM to infer grasp sequences and decide whether to grasp directly or remove occluding objects first. We further construct a synthetic dataset with annotated instructions and grasp sequences, and validate our method in both simulated and real-world settings with a robotic arm. Experiments demonstrate state-of-the-art performance in grasp reasoning and execution, highlighting the potential of VLMs for instruction-based reasoning and grasping. Project website: https://tev-fbk.github.io/FreeGrasp/.**

I. Introduction

Vision-Language Models (VLMs) encode rich semantic knowledge essential for interpreting nuanced human free-form language instructions [1], enabling robots to operate in diverse, unseen, and unstructured environments [2]. Beyond interpreting instructions, grounding them in the physical world requires spatial reasoning, *i.e.* understanding and acting upon spatial object relationships [3].

Recent work has explored large-scale pre-trained models for robotic control [4]–[6]. We build on this by evaluating VLMs ' robustness to free-form instructions combined with spatial reasoning for robotic grasping. Our setup (Fig. 1) uses a robotic arm, RGB-D camera, and a cluttered bin with possible multiple instances of the same object, aiming to determine the minimal grasp sequence to retrieve a specified target. 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.

We present *FreeGrasp*, which leverages VLMs ' world knowledge for reasoning about instructions and spatial arrangements. Objects are detected as keypoints, marked on images via visual prompting [5], and analyzed by GPT-4o to identify the correct target, even with multiple similar instances. Depth and segmentation then yield grasp poses [7]. No task-specific training is used.

To evaluate, we extend MetaGraspNetV2 [8] with human-annotated free-form instructions, forming FreeGraspData with varied difficulty and object ambiguity. We compare against ThinkGrasp [6] and show superior reasoning and grasping in synthetic and real-world experiments. In summary, our contributions are:

- A robotic grasping setup combining free-form instruction interpretation and spatial reasoning.
- A method leveraging VLMs ' knowledge to address this task setup without task-specific training.
- A new evaluation dataset extending MetaGraspNetV2 [8] with free-form instructions.
- Real-world experiments confirming the advantages of FreeGrasp in reasoning and grasping under real-world conditions with clutter and object ambiguity.

II. Related works

VLM-driven robotic grasping. Recent work shows that multimodal learning enables robots to interpret natural language instructions and perform grasp reasoning. RoboPoint [9] integrates VLMs for spatial affordance prediction, while ThinkGrasp [6] links visual recognition with language reasoning for cluttered grasping, whereas FreeGrasp further handles free-form instructions, stronger spatial reasoning, and object ambiguities in cluttered or repetitive-object environments.

[†] 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.

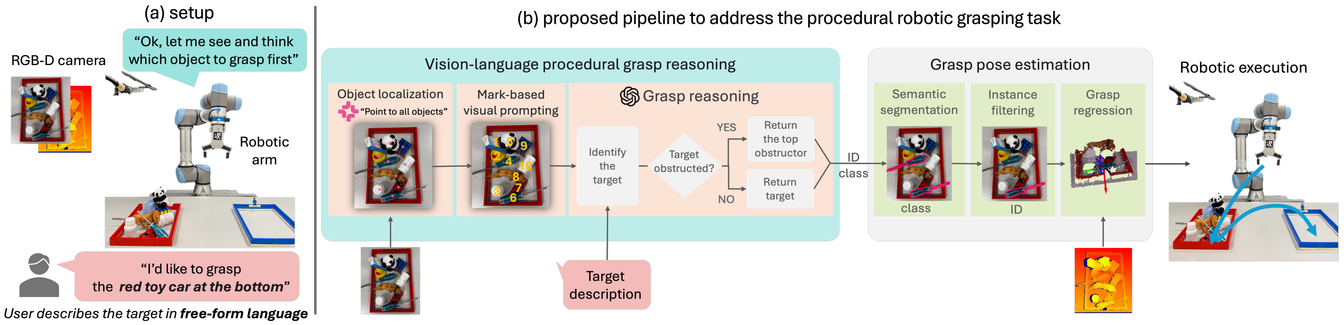

(a) setup
RGB-D camera
"Ok, let me see and think which object to grasp first"
Robotic arm
"I'd like to grasp the red toy car at the bottom"
User describes the target in free-form language
(b) proposed pipeline to address the procedural robotic grasping task
Vision-language procedural grasp reasoning
Object localization
"Point to all objects"
Mark-based visual prompting
Grasp reasoning
Identify the target
Target obstructed?
YES
Return the top obstructor
NO
Return target
Target description
ID
class
Grasp pose estimation
Semantic segmentation
class
Instance filtering
ID
Grasp regression
Robotic execution

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.

Learning-based and traditional grasping. Traditional heuristic or geometry-based approaches often fail in complex clutter. Learning-based methods therefore emphasize relationship reasoning to enhance robustness. Large-scale datasets like MetaGraspNetV2 [8] have become important resources for advancing grasp research.

A. *Overview*

We address the robotic reasoning and grasping task and propose a modular pipeline that leverages the world knowledge of pre-trained VLMs without additional training. The setup consists of a cluttered bin, an exocentric RGB-D camera with top-down view, a robotic arm, and a user who provides a free-form language instruction (e.g., "`grasp the red toy car at the bottom`"). An episode begins upon receiving the instruction and proceeds step by step until the target object is retrieved or termination conditions are met.

As shown in Fig. 2, the pipeline integrates three main modules. First, the *vision-language grasp reasoning* module grounds all objects and reasons about their spatial relationships to decide the next object to grasp. Second, the identified object is segmented at the instance level to provide a precise mask. Third, the *grasp estimation* module computes the optimal grasp pose from the point cloud, and the robot executes the grasp and places the object in a predefined location. This process repeats, with new observations captured after each action, enabling the pipeline to adapt to changes in the scene as objects are removed.

B. *Vision-language grasp reasoning*

Given the RGB observation and the user's free-form instruction, this module leverages state-of-the-art VLMs to determine which object to pick.

Object localization. We explored two strategies: (i) prompting a VLM (e.g., GPT-4o [1]) to list object names and then apply LangSAM [10] for segmentation, and (ii) directly apply Molmo [11] for visual grounding. As shown in Fig. 3, Molmo outperforms the GPT+LangSAM pipeline on MetaGraspNetV2, where the latter often produces fragmented masks. We therefore adopt Molmo for object pointing.

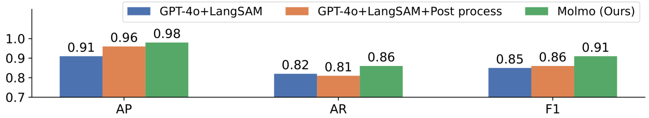


Fig. 3: Object localization performance with different VLM-based method on MetaGrasNetv2.

Mark-based visual prompting. VLMs reason more effectively when the problem is presented in a multiple-choice format [5], [12]. Hence, to each localized object is assigned a unique numeric marker, and the input image is augmented accordingly. This mark-based visual prompt provides GPT with an explicit set of options to choose from, significantly reducing hallucinations and improving spatial reasoning.

Grasp reasoning. The marked image and instruction are then passed to GPT-4o, which decides whether the target is directly graspable or obstructed, and outputs the ID and class name of the next object. To handle ambiguous or varied instructions (e.g., "`juice box`" vs. "`refill pouch`"), the prompt explicitly incorporates spatial references and task context, enabling the model to resolve ambiguities even in cluttered or repetitive-object scenes.

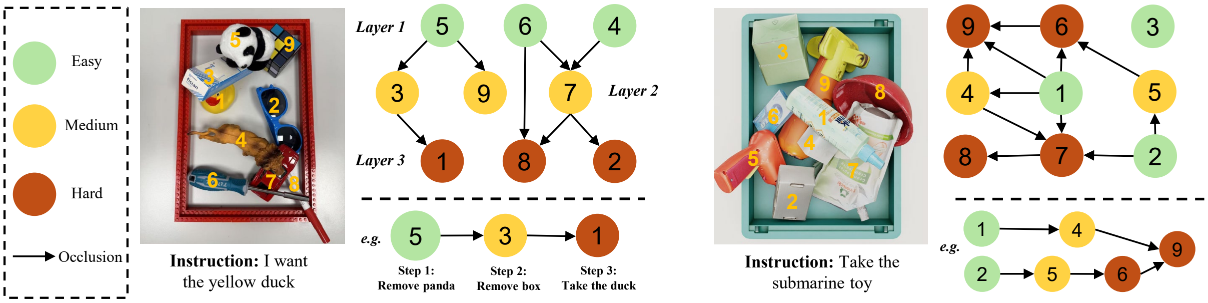
Object segmentation. Finally, the selected object is segmented using LangSAM with the predicted class name and filtered by the coordinates corresponding to the numeric ID to ensure the correct instance mask is obtained.

C. *Grasp estimation*

With the segmented instance, GraspNet [7] estimates the optimal grasp pose from the cropped point cloud. The robot then executes the grasp and places the object in a predefined location. After each grasp, new observations are taken, and the pipeline continues until the user-specified target object is retrieved.

III. Free-form language grasping dataset

We introduce free-from language grasping dataset (FreeGraspData), an evaluation dataset built upon MetaGraspNetV2 [8] to study robotic grasping with free-form language instructions. MetaGraspNetV2 provides simulated bin-picking scenes with RGB-D images and occlusion graphs. To construct FreeGraspData, we select scenes with at least four objects to ensure clutter, and extend the dataset in three aspects: i) derive ground-truth grasp sequences from occlusion graphs, ii) categorize difficulty by obstruction level and object ambiguity, and iii) collect diverse free-form instructions from human annotators. Occlusion graphs are shown in Fig. 4.



(a) Real world (b) FreeGraspData

Fig. 4: Occlusion graphs in representative scenarios. (a) Real-world and (b) synthetic (FreeGraspData) scenes with corresponding occlusion graphs. Nodes are colored by difficulty and edges denote occlusion dependencies; example instructions and action plans are shown below.

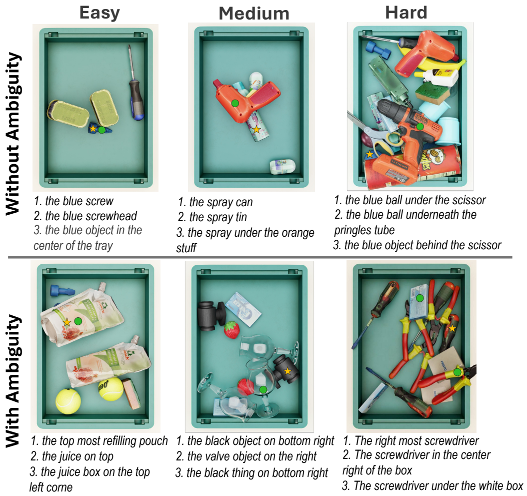

Easy
Medium
Hard
Without Ambiguity
With Ambiguity
1. the blue screw
2. the blue screwhead
3. the blue object in the center of the tray
1. the spray can
2. the spray tin
3. the spray under the orange stuff
1. the blue ball under the scissor
2. the blue ball underneath the pringles tube
3. the blue object behind the scissor
1. the top most refilling pouch
2. the juice on top
3. the juice box on the top left corne
1. the black object on bottom right
2. the valve object on the right
3. the black thing on bottom right
1. The right most screwdriver
2. The screwdriver in the center right of the box
3. The screwdriver under the white box

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

As illustrated in Fig. 5, difficulty is grouped into six categories (Easy/Medium/Hard × with/without Ambiguity). In total, FreeGraspData contains 300 scenarios, each annotated with three user instructions, resulting in 900 evaluation cases. The instructions show substantial linguistic diversity, ranging from synonyms to different descriptive styles. Despite this variation, our GPT-based analysis shows that GPT-4o can still robustly identify the intended target across all difficulty levels.

IV. Experiments

We evaluate FreeGrasp on both the synthetic dataset FreeGraspData and a robotic arm in the real world. We compare against ThinkGrasp [6], which also uses GPT-4o for reasoning. Metrics assess both intermediate reasoning and final grasping performance, and ablation studies validate our design choices.

A. Experiments on FreeGraspData

Setup. FreeGraspData contains 900 evaluation scenarios across six difficulty levels. Each method receives a top-down RGB image and user instructions; results are averaged across three instructions per target.

Metrics. We report the Reasoning Success Rate (RSR), i.e., whether the predicted object ID matches ground truth, and the Segmentation Success Rate (SSR), i.e., whether the output mask has IoU $\geq$0.5 with ground truth. ThinkGrasp only produces segmentation, so we report its SSR. We also evaluate a variant of FreeGrasp using ground-truth localization.

As shown in Tab. I, FreeGrasp consistently outperforms ThinkGrasp, especially in ambiguous and cluttered scenes, demonstrating the benefits of mark-based visual prompting. Both methods perform similarly in simple cases without ambiguity. Interestingly, Molmo-based localization achieves higher RSR than ground-truth localization, as it naturally prioritizes visible objects, improving reasoning robustness.

B. Experiments in the real world

Setup. We evaluate FreeGrasp and ThinkGrasp [6] on a UR5e arm with a parallel gripper and a top-down RGB-D camera, following the same tabletop bin-picking setup as prior work for fair comparison. Ten scenarios are composed for each of the six difficulty levels (Fig. 6), each paired with a free-form language instruction.

Evaluation. In each episode, the robot's task is to grasp and place a user-specified object. We categorize failures into three types: segmentation (S), pose estimation (P), and motion planning (M). To evaluate the system's robustness, we introduce three operational settings with varying strictness: (i) (S, P, M): The episode terminates upon any failure. (ii) (P, M): Segmentation errors are ignored, and the episode only terminates on a pose or motion failure. (iii) (P): The episode terminates only on a pose estimation failure. Performance is measured by Success Rate (SR), Path Efficiency (PE), and Success-weighted by Path Length (SPL) [13].

As shown in Tab. II, FreeGrasp consistently outperforms ThinkGrasp, especially in medium and hard scenarios with clutter and ambiguity. Under the strict (S,P,M) setting, ThinkGrasp fails all medium/hard cases, while FreeGrasp succeeds in many thanks to its vision-language reasoning. Relaxing the segmentation constraint (P,M) slightly improves results when masks include multiple objects but still yield valid grasps. As expected, performance is highest under the relaxed (P) setting, but overall results indicate that real-world robotic grasping with free-form instructions remains highly challenging.

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 [6]	✓	✓	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

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.		
	S	P	M	SR	PE	SPL	SR	PE	SPL	SR	PE	SPL	SR	PE	SPL	SR	PE	SPL	SR	PE	SPL
ThinkGrasp [6]	✓	✓	✓	*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
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	*0.10*	*1.0*	*0.10*
ThinkGrasp [6]		✓	✓	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
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*	*1.0*	*0.10*
ThinkGrasp [6]		✓		*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	0.0	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*	*1.0*	*0.10*

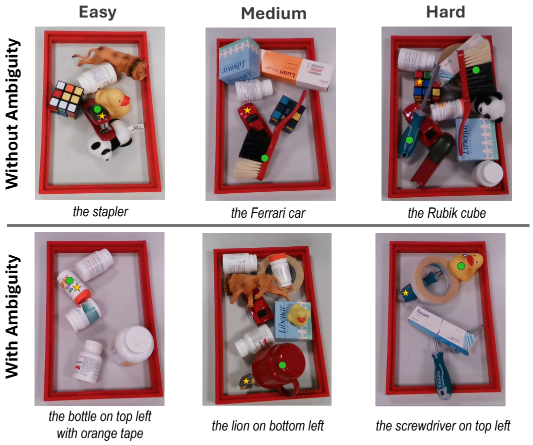


Fig. 6: Samples from real-world experiments for different task difficulties. ★ indicates the user-described target object, and ● are the GT objects to pick.

C. Computational analysis

We benchmark FreeGrasp on a workstation with an RTX 4500 GPU over 60 real-world episodes. The average execution time per step is 15.4s, dominated by object localization (≈9s) and reasoning with GPT-4o (≈5s), while segmentation and grasp estimation are negligible ($<$ 1s). As the camera is externally mounted, reasoning and pose estimation can run in parallel with robotic motion after the first step, reducing overall task latency.

V. Conclusions

We presented FreeGrasp, a novel approach that leverages pre-trained VLMs for robotic grasping by interpreting free-form instructions and reasoning about spatial relationships. While VLMs like GPT-4o show strong general reasoning, our study reveals their limitations in visual-spatial understanding, especially for occlusion. By introducing mark-based visual prompting and contextualized reasoning, FreeGrasp mitigates this gap and outperforms the state-of-the-art ThinkGrasp in both synthetic and real-world evaluations.

Limitations and future work. FreeGrasp still relies on GPT-4o, which struggles with fine-grained occlusion reasoning, and lacks mechanisms to adapt instructions as scenes change with object removal. Future directions include exploring specialized spatial VLMs, memory mechanisms to track scene dynamics, and adaptive instruction updates to improve robustness in evolving environments.

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, 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.

[4] 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.

[5] F. Liu, K. Fang, P. Abbeel, and S. Levine, "MOKA: Open-World Robotic Manipulation through Mark-Based Visual Prompting," in *RSS*, 2024.

[6] 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.

[7] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *CVPR*, 2020.

[8] 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.

[9] 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.

[10] L. Medeiros, "Language Segment-Anything," 2025. [Online]. Available: https://github.com/luca-medeiros/lang-segment-anything.

[11] M. Deitke *et al.*, "Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Multimodal Models," *arXiv:2409.17146*, 2024.

[12] 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.

[13] P. Anderson *et al.*, "On evaluation of embodied navigation agents," *arXiv:1807.06757*, 2018.