Open-Vocabulary and Semantic-Aware Reasoning for Search and Retrieval of Objects in Dynamic and Concealed Spaces

Rohit Menon Yasmin Schmiede Maren Bennewitz Hermann Blum

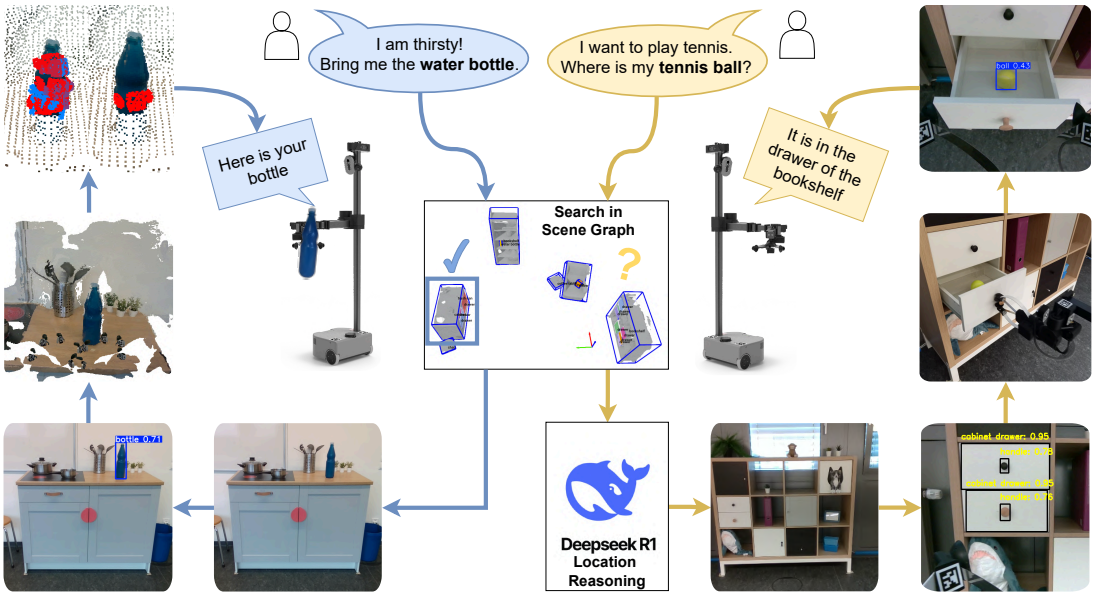

Fig. 1. Overview of Stretch-Compose. The framework addresses two types of open-vocabulary queries: (a) retrieval requests, such as "Bring me the water bottle," where the robot searches and fetches the target, and (b) search requests, such as "Where is my tennis ball?", where concealed-space reasoning and dynamic 3D scene-graph updates are required. Both query types are resolved through the integration of spatial, semantic, and geometric reasoning on the Stretch SE3 platform.

***Abstract*— We present an open-vocabulary framework that combines spatial, semantic, and geometric reasoning to solve a highly relevant, yet under-studied problem: Given an (outdated) map of an environment, how can a robot efficiently retrieve relocated or unmapped items? By unifying spatial cues about proximity and topology, semantic priors on typical placements, and geometric constraints that rule out infeasible locations, particularly within concealed spaces, our approach finds objects even when they are relocated or hidden in drawers or cabinets. We further propose in-situ viewpoint planning to model new objects for manipulation, and to add the object to our dynamic 3D scene graph. We validate our framework through extensive real-world trials on the Stretch SE3 mobile manipulator, evaluating search and retrieval in various conditions. Results demonstrate robust navigation (100%) and open-space detection (100%), with semantic-geometric reasoning reducing concealed space search time by 68% versus semantic-only approaches. Implemented on a low-cost, compact mobile manipulator, our solution combines sophisticated cognitive capabilities with practical deployability, representing a significant step toward accessible service robots for everyday homes.**

All authors are with the University of Bonn, Germany. R. Menon and M. Bennewitz are also with the Humanoid Robots Lab. Y. Schmiede and H. Blum are also with the Robot Perception and Learning Lab. R. Menon, M. Bennewitz, and H. Blum are further affiliated with the Lamarr Institute and the Center for Robotics, Bonn.

This work has partially been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy, EXC-2070 – 390732324 – PhenoRob, and by the BMBF within the Robotics Institute Germany, grant No. 16ME0999.

I. Introduction

Object retrieval is an essential task that any general-purpose household service robot must solve robustly and efficiently. Yet the problem is mostly studied in two simplified variants that are unrealistic: one assumes a static environment with detailed maps from which the previous location of a queried item can be retrieved [1], [2], whereas the other assumes an entirely unknown environment where items must be found without prior information [3]. In reality, domestic settings lie between these extremes. Humans routinely move, conceal, and replace objects [4], while furniture typically remains static for years. A realistic scenario is thus retrieval from an outdated map where some items are unchanged, others relocated, and new ones introduced.

Beyond being moved, objects in homes are often concealed in furniture. This aspect has only recently gained attention [5], typically by mapping once which items lie in drawers or behind doors. Such maps, however, do not help when objects are moved into concealed spaces, and exhaustive search across all compartments is prohibitively time-intensive. An effective service robot must exploit contextual cues, semantic priors, and geometric checks to prioritize which furniture to open and thereby find concealed objects more efficiently.

We introduce *Stretch-Compose*, a modular framework that

unifies spatial, semantic, and geometric reasoning with online viewpoint planning. A 3D semantic scene graph drives a three-stage search: (i) spatial reasoning on the graph ranks reachable open volumes and verifies nearby candidates; (ii) semantic reasoning queries a language model with graph context to propose likely open locations for unseen or relocated objects; and (iii) semantic–geometric reasoning targets concealed spaces by combining semantic priors with size and kinematic feasibility. When the robot observes a previously unseen or relocated object, it performs multi-view open-vocabulary synthesis with onboard RGB-D sensors to reconstruct a 3D object model, updates the scene graph, and supports grasping or the next search step. This integration narrows the search space, limits costly actions, and adapts after each observation (see Fig. 1). Our contributions are:

- A unified framework for open-vocabulary object retrieval that can find relocated, unknown, or concealed objects and maintains an incrementally updated 3D semantic scene graph.
- Spatial, semantic, and geometric reasoning strategies that narrow down and speed up the object search by using contextual and geometric cues.
- An open-source implementation on the low-cost Stretch SE3 platform with real-world experiments verifying the effectiveness of the proposed approach.

II. Related Work

Object retrieval in robotics has advanced through modular point-cloud systems assuming static scenes. SpotCompose [6] segments a pre-scanned 3D map to locate and grasp objects and open drawers using YOLOv8 and OWLv2 [7], while Spotlight [8] extends this with a scene graph for affordance reasoning on switches and doors. Both achieve high success in static environments but fail to address online changes or open-vocabulary queries. In contrast, our approach updates maps and priors on the fly to handle moved or novel objects.

Open-vocabulary mobile manipulation has been explored on real robots using vision–language models. HomeRobot [9] applied DETIC [10] with DDPO policies for navigation and placement, OK-Robot [2] combined Owl-ViT [7], SAM [11], and AnyGrasp [12] for perception and grasp synthesis, whereas CoMeRobot [13] leveraged GPT-4V to generate perception and execution code at run time. These systems validated zero-shot perception, but they typically assumed static maps, relied on heuristic or exhaustive search, and treated 2D detection and 3D reachability as separate stages.

Recent works address dynamic environments by maintaining semantic memories or maps during task execution. DynaMem [14] builds a voxel-based spatio-semantic memory for Stretch SE3, JSR-1 [15] constructs a layered 3D map with LLM-guided proposals, whereas MoMa-LLM [16] reasons over dynamic scene graphs with LLM planning for drawer interactions. These methods adapt to environmental changes, but they do not integrate semantic cues with geometric feasibility and viewpoint planning, which are key to reducing concealed-space search time.

III. Our Approach

Our approach consists of two components: a preprocessing pipeline that builds a 3D environment representation, and a search-and-retrieve strategy that integrates spatial, semantic, and geometric reasoning with grasp planning.

A. Preprocessing Pipeline

The preprocessing pipeline involves scanning the environment, merging and aligning point clouds, segmenting object instances in the 3D point cloud, and organizing the segmented objects in a 3D scene graph.

1) Data Acquisition and Alignment: In the preprocessing stage, we scan the environment with both an iPad Pro and with the robot's onboard RGB-D camera through FUNMAP [17]. In theory only one scan would suffice, but like other works before [2] we find that the SE3's sensor is quite noisy and the iPad yields a more useful initial map. The two point clouds are then aligned with ICP, initialized from a fiducial marker.

2) 3D Instance segmentation and Scene Graph Generation: We first apply OpenMask3D [18] on the iPad scan data to compute CLIP [19] features and generate initial scene masks. To identify furniture items, we also run Mask3D [20], trained on ScanNet200 , to obtain refined instance masks $\mathcal{M} = \{M_k\}$ with semantic labels c_k. From the 200 classes we select those representing furniture. Since contiguous predictions often merge furniture and drawer fronts, we further refine them with a YOLO-Drawer detector [21] that partitions these regions into separate instances.

From the resulting set of instances, we construct a directed scene graph $\mathcal{G} = (V, E)$, where

- $V = \{v_k\}$ are nodes representing furniture, drawers, and objects,
- $(v_i \to v_j) \in E$ denotes a containment relation (e.g., object in drawer, drawer in furniture).

Each node v_k is annotated with centroid $\bar{\mathbf{x}}_k \in \mathbb{R}^3$, axis-aligned bounding box $\mathbf{B}_k$, and semantic class c_k.

B. Search and Retrieval of Objects

We initialize our system with the scene graph created in the preprocessing stage. Our system then updates this scene graph online through detection and interaction to locate, grasp, and retrieve queried object.

1) Spatial Reasoning and Initial Detection: Given a user query for object o^*, we first run a search through the scene graph to check if the object has already been mapped. If successful, this search returns a last-seen node v_ℓ. If more than one node in the scene graph match the query, the node closest to the robot is returned. From v_ℓ we then compute a candidate robot base pose (R_b, t_b) in front of v_ℓ:

$$t_b \;=\; \bar{\mathbf{x}}_\ell + \delta\,\mathbf{n}_\ell, \quad R_b\,\mathbf{e}_x = -\mathbf{n}_\ell, \tag{1}$$

where $\mathbf{n}_\ell$ is the outward normal of the furniture front face (the furniture on or in which v_ℓ sits is simply the parent node) and $\delta > 0$ is a collision radius plus margin. From this pose, the head camera acquires RGB–D data of the

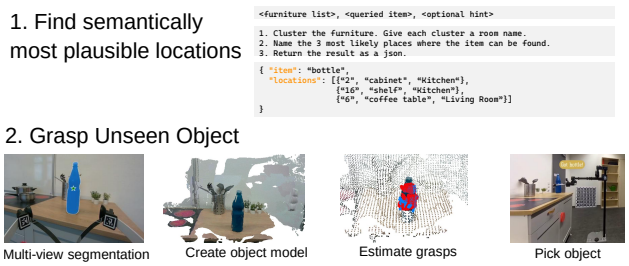

Fig. 2. Semantic reasoning predicts the most plausible furniture locations for a queried item, providing targeted hypotheses that guide subsequent perception and grasping.

full facade and YOLO-World [22] runs on the RGB image. Spatial reasoning assumes limited displacement relative to the last known location and inspects the entire open volume rather than only the stored object position. If o^* is detected, the pipeline proceeds to manipulation; otherwise, it continues with semantic reasoning.

2) Semantic Reasoning: If the object cannot be detected at its previous location or if the scene graph has no entry for o^*, we query DeepSeek-R1 [23] with scene context to retrieve a set of likely search locations. We cluster furniture into room groups using k-means on their (x, y) positions and request a ranked list $\{v_{(1)}, v_{(2)}, v_{(3)}\}$ of plausible furniture conditioned on room and object class. Optional user hints bias the ranking for atypical locations. The system caches the result to avoid repeated queries for the same object. We visit candidates in order, applying the same base-pose computation and YOLO-World check at each location.

3) Geometric Constraint Reasoning: If o^* remains undetected in plain sight, we inspect the occluded spaces in the already visited furniture. For easier readability, we describe in the following the case for drawers, where door compartments mostly work the same except for the motion to open them. When size information for o^* is available, we test fitability against interior drawer dimensions:

$$\text{fit}(o^*, v_d) \;=\; \mathbb{I}\big[l^* \le L_d,\; w^* \le W_d,\; h^* \le H_d\big], \qquad (2)$$

Note that even without drawer manipulation in the mapping stage, we can estimate the upper bound of the drawer's inner dimensions from the bounding box of the front together with the depth of the furniture in the 3D scan.

To interact with a specific drawer we mostly follow the heuristic policy of [6], [8], which we adapt for the Stretch SE3 robot: We move the gripper to the front of the drawer, such that the gripper camera can capture an image to improve drawer-front perception. A YOLO-based drawer detector [21] localizes handles; the handle placement indicates container type (door or drawer) and opening direction. The robot moves the end effector in front of the handle, extends the arm until contact with the front is detected via effort feedback, closes the gripper, and pulls to open until effort feedback indicates the articulation limit. The gripper then moves above the drawer to examine contents, followed by another YOLO-World based object existence check.

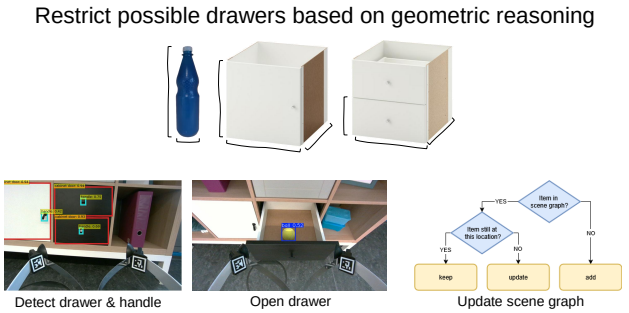

Fig. 3. Geometric reasoning restricts the set of candidate drawers or doors by ruling out infeasible compartments, after which the robot detects and opens the selected drawer and updates the scene graph.

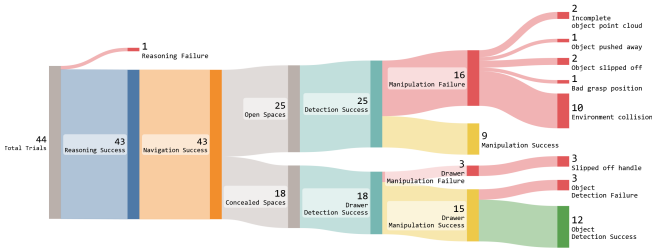

Fig. 4. Sankey diagram of the 44 real-world trials, showing stage-wise outcomes of the search-and-retrieve pipeline. Reasoning and navigation succeed almost universally, while failures occur mainly in the manipulation stage, dominated by collisions with the environment and infeasible grasps.

4) Grasp Planning and Execution: When o^* is visible in an open space or inside a drawer, we estimate its position from depth and re-plan a body pose at a fixed stand-off. The gripper camera collects multi-view RGB–D observations around the object. We segment o^* in the RGB views using SAM2 [24] with multi-point prompts and align the depth observations via ICP to obtain a consolidated object point cloud. For this point cloud, GPD [25] proposes grasps with scores and widths. We filter infeasible grasps imposed by the gripper design, discarding approaches that deviate by more than 12° to the right, and mirroring grasps from behind the object. The robot executes a pre-grasp, uses ArUco tags on the fingers to verify the pre-grasp pose, closes the gripper, lifts the object, and transitions to a carry pose.

5) Scene Graph Update: After each detection, the system updates the scene graph by adding o^* if new or revising its stored location if moved, maintaining consistency between search hypotheses and subsequent actions.

IV. Experimental Evaluation

We evaluate the framework in a lab environment containing a kitchen and living area with multiple pieces of furniture and household objects placed in semantically plausible locations. We test search-and-retrieve tasks for 14 objects under four conditions: objects absent from the scene graph, objects relocated, objects locally displaced on the same furniture, and objects placed inside concealed spaces. Each condition is repeated across different objects, resulting in 44 full trials.

1) Spatial and Semantic Reasoning: The full pipeline achieves 47.7% success across trials. Reasoning performance is consistently strong: semantic reasoning identifies plausible furniture in nearly all cases (97.7%), even when objects are absent from or displaced in the scene graph. Multi-view integration detects 80 % of objects hidden in drawers.

Model	Accuracy	Number of tokens
DeepSeek R1	**0.91**	**613**
OpenAI o1	0.89	911
OpenAI o3-mini	0.74	1086

TABLE I

ABLATION OVER DIFFERENT LLMS FOR SEMANTIC REASONING.

2) Geometric Reasoning: To evaluate the effect of geometric reasoning, which just adjusts the priority order of the different drawers and therefore has no direct effect on the success rate, we measure task execution times with and without geometric reasoning: Over 18 additional trials, geometric reasoning reduces the average execution time per task from $412\,\mathrm{s}$ to $129\,\mathrm{s}$. This is a $68\,\%$ gain by pruning infeasible drawers and doors, turning otherwise exhaustive searches into targeted exploration.

3) LLM Ablation: We also ran a small ablation study on the semantic reasoning module, where we consider an answer correct if the model returns the correct furniture item as the most likely location. Table I shows that we got the best results with DeepSeek R1. We note that DeepSeek also had the highest inference time in our case, but due to different hosting services between the models, an objective comparison of inference time was not possible.

4) Limitations: As shown in Figure 4, most errors stem from manipulation rather than reasoning. Navigation and open-space detection are consistently reliable, but many GPD-proposed grasps are infeasible in clutter or collide with furniture. The limited orientation range of the Stretch gripper, its restricted manipulability, and the fixed pre-grasp stance further constrain feasible grasps, making manipulation the dominant source of failures despite near-perfect reasoning.

In summary, semantic reasoning yields plausible placements and geometric reasoning reduces concealed-space search, while grasp generation remains challenging due to reachability and collision constraints.

V. CONCLUSION

We present Stretch-Compose, a unified framework for open-vocabulary mobile manipulation that combines spatial, semantic, and geometric reasoning with online viewpoint planning. In 44 real-world trials on the Stretch SE3, the system achieves an overall success rate of 47.7%, with semantic reasoning reliably generating plausible placements, geometric reasoning reducing concealed-space search time by 68%, and multi-view integration detecting 80% of hidden objects. These results demonstrate that lightweight reasoning pipelines enable efficient object retrieval in dynamic environments, even on compact and low-cost platforms.

REFERENCES

[1] Q. Gu, A. Kuwajerwala, S. Morin, K. M. Jatavallabhula, B. Sen, A. Agarwal, C. Rivera, W. Paul, K. Ellis, R. Chellappa, *et al.*, "Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning," in *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2024.

[2] P. Liu, Y. Orru, C. Paxton, N. M. M. Shafiullah, and L. Pinto, "Ok-robot: What really matters in integrating open-knowledge models for robotics," *arXiv preprint*, 2024.

[3] K. Zheng, A. Paul, and S. Tellex, "A system for generalized 3d multi-object search," in *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*. IEEE, 2023.

[4] A. J. Miller, "Understanding clutter: Geographies of everyday homes and objects," Ph.D. dissertation, University of Leeds, 2018.

[5] H. Jiang, B. Huang, R. Wu, Z. Li, S. Garg, H. Nayyeri, S. Wang, and Y. Li, "Roboexp: Action-conditioned scene graph via interactive exploration for robotic manipulation," in *Conference on Robot Learning*. PMLR, 2025, pp. 3027–3052.

[6] O. Lemke, Z. Bauer, R. Zurbrügg, M. Pollefeys, F. Engelmann, and H. Blum, "Spot-compose: A framework for open-vocabulary object retrieval and drawer manipulation in point clouds," *arXiv preprint*, 2024.

[7] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, *et al.*, "Simple open-vocabulary object detection," in *Proc. of the Europ. Conf. on Computer Vision (ECCV)*. Springer, 2022.

[8] T. Engelbracht, R. Zurbrügg, M. Pollefeys, H. Blum, and Z. Bauer, "Spotlight: Robotic scene understanding through interaction and affordance detection," *arXiv preprint*, 2024.

[9] S. Yenamandra, A. Ramachandran, K. Yadav, A. Wang, M. Khanna, T. Gervet, T.-Y. Yang, V. Jain, A. W. Clegg, J. Turner, *et al.*, "Homerobot: Open-vocabulary mobile manipulation," *arXiv preprint*, 2023.

[10] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra, "Detecting twenty-thousand classes using image-level supervision," in *Proc. of the Europ. Conf. on Computer Vision (ECCV)*. Springer, 2022.

[11] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[12] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE Trans. on Robotics (TRO)*, 2023.

[13] P. Zhi, Z. Zhang, M. Han, Z. Zhang, Z. Li, Z. Jiao, B. Jia, and S. Huang, "Closed-loop open-vocabulary mobile manipulation with gpt-4v," *arXiv preprint*, 2024.

[14] P. Liu, Z. Guo, M. Warke, S. Chintala, N. M. M. Shafiullah, and L. Pinto, "Dynamem: Online dynamic spatio-semantic memory for open world mobile manipulation," *arXiv preprint*, 2024.

[15] D. Qiu, W. Ma, Z. Pan, H. Xiong, and J. Liang, "Open-vocabulary mobile manipulation in unseen dynamic environments with 3d semantic maps," *arXiv preprint*, 2024.

[16] D. Honerkamp, M. Büchner, F. Despinoy, T. Welschehold, and A. Valada, "Language-grounded dynamic scene graphs for interactive object search with mobile manipulation," *IEEE Robotics and Automation Letters (RA-L)*, 2024.

[17] H. Robot, "Stretch funmap: Fast unified navigation, manipulation, and planning," 2023, accessed: 2025-05-15. [Online]. Available: https://github.com/hello-robot/stretch_ros2/tree/humble/stretch_funmap

[18] A. Takmaz, E. Fedele, R. W. Sumner, M. Pollefeys, F. Tombari, and F. Engelmann, "OpenMask3D: Open-Vocabulary 3D Instance Segmentation," in *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*, 2023.

[19] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. of the Intl. Conf. on Machine Learning*. PmLR, 2021.

[20] J. Schult, F. Engelmann, A. Hermans, O. Litany, S. Tang, and B. Leibe, "Mask3d: Mask transformer for 3d semantic instance segmentation," in *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*. IEEE, 2023.

[21] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[22] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "Yolo-world: Real-time open-vocabulary object detection," *arXiv preprint*, 2024.

[23] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint*, 2025.

[24] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "Sam 2: Segment anything in images and videos," *arXiv preprint*, 2024.

[25] A. Ten Pas, M. Gualtieri, K. Saenko, and R. Platt, "Grasp pose detection in point clouds," *Intl. Journal of Robotics Research (IJRR)*, vol. 36, no. 13-14, 2017.