Active Perception for Robotic Grasping under Occlusion: Semantic-Aware and VLM-Distilled Approaches

by

Tae Hyeon Kweon

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Mechanical and Mechatronics Engineering

Waterloo, Ontario, Canada, 2026
© Tae Hyeon Kweon 2026

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Robotic grasping in cluttered environments is challenging when target objects are heavily occluded, as a single viewpoint is often insufficient for reliable grasp planning. This thesis addresses this problem through two approaches to next-best-view (NBV) planning for active grasping under occlusion. The first approach is a semantic-aware NBV framework that maintains a unified semantic–geometric volumetric representation, integrating geometric and semantic information gain to guide viewpoint selection toward regions most likely to reveal graspable target surfaces. Semantic detections are incrementally accumulated across views through a confidence-weighted update strategy, which adds robustness to false positives and enables autonomous target localization without any prior knowledge of object location. The second approach is a learned NBV policy distilled from Vision-Language Model (VLM) viewpoint preferences. By ranking candidate views offline based on target visibility, occlusion, and grasp feasibility, the VLM provides training labels for a lightweight view scoring network that efficiently selects NBVs at inference without further VLM queries or ray casting. Experiments in simulation and real-world settings validate both approaches, demonstrating that semantic volumetric reasoning and learned VLM preferences are effective and complementary strategies for active perception in robotic grasping.

Acknowledgements

I would like to thank my supervisor, Professor Soo Jeon, for his guidance, support, and encouragement throughout my Master's degree. His feedback and encouragement were invaluable to this work and to my growth during this program.

I would also like to thank the members of the Waterloo Mechanical Systems & Control Laboratory for their support and friendship throughout my time in the lab.

I am grateful to the Waterloo RoboHub for their support with the equipment and facilities that made the experimental work in this thesis possible. I would also like to acknowledge the University of Waterloo and Professor Soo Jeon for their financial support during my program.

Finally, I would like to thank my family and friends for their constant support and encouragement.

Table of Contents

Author's Declaration ii

Abstract iii

Acknowledgements iv

List of Figures viii

List of Tables xi

1 Introduction 1

1.1 Motivation 1

1.2 Contributions 2

1.3 Thesis Outline 3

2 Background and Related Works 5

2.1 Map Representations for RGB-D Perception 5

2.1.1 Point Cloud Representation 5

2.1.2 Truncated Signed Distance Function 6

2.1.3 Semantic TSDF 7

2.2 Grasp Detection for Robotic Manipulation 8

2.3 Active Perception and Next-Best-View Planning 9

2.3.1 Geometric and Task-Oriented NBV . . . 10
2.3.2 Semantic NBV . . . 10
2.3.3 VLM-Based NBV . . . 12
2.3.4 Learning-Based NBV . . . 13
2.4 Vision-Language Models . . . 13
2.4.1 VLMs for Robotic Grasping . . . 13
2.4.2 Knowledge Distillation for Robot Policies . . . 13
2.4.3 Learning from VLM Preferences . . . 14

3 A Semantic-Aware NBV Framework for Active Grasping 16
3.1 Introduction . . . 16
3.2 Methodology . . . 18
3.2.1 Volumetric Mapping with Semantic Integration . . . 19
3.2.2 Semantic-Aware NBV Formulation . . . 21
3.2.3 Semantic Grasp Filtering . . . 24
3.3 Experiments . . . 24
3.3.1 Experimental Setup and Performance Metrics . . . 24
3.3.2 Simulation Results . . . 28
3.3.3 Experimental Results . . . 32
3.4 Conclusion . . . 33

4 Learning Next-Best-View Policy from Vision-Language Model Preferences 34
4.1 Introduction . . . 34
4.2 Methodology . . . 35
4.2.1 Data Collection . . . 36
4.2.2 VLM Labeling . . . 37
4.2.3 View Scoring Network . . . 38

4.2.4 Training . . . 40
4.2.5 Inference . . . 40
4.3 Experiments . . . 41
4.3.1 Teacher Label Validity . . . 41
4.3.2 Position Bias in VLM Labeling . . . 42
4.3.3 Backbone Comparison and the Pose-Only Baseline . . . 44
4.3.4 Context-Reliance Ablation . . . 44
4.3.5 Real-World System Demonstration . . . 45
4.4 Conclusion . . . 45

5 Conclusion and Future Work **47**
5.1 Conclusion . . . 47
5.2 Future Work . . . 48

References **49**

List of Figures

1.1 Example of occlusion in cluttered scenes. From the initial viewpoint (a), the target object (yellow block) is fully hidden behind other items, and the robot cannot determine its location. After moving to an informative viewpoint (b), the target becomes visible, enabling detection and grasp planning. 2

2.1 Comparison of common 3D representations. Point clouds represent sparse surface samples, voxel grids (of which TSDF is a signed-distance variant) represent volumetric occupancy, and RGB-D provides a single-view color and depth capture [17]. 6

2.2 Illustration of the TSDF. Blue regions indicate free space (TSDF > 0), red regions indicate space behind the surface (TSDF < 0), and the zero-level set (TSDF ≈ 0) corresponds to the reconstructed surface. Each box represents a voxel sampled at its center. 7

2.3 Example of a 6-DoF grasp pose (green) predicted by VGN [8] on an object. 8

2.4 Overview of a next-best-view planning loop for grasping. The system iteratively updates its perception of the scene, computes grasp candidates, and plans the next viewpoint until a stopping criterion is met [60]. 9

2.5 Uncertainty-informed viewpoint and action selection for semantic mapping in a cluttered shelf. White regions in the belief map indicate high uncertainty. The robot first selects informative viewpoints to reduce uncertainty, then pushes occluding objects to reveal remaining hidden areas [16]. 11

2.6 VLM-based viewpoint scoring loop, one stage of a larger active perception pipeline. Candidate views are generated by segmenting and projecting the scene, then scored by a VLM based on target visibility, grasp suitability, and occlusion. The highest-scoring candidate is used for subsequent next-best-view generation and 3D scene reconstruction [56]. 12

2.7 Open-world grasping pipeline combining a VLM (GPT-4V) with segmentation and grasp synthesis models. The VLM grounds open-ended language instructions, plans a grasp, and ranks candidates via contact reasoning. (Figure adapted from [54].) 14

2.8 RL-VLM-F generates reward functions for policy learning using a text task description and the agent's visual observations. A VLM is queried to provide preferences over pairs of image observations based on the task description, and a reward function is learned from these preference labels [57]. 15

3.1 The proposed semantic-aware NBV framework actively selects informative viewpoints to refine the geometric and semantic representation of a target object in a cluttered scene prior to grasp execution. 17

3.2 Overview of the proposed semantic-aware NBV planning pipeline for active grasping. The system integrates semantic and geometric information for voxel-based mapping, viewpoint sampling, and grasp execution. Here, ξ denotes a candidate viewpoint, $\mathcal{X}_t$ the set of candidate viewpoints at iteration t, $G(\xi)$ the total information gain combining geometric and semantic terms, $\alpha \in [0, 1]$ the weighting between geometric and semantic information gain, and ξ^*_{t+1} the selected next-best viewpoint for the following iteration. . . . 18

3.3 Semantic region expansion through multi-view observation. (a) Partial observation. (b) Expanded observation. The semantic bounding box (green) expands as more target voxels are observed from different viewpoints. . . . 20

3.4 Geometric and semantic information gain (IG) via ray casting. Here, $\mathrm{TSDF}(v)$ denotes the truncated signed distance value at voxel v (positive in free space, near zero on the surface, negative behind surfaces), $\mathcal{B}_t$ is the semantic bounding box of the detected target at iteration t, and $\alpha \in [0, 1]$ is the semantic IG weight. The gray block extending below the visible target surface represents the unobserved portion of the target object. (a) Geometric IG counts all occluded voxels ($\mathrm{TSDF}(v) < 0$) equally across the workspace. (b) Semantic IG counts occluded voxels within the detected target region, weighted by α, while voxels outside the target region are weighted by $(1 - \alpha)$. In 3D, all rear-side voxels ($\mathrm{TSDF}(v) < 0$) within $\mathcal{B}_t$ contribute to the semantic information gain G_s. 22

3.5 Experimental setup. The robot observes a $30\,\mathrm{cm} \times 30\,\mathrm{cm}$ workspace containing a target object and occluding objects. 25

3.6 Arrangement of objects for (a) experiment and (b) simulation environments. Top row: scene setup. Bottom row: initial camera view. 26

3.7 Sequence of RGB observations, semantic detections, TSDF map updates, and robot motion across multiple viewpoints during semantic-aware NBV planning. The last column ($t = t_f$) shows the final grasp execution. 31

4.1 During data collection, the VLM teacher ranks candidate views based on target visibility. At inference, our distilled ViewScorer selects the next-best view from a single observation in a single forward pass. 35

4.2 Stage 1: Data Collection. The robot observes cluttered scenes from a randomized initial viewpoint, rendering the current RGB image, target ROI, and candidate view images. 36

4.3 Example candidate viewpoint with the predicted grasp proposal overlaid in green, computed by fusing the candidate's depth into a TSDF and querying VGN [8]. 37

4.4 Stage 2: VLM Labeling and Network Training. GPT-4o ranks candidate viewpoints based on target visibility, generating soft labels used to train the view scoring network. 37

4.5 Overview of the VLM labeling prompts. (a) The scoring prompt is applied to every candidate view, producing a score from 0 to 10. (b) When multiple candidates share the highest score, the pairwise tie-resolution prompt selects a winner between the tied views. 39

4.6 Stage 3: Deployment. The trained view scoring network scores candidate viewpoints from the current observation and pose parameters, selecting the next best view without VLM queries. 40

4.7 The initial grid-based labeling approach, in which the VLM selects the highest-scoring candidate from a grid of viewpoints shown together. This approach was later replaced by one-at-a-time candidate scoring due to the position bias observed here. 43

4.8 Real-world pipeline verification. (a, b) External camera views of the scene before and after the NBV policy selects a new viewpoint. (c, d) Corresponding wrist-mounted camera views. The selected view reveals the target object, enabling a successful grasp. 46

List of Tables

3.1 Implementation Parameters 27

3.2 Performance of simulation results 28

3.3 Ablation study 29

3.4 Per-class score evolution under a false positive at $p = 0.9$ followed by correct detections at $p = 0.6$. T and F denote the target and false positive classes, respectively. Bold indicates first correct target class assignment. 30

3.5 Performance of experimental results 32

4.1 Validity of VLM teacher labels against simulator-computed ground truth. . 41

4.2 Positional bias in grid-based and pairwise VLM querying. 42

4.3 Ranking accuracy of vision-conditioned models versus a pose-only baseline. 44

4.4 Context sensitivity under input shuffling 44

Chapter 1

Introduction

1.1 Motivation

When reaching for an object that is partially hidden behind other items, humans do not simply guess and grab blindly. Instead, we instinctively shift our head or body to find a better angle. For example, we move our head around obstacles until we have enough visual information to confidently reach for what we want. This intuitive behavior is an example of active perception, which involves deliberately moving to gather better information before acting.

Robots performing grasping tasks in real-world environments face exactly the same challenge. In settings such as warehouse logistics, domestic assistance, or agricultural harvesting, a robot must frequently pick a specific target object from a cluttered scene where other objects partially or fully block the view. A robot equipped with a camera mounted near its hand sees only what is directly in front of it. If the target object is hidden behind a box or another item, the robot has no way of knowing where to grasp it, or even whether the object is there at all (Fig. 1.1).

One common solution is to move the camera to a better viewpoint before attempting a grasp, much like a person leaning sideways to peek around an obstacle. This way, the robot builds up a more complete picture of the scene and is able to locate where the target object is. This is the core idea behind *Next-Best-View (NBV) planning*, which is actively selecting the most informative camera position at each step until the robot has gathered enough information to act reliably.

However, existing NBV approaches have limitations. Geometric NBV methods have no semantic understanding of the scene and require a predefined 3D bounding box specifying



(a) Initial Viewpoint (b) Next-Best View

Figure 1.1: Example of occlusion in cluttered scenes. From the initial viewpoint (a), the target object (yellow block) is fully hidden behind other items, and the robot cannot determine its location. After moving to an informative viewpoint (b), the target becomes visible, enabling detection and grasp planning.

where the target object is located before exploration begins [9, 48]. More recent semantic NBV approaches incorporate object-level understanding, but typically assume the target object is already visible and do not perform global scene exploration or do not maintain a persistent semantic representation across viewpoints [15, 49, 10]. There remains a need for a framework that can search for, identify, and localize a target object autonomously during exploration without any prior knowledge of its location.

This thesis proposes two complementary approaches to active grasping under occlusion. The first integrates semantic understanding directly into the NBV planning process through a semantic volumetric map, enabling autonomous exploration and localization of the target object without prior knowledge of its location. The second distills VLM viewpoint preferences offline into a lightweight view scoring network, enabling semantic-aware view selection without VLM queries or ray casting at inference.

1.2 Contributions

This thesis presents two complementary approaches to active grasping under occlusion.

The first contribution is a semantic-aware NBV framework that integrates geometric and semantic information gain for active grasping in cluttered scenes, enabling the robot to autonomously explore, localize, and grasp a target object under heavy occlusion without any prior knowledge of its location:

- A semantic-aware NBV framework that combines workspace-level and target-centric geometric information gain through a semantic bounding box that updates as target detections accumulate.
- A unified volumetric representation with aligned geometric and semantic grids that preserves persistent semantic labels and confidence scores across multiple viewpoints.
- A confidence-weighted semantic update strategy that accumulates per-class observation history across views, enabling recovery from high-confidence false positives through consistent multi-view observations.

This work has been accepted for publication in the International Journal of Precision Engineering and Manufacturing (IJPEM) under the title "Semantic-Aware Active Perception for Next-Best-View Grasp Planning".

The second contribution is a learned NBV policy that distills VLM viewpoint preferences for efficient view selection:

- A VLM-based offline labeling strategy that ranks candidate viewpoints based on target visibility, occlusion, and grasp feasibility from 2D images
- A lightweight view scoring network that predicts the next-best view from the current observation and candidate pose parameters in a single forward pass, replacing expensive ray casting or VLM queries at inference time.

1.3 Thesis Outline

This thesis is organized as follows.

Chapter 2 provides the background and related work, covering volumetric representations, robotic grasp detection, active perception and NBV planning, and vision-language models for robotics.

Chapter 3 presents the first proposed method: a semantic-aware NBV framework for active grasping. It describes the volumetric mapping with semantic integration, the combined

information gain formulation, and the semantic grasp filtering approach, along with simulation and real-world experimental results.

Chapter 4 presents the second proposed method: a learned NBV policy distilled from VLM viewpoint preferences. It describes the data collection pipeline, VLM labeling strategy and its validation against simulator ground truth, the view scoring network, and real-world demonstration.

Chapter 5 concludes the thesis with a summary of findings across both methods and outlines directions for future work.

Chapter 2

Background and Related Works

2.1 Map Representations for RGB-D Perception

RGB-D cameras provide both colour and depth information, enabling the reconstruction of 3D structure from a single viewpoint. The depth channel encodes the distance from the camera to each surface point, which can be back-projected into 3D space to form a point cloud. This thesis uses two forms of 3D representation built from RGB-D data: TSDF volumes and point clouds (Fig. 2.1). TSDF volumes are constructed by fusing depth measurements across one or more viewpoints and are used for volumetric grasp prediction. Point clouds are extracted from the fused TSDF and used to generate collision geometry for motion planning. The RGB channel additionally provides appearance information for semantic understanding of the scene.

2.1.1 Point Cloud Representation

A point cloud is a set of 3D points reconstructed by back-projecting depth measurements using the known camera intrinsics. Each point corresponds to a surface sample in the scene and can optionally carry colour information from the RGB channel. Point clouds are lightweight and directly reflect observed geometry, and are commonly used in the literature for single-view grasp detection. However, point clouds only represent surfaces that have been directly observed. They provide no information about free space, occluded regions, or unobserved parts of the scene, which limits their ability to reason about the full 3D structure of an object from a single or partial viewpoint. This limitation is instead



(a) Point Cloud (b) Voxels (c) RGB-D

Figure 2.1: Comparison of common 3D representations. Point clouds represent sparse surface samples, voxel grids (of which TSDF is a signed-distance variant) represent volumetric occupancy, and RGB-D provides a single-view color and depth capture [17].

addressed by the TSDF representation, described in the next subsection. Accordingly, in this thesis, point clouds are used only for collision-geometry generation.

2.1.2 Truncated Signed Distance Function

A TSDF [14] is a volumetric representation that discretizes 3D space into a regular grid of voxels. Unlike point clouds, a TSDF explicitly models free space, surface boundaries, and occluded regions within a single unified volume. Each voxel stores the signed distance from its center to the nearest observed surface, truncated within a predefined margin and normalized to the range $[-1, 1]$. The volumetric grid is constructed by integrating depth observations from multiple viewpoints, allowing new measurements to update existing voxels coherently. This makes TSDF particularly well suited for active perception tasks where the scene is observed incrementally and occlusion awareness is critical. Voxels near zero correspond to surface boundaries, positive values (TSDF > 0) indicate free space, and negative values (TSDF < 0) indicate space behind the observed surface, as illustrated in Figure 2.2.

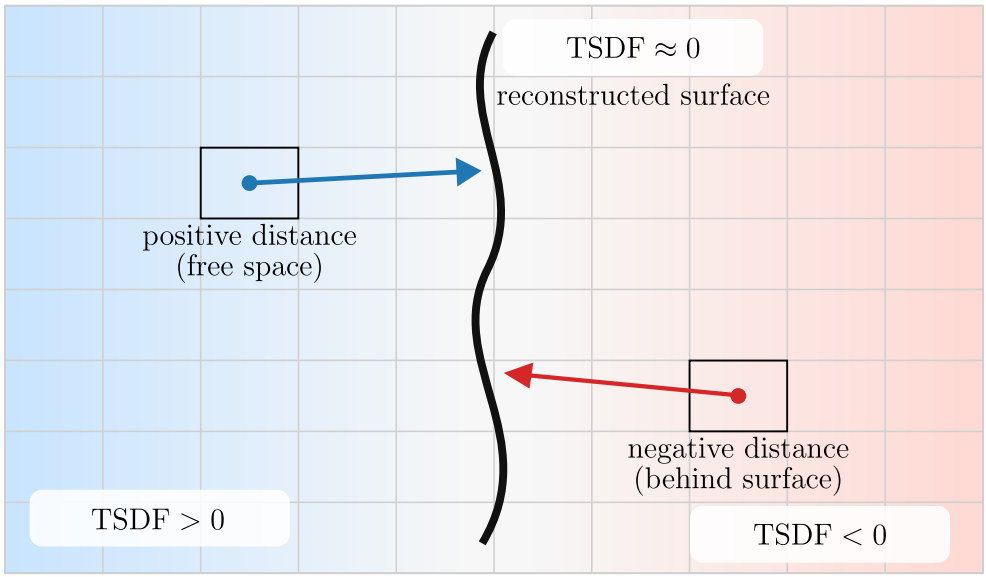


Figure 2.2: Illustration of the TSDF. Blue regions indicate free space (TSDF > 0), red regions indicate space behind the surface (TSDF < 0), and the zero-level set (TSDF ≈ 0) corresponds to the reconstructed surface. Each box represents a voxel sampled at its center.

2.1.3 Semantic TSDF

While a standard TSDF captures the geometric structure of a scene, it contains no information about the semantic meaning of the observed surfaces. A semantic TSDF extends the standard formulation by associating each voxel with a class label and a confidence value in addition to the signed distance. This allows the volumetric map to represent not only where surfaces are, but also what objects occupy each region of space. Semantic labels are assigned to voxels by projecting class predictions from an image-based semantic segmentation model into the volumetric grid. As new observations arrive from different viewpoints, both the geometric and semantic information are updated incrementally, allowing label confidence to accumulate over time.

Prior work has explored integrating semantic labels directly into volumetric maps. SemanticFusion [40] jointly reconstructs geometry and semantics by projecting CNN-based predictions into a surfel map, while Voxblox++ [21] extends voxel-based TSDF reconstruction with dense semantic labelling of all surfaces. These approaches tightly couple geometric and semantic reconstruction into a single unified map.



Figure 2.3: Example of a 6-DoF grasp pose (green) predicted by VGN [8] on an object.

2.2 Grasp Detection for Robotic Manipulation

Robotic grasp detection aims to estimate feasible grasp poses from visual observations. In 6-DoF grasp detection, each grasp candidate is typically represented by a position, orientation, gripper width, and grasp quality score. Early analytical methods rely on geometric heuristics [7, 41], while modern learning-based approaches leverage deep neural networks trained on large-scale datasets [18, 39].

Volumetric grasp prediction methods represent the scene as a 3D voxel grid and predict grasp quality over the reconstructed volume. Volumetric methods such as Volumetric Grasping Network (VGN) [8] and GIGA [26] employ 3D convolutional or implicit neural networks on Truncated Signed Distance Function (TSDF) representations to predict 6 Degrees of Freedom (DoF) grasps from partial observations (Fig. 2.3). Specifically, VGN [8] is a fully convolutional network that takes a TSDF volume as input and predicts a per-voxel grasp quality, orientation, and gripper width at the same spatial resolution. Candidates are extracted by thresholding the quality map and applying non-maximum suppression. The network is trained on simulated grasp trials, with ground-truth labels obtained by physically simulating gripper closure at sampled candidate poses, allowing it to operate on partial or fused multi-view TSDF reconstructions without requiring complete object models. This formulation is particularly useful for active perception since multi-view observations can be fused into the TSDF before grasp prediction.

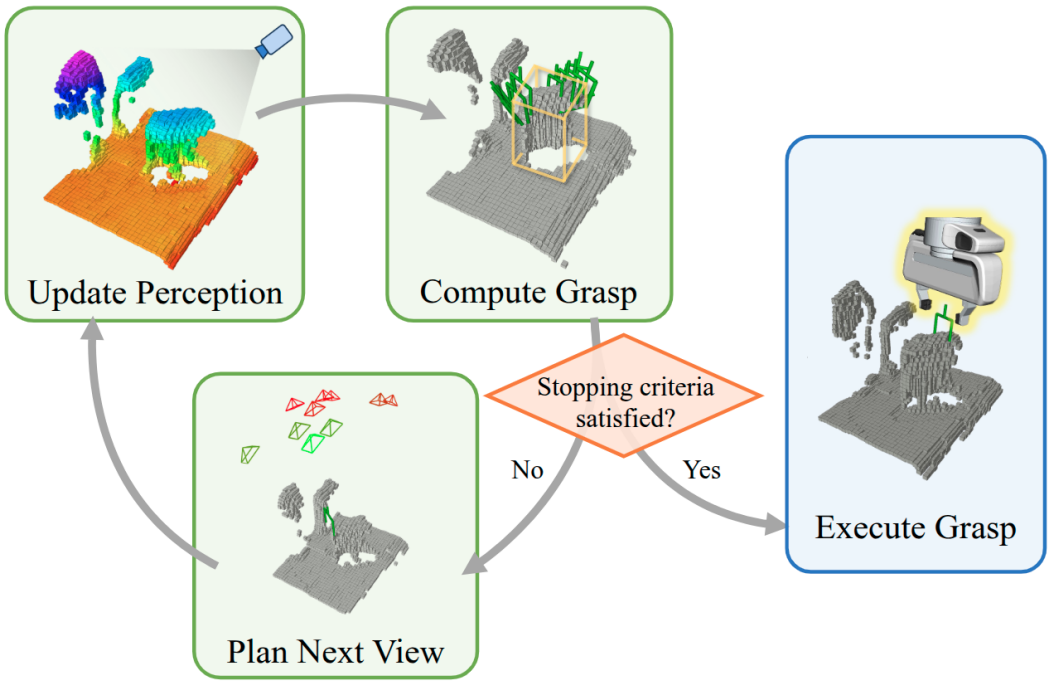


Figure 2.4: Overview of a next-best-view planning loop for grasping. The system iteratively updates its perception of the scene, computes grasp candidates, and plans the next viewpoint until a stopping criterion is met [60].

Point-cloud-based methods directly operate on 3D points reconstructed from depth observations. AnyGrasp [2] is an example of a general 6-DoF grasp detector that predicts grasp candidates from point cloud inputs.

Multi-view approaches have been explored to improve grasp detection under occlusion. Morrison et al. [42] propose a sequential observation strategy that aggregates grasp predictions from multiple viewpoints.

2.3 Active Perception and Next-Best-View Planning

Active perception approaches, particularly Next-Best-View (NBV) planning, provide a principled framework to address partial observability by iteratively selecting informative viewpoints that reveal hidden regions [25]. By moving the camera to observe the scene

from different viewpoints, NBV methods improve visibility of occluded regions and obtain more informative observations for grasp planning (Fig. 2.4).

2.3.1 Geometric and Task-Oriented NBV

Early NBV methods for grasping focus on geometric reconstruction. Breyer et al. [9] propose a closed-loop NBV framework that maximizes volumetric information gain via ray casting, while Schaub et al. [48] extend this with probabilistic TSDF formulations, improving robustness to noise and occlusion. More recent approaches incorporate task-oriented objectives beyond geometry. ACE-NBV [60] proposes an affordance-driven policy based on implicit grasp detection, while Ma et al. [38] propose a neural graspness field that estimates grasp quality over the scene for NBV planning.

2.3.2 Semantic NBV

Other works have integrated semantic information into NBV planning frameworks to focus exploration on task-relevant objects. Kay et al. [28] extend volumetric information gain with semantic weighted entropy for aerial reconstruction, multiplying geometric entropy by class-specific utility weights to distinguish target from nuisance objects. Burusa et al. [10] develop a semantics-aware NBV planner that maintains semantic voxel grids with class labels and confidence scores, using entropy-based semantic information gain to prioritize viewpoints that reduce uncertainty about task-relevant plant parts. For manipulation-oriented tasks, Song et al. [50] propose a geometry-based, semantics-aware NBV planner for avocado harvesting that evaluates candidate viewpoints using semantic entropy under task-specific geometric constraints.

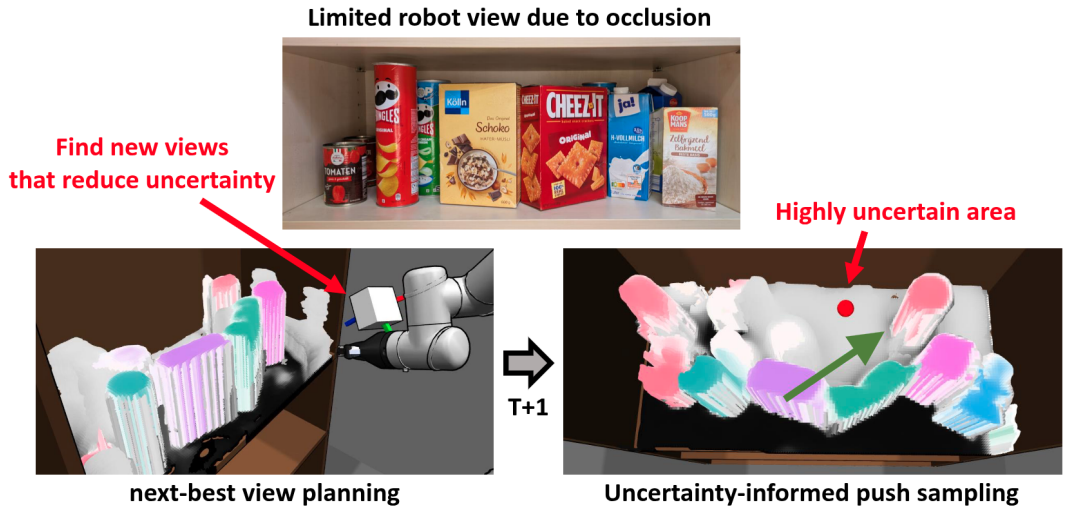


Figure 2.5: Uncertainty-informed viewpoint and action selection for semantic mapping in a cluttered shelf. White regions in the belief map indicate high uncertainty. The robot first selects informative viewpoints to reduce uncertainty, then pushes occluding objects to reveal remaining hidden areas [16].

Recent advances in semantic segmentation and object detection have enabled real-time identification of objects of interest [23, 29]. Several works incorporate semantic information into active perception for manipulation tasks. Dengler et al. [16] maintain semantic volumetric maps with evidential uncertainty estimates, combining uncertainty-aware viewpoint selection and push actions for semantic mapping in cluttered shelf environments (Fig. 2.5). Koc and Sariel [31] enrich volumetric representations with object-aware semantic information to evaluate active and interactive actions for tabletop scene exploration. These approaches focus on complete scene understanding rather than target-object-specific exploration.

2.3.3 VLM-Based NBV

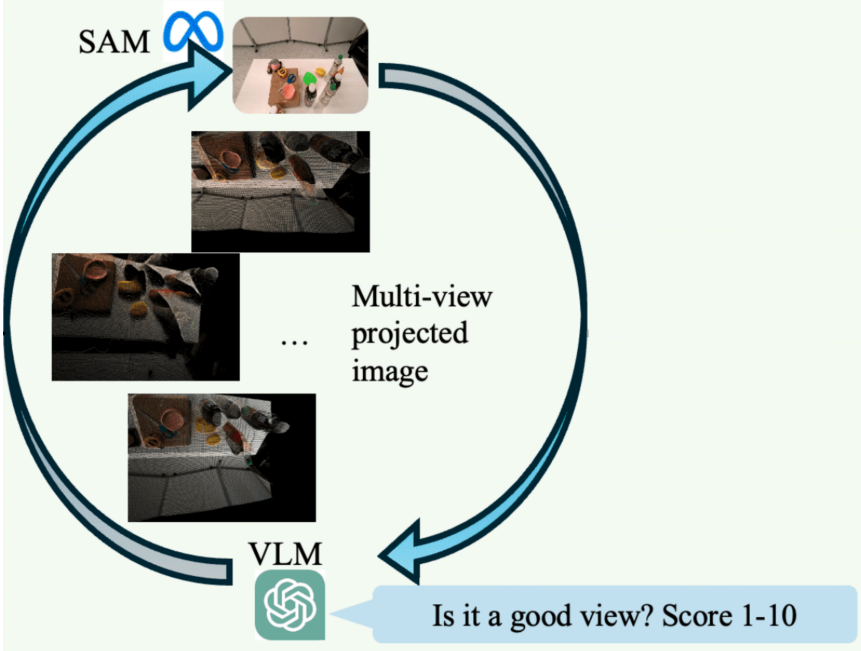


Figure 2.6: VLM-based viewpoint scoring loop, one stage of a larger active perception pipeline. Candidate views are generated by segmenting and projecting the scene, then scored by a VLM based on target visibility, grasp suitability, and occlusion. The highest-scoring candidate is used for subsequent next-best-view generation and 3D scene reconstruction [56].

More recently, vision-language models (VLMs) have been applied to semantic-aware viewpoint planning. Wang et al. [56] leverage VLMs to score candidate viewpoints based on target visibility, grasp suitability, and occlusion (Fig. 2.6), while Liu et al. [34] use a VLM to estimate a per-view ambiguity score for 6-DoF pose estimates, combining it with geometric pose entropy to select a next-best-view that resolves viewpoint-induced pose ambiguity. APeG [15] incorporates CLIP [47] text-image similarity for language-guided grasping and selects semantic-aware viewpoints that reveal occluded regions around the target. VISO-Grasp [49] reasons about occlusions by identifying likely occluding objects when the target object is not detected and uses NBV fields to guide sequential grasping. Across these methods, the role of the VLM varies: Wang et al. and Liu et al. use the VLM to directly score or rank candidate viewpoints, while APeG and VISO-Grasp use the

VLM to inform semantic understanding or occlusion reasoning, with the viewpoint itself computed by a separate geometric optimization. These object-centric approaches handle occlusion locally, without performing global scene exploration or maintaining a persistent semantic representation over unobserved regions.

2.3.4 Learning-Based NBV

Beyond VLM-based approaches, active perception has also been explored in the context of foundation models and imitation learning. I-Perceive [5] frames active viewpoint selection as a foundation model conditioned on language instructions, while ActiveVLA [6], EgoMI [3], and ActiveUMI [4] learn active viewpoint policies from demonstrations.

2.4 Vision-Language Models

2.4.1 VLMs for Robotic Grasping

Recent works have demonstrated that VLMs can reason about grasping from visual observations (Fig. 2.7). OWG [54] uses GPT-4V to rank grasp candidates based on contact reasoning, showing that VLMs can assess physical grasp poses when presented with grasp candidates visually. ThinkGrasp [45] leverages GPT-4o to select grasp regions in cluttered scenes, while AffordGrasp [51] and OVAL-Grasp [53] use VLMs to identify graspable affordance regions from implicit user instructions.

2.4.2 Knowledge Distillation for Robot Policies

Knowledge distillation [24] compresses expensive teacher models into lightweight students by training the student to reproduce the teacher's outputs. In robotics, distillation has been used to compress large generalist policies into specialized experts [27] and to distill LLM-guided robot demonstrations into multi-task visuomotor policies [22]. A related variant, privileged-information distillation [36], allows the teacher to access additional inputs during training that are unavailable to the student at deployment, enabling the student to benefit from richer supervision without requiring those inputs at inference time.

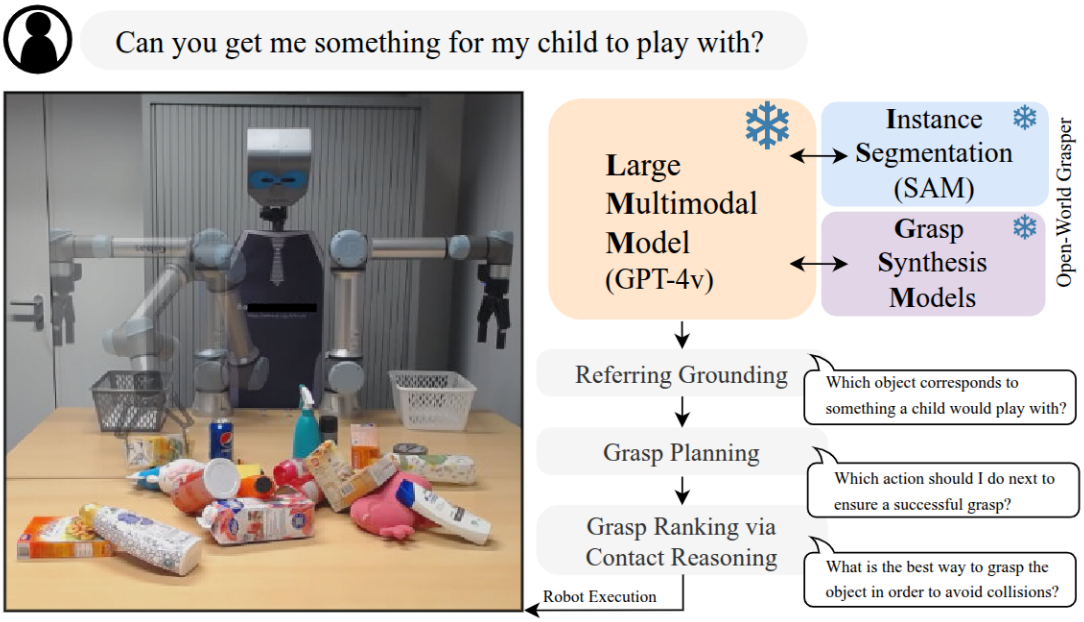


Figure 2.7: Open-world grasping pipeline combining a VLM (GPT-4V) with segmentation and grasp synthesis models. The VLM grounds open-ended language instructions, plans a grasp, and ranks candidates via contact reasoning. (Figure adapted from [54].)

2.4.3 Learning from VLM Preferences

There has been growing interest in using VLMs as preference annotators to generate training supervision. RL-VLM-F [57] established this paradigm by prompting a VLM to provide pairwise preferences over agent observations given a task description, learning a reward function from those labels without any human annotation (Fig. 2.8). Offline RL-VLM-F [55] extends this to the offline setting, first collecting a dataset of trajectories and then using a VLM offline to generate preference labels, eliminating the need for online VLM queries during training. PrefVLM [19] combines VLM-generated trajectory preferences with targeted human feedback to handle label noise. In a related direction, PLARE [37] queries a VLM offline for pairwise trajectory preferences and trains a policy directly from those labels. This line of work demonstrates that VLMs can be used to generate supervision offline.

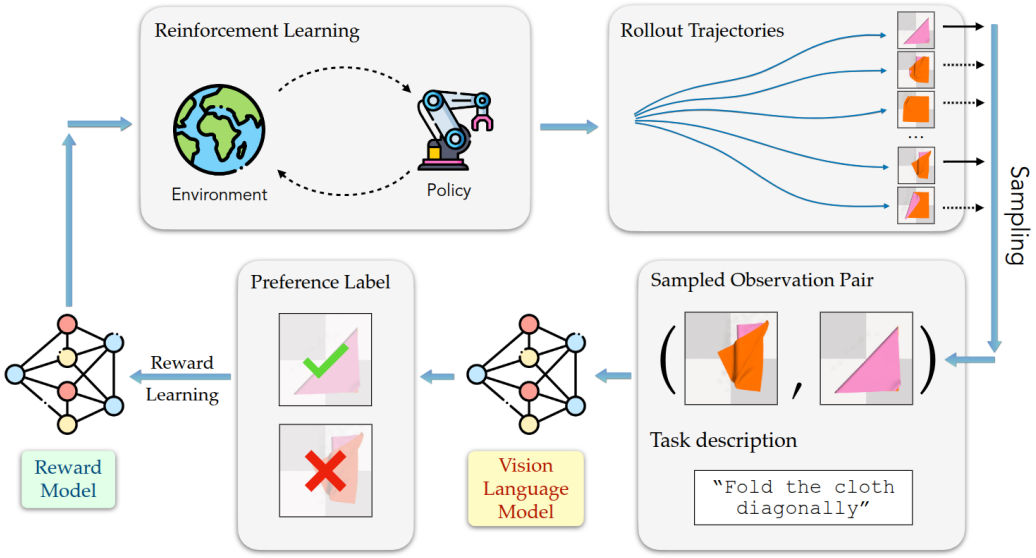


Figure 2.8: RL-VLM-F generates reward functions for policy learning using a text task description and the agent's visual observations. A VLM is queried to provide preferences over pairs of image observations based on the task description, and a reward function is learned from these preference labels [57].

Chapter 3

A Semantic-Aware NBV Framework for Active Grasping

3.1 Introduction

Vision-guided robotic grasping in unstructured environments remains a fundamental challenge in robotics, particularly when target objects are partially occluded or have significant hidden surfaces [58, 59]. In real-world settings such as manufacturing, logistics, domestic assistance, and agricultural harvesting, robots must frequently manipulate objects in cluttered scenes where target objects are often partially occluded by other items. A typical configuration of vision-guided robotic grasping uses an RGB-D camera mounted near the end-effector (i.e., eye-in-hand camera). With this configuration, single-view grasp planning methods cannot observe hidden surfaces that may contain the most graspable regions, leading to frequent failures when the visible surfaces offer poor grasp affordances [42].

To address this, in this chapter, we propose a semantic-aware NBV planner for grasping that leverages semantic information to focus sensing on objects of interest (see Fig. 3.1 for the overall procedure).[1] We maintain a unified volumetric representation that fuses geometric TSDF and semantic voxel grids, accumulating semantic labels and confidence scores across multiple viewpoints. Semantic detections construct an adaptive spatial bounding box, within which geometric information gain is evaluated to guide target-focused exploration. Finally, we introduce semantic grasp filtering that leverages the accumulated

[1] This chapter has been accepted for publication in the *International Journal of Precision Engineering and Manufacturing*.



Figure 3.1: The proposed semantic-aware NBV framework actively selects informative viewpoints to refine the geometric and semantic representation of a target object in a cluttered scene prior to grasp execution.

semantic representation to ensure that grasps target the intended object rather than occluders.

The remainder of this chapter is organized as follows. Section 3.2 describes the proposed methodology. Section 3.3 presents simulation and real-world experimental results. Section 3.4 concludes the chapter.

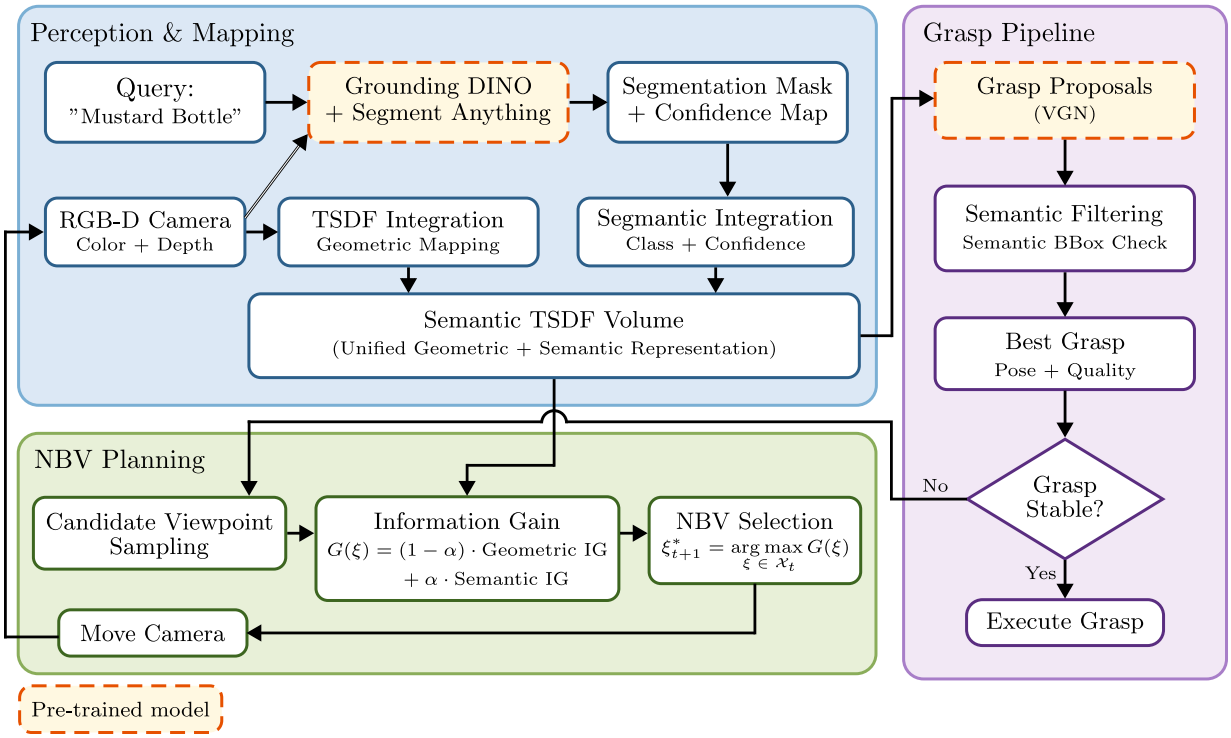


Figure 3.2: Overview of the proposed semantic-aware NBV planning pipeline for active grasping. The system integrates semantic and geometric information for voxel-based mapping, viewpoint sampling, and grasp execution. Here, ξ denotes a candidate viewpoint, $\mathcal{X}_t$ the set of candidate viewpoints at iteration t, $G(\xi)$ the total information gain combining geometric and semantic terms, $\alpha \in [0, 1]$ the weighting between geometric and semantic information gain, and ξ^*_{t+1} the selected next-best viewpoint for the following iteration.

3.2 Methodology

As shown in Fig. 3.2, the system operates in a closed-loop exploration-grasp cycle targeting a specified object in a cluttered scene, leveraging VGN [8] for grasp candidate generation and GroundingDINO [35] with the Segment Anything Model (SAM) [29] for open-vocabulary semantic segmentation. At each iteration, the robot: (1) acquires RGB-D observations with object detections and segmentation, and updates the geometric TSDF and semantic voxel representations (Perception & Mapping), (2) evaluates grasp candidates on the target object (Grasp Pipeline), and (3) either executes the highest-quality valid grasp if the grasp condition is met, or selects the next best viewpoint via information

gain maximization to continue exploration (NBV Planning). The cycle continues until a successful grasp is executed. VGN [8] predicts 6-DoF grasp candidates from the accumulated TSDF volume and outputs grasp candidates with associated quality scores and 6-DoF poses. More details of each step are explained below.

3.2.1 Volumetric Mapping with Semantic Integration

We represent the geometry of the scene using a TSDF volume [13], where each voxel stores the signed distance to the nearest surface. The TSDF is incrementally constructed through volumetric integration of depth observations from multiple viewpoints, providing a continuous volumetric representation of the scene geometry. This representation enables the identification of occluded voxels behind observed surfaces for the computation of information gain and serves as input to the VGN [8] for the prediction of 6-DoF grasp poses. Although there are some newer grasping networks, e.g. [2], we employed VGN in this work because of its direct compatibility with the TSDF representation, enabling tight integration between the volumetric mapping and grasp generation modules.

In this work, we take an approach to maintain a unified voxel-based representation that stores both geometric and semantic information in aligned voxel grids. Let $\mathcal{V}$ denote the set of all voxels within a bounding box for the entire workspace $\mathcal{W}$. At each iteration step t, a set of semantic voxels $\mathcal{S}_t \subset \mathcal{V}$ is identified (see Fig. 3.3). When the semantic voxel set $\mathcal{S}_t$ is constructed (as explained below in 3.2.1), each voxel element $v \in \mathcal{S}_t$ is associated with a class label $c_t(v) \in \{0, 1, ..., N_c\}$ and detection confidence $p_t(v) \in [0, 1]$, where N_c is the number of identifiable classes and $c_t(v) = 0$ indicates unlabeled voxels.

While the semantic voxel representation can accommodate multiple object classes, in this work semantic integration is restricted to the specified target object class, as the active grasping task requires identifying only the grasp target.

Construction of Semantic Voxels

At each iteration time t, the camera pose $\xi_t \in SE(3)$ generates the RGB-D observation. Given a text prompt for a single target object class, GroundingDINO [35] produces a 2D detection box, a class label for the target c_t, and its confidence score p_t. The detected regions are then segmented using a segmentation algorithm (e.g., SAM [29]), producing pixel-wise masks in 2D. These masks are projected onto 3D space through ray-casting from the camera pose $\xi_t \in SE(3)$ taking into account the camera intrinsics. This process

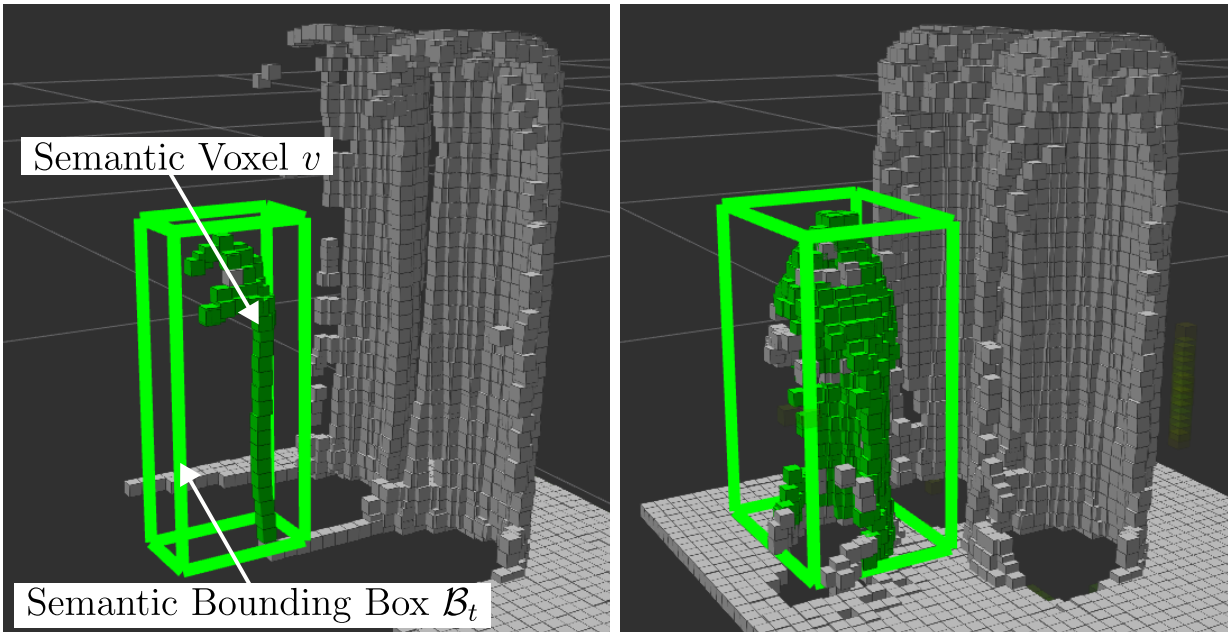


Figure 3.3: Semantic region expansion through multi-view observation. (a) Partial observation. (b) Expanded observation. The semantic bounding box (green) expands as more target voxels are observed from different viewpoints.

generates the semantic voxel set $\mathcal{S}_t$, where each voxel $v \in \mathcal{S}_t$ is assigned a class label $c_t(v)$ and a confidence score $p_t(v)$.

Semantic labels are integrated only when the detector confidence exceeds a threshold τ_p. By discarding low-confidence detections, this reduces incorrect semantic labels from false-positive target detections (e.g., occluders misidentified as the target object) and ensures that semantic information gain is driven primarily by high-confidence target observations.

Unlike approaches that tightly couple geometric and semantic reconstruction such as SemanticFusion [40] and Voxblox++ [21], we maintain separate but aligned representations. This modular design preserves standard TSDF reconstruction while restricting semantic integration on detected objects, rather than performing dense semantic labeling of all surfaces.

Update of Semantic Voxels

As the robot observes the scene from multiple viewpoints, the same voxel may receive different semantic labels due to occlusion, viewpoint variation, or detector noise. To maintain consistent semantic assignments, the semantic voxel set $\mathcal{S}_t$ accumulates labeled voxels across iterations, i.e., $\mathcal{S}_{t-1} \subseteq \mathcal{S}_t$, preserving semantic information from previous viewpoints.

When a voxel is re-observed, we incrementally accumulate per-class observation counts and confidence scores across all views. Let $N_t(v) = \sum_{k=1}^{t} \mathbb{I}[c_k(v) \neq 0]$ denote the total number of views at which voxel v received a detection, and $\sigma_t(v, c) = \sum_{k=1}^{t} p_k(v) \cdot \mathbb{I}[c_k(v) = c]$ the accumulated confidence for class c at voxel v. The per-class confidence score, normalized by the total number of observations across all classes is computed as:

$$\bar{p}_t(v, c) = \frac{\sigma_t(v, c)}{N_t(v)}, \quad \forall v \in \mathcal{S}_{t-1} \tag{3.1}$$

and the semantic label and confidence are assigned as:

$$c_t(v) = \arg\max_{c} \ \bar{p}_t(v, c), \quad \forall v \in \mathcal{S}_{t-1} \tag{3.2}$$

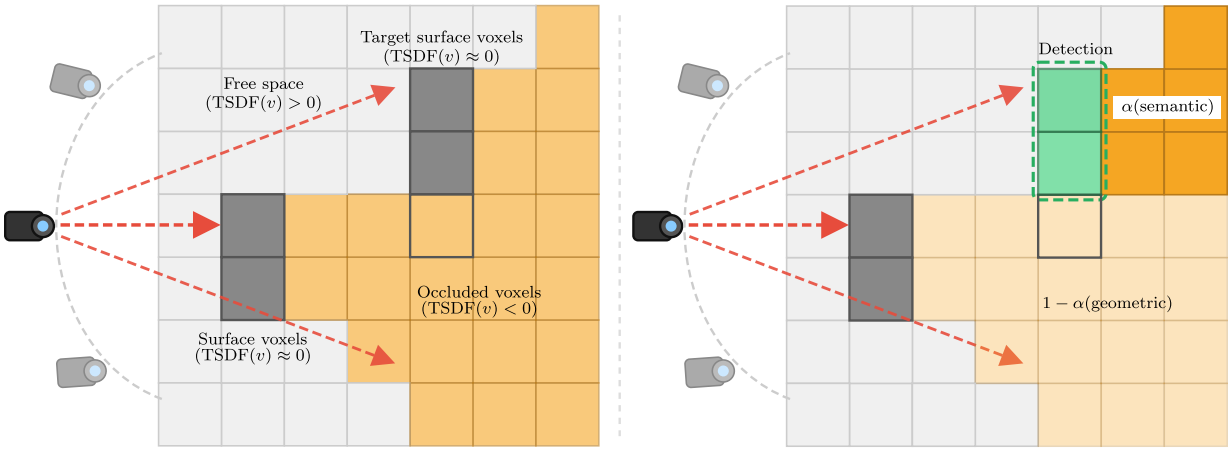
and $p_t(v)$ is set to the raw detection confidence from the most recent view at which class $c_t(v)$ was detected at v. The normalized score $\bar{p}_t(v, c)$ is used for class selection to ensure robustness to early false positives, while the stored confidence $p_t(v)$ reflects the latest observation of the assigned class.

3.2.2 Semantic-Aware NBV Formulation

Our key contribution is extending geometric NBV planning by focusing exploration on semantically detected target regions through an adaptive bounding box that guides geometric information gain computation. All information gain functions and the TSDF are evaluated relative to the current map state at iteration t, and thus implicitly depend on t through the accumulated observations.

Geometric Information Gain

Following Breyer et al. [9], the geometric information gain (IG) quantifies the number of occluded voxels that would become visible from a candidate viewpoint, computed via rear-side voxel counting through ray casting. In our formulation, this gain is computed over



(a) Geometric IG

(b) Semantic-aware IG

Figure 3.4: Geometric and semantic information gain (IG) via ray casting. Here, TSDF(v) denotes the truncated signed distance value at voxel v (positive in free space, near zero on the surface, negative behind surfaces), $\mathcal{B}_t$ is the semantic bounding box of the detected target at iteration t, and $\alpha \in [0, 1]$ is the semantic IG weight. The gray block extending below the visible target surface represents the unobserved portion of the target object. (a) Geometric IG counts all occluded voxels (TSDF(v) < 0) equally across the workspace. (b) Semantic IG counts occluded voxels within the detected target region, weighted by α, while voxels outside the target region are weighted by $(1 - \alpha)$. In 3D, all rear-side voxels (TSDF(v) < 0) within $\mathcal{B}_t$ contribute to the semantic information gain G_s.

the entire workspace region $\mathcal{W}$ to promote broad scene exploration and initial detection of the target object.

Let $\mathcal{R}(\xi) \subset \mathcal{W}$ denote the set of voxels visible from viewpoint ξ within the workspace $\mathcal{W}$, determined via ray casting through the camera. For each candidate viewpoint, rays are cast from the camera center, traversing voxels until the first observed surface is encountered (TSDF ≈ 0). Voxels behind this surface, where TSDF < 0, are considered occluded (rear-side voxels).

The geometric IG is computed as:

$$G_{\mathrm{g}}(\xi) = \frac{1}{|\mathcal{W}|} \sum_{v \in \mathcal{R}(\xi)} \mathbb{I}[\mathrm{TSDF}(v) < 0], \tag{3.3}$$

where $\mathbb{I}[\bullet]$ is the indicator function whose value is 1 if the statement $\bullet$ is true and 0 otherwise, and $|\mathcal{W}|$ denotes the total number of voxels in the workspace $\mathcal{W}$. This quantifies

the proportion of currently occluded voxels within the workspace that would become observable from viewpoint ξ. Normalizing by $|\mathcal{W}|$ expresses the gain as a fraction, enabling direct comparison with the semantic IG computed over the smaller target region.

Semantic-aware Geometric Information Gain

To focus semantic exploration on the target object, we define the semantic bounding box $\mathcal{B}_t \supset \mathcal{S}_t$ as the smallest axis-aligned bounding box containing the semantic voxel set $\mathcal{S}_t$ (see the green box in Fig. 3.3). As shown in Fig. 3.3, the semantic bounding box $\mathcal{B}_t$ dynamically updates as the robot observes the target from multiple viewpoints.

The semantic-aware IG applies the same geometric rear-side voxel counting, but restricts the computation to voxels within the semantic bounding box (see Fig. 3.4):

$$G_{\mathrm{s}}(\xi) = \frac{1}{|\mathcal{B}_t|} \sum_{v \in \mathcal{R}(\xi) \cap \mathcal{B}_t} \mathbb{I}[\mathrm{TSDF}(v) < 0], \tag{3.4}$$

where $|\mathcal{B}_t|$ denotes the total number of voxels in the semantic bounding box $\mathcal{B}_t$. This quantifies the proportion of currently occluded voxels within the semantic region that would become observable from viewpoint ξ.

Combined Information Gain and View Selection

This dual-scale formulation enables workspace-level exploration via G_{g} for initial scene understanding, while the semantic-aware geometric information gain G_{s} guides exploration toward the detected target region. The total IG denoted by $G(\xi)$ is then computed as:

$$G(\xi) = (1 - \alpha) \cdot G_{\mathrm{g}}(\xi) + \alpha \cdot G_{\mathrm{s}}(\xi) \tag{3.5}$$

where $\alpha \in [0, 1]$ balances workspace-level exploration and semantic refinement toward the target region. A higher α focuses the robot on reducing occlusions within the target region, while a lower α favors broader workspace exploration. Since both G_{g} and G_{s} are normalized as fractions within their respective regions, they lie on the same $[0, 1]$ scale, ensuring α provides meaningful control over the balance between the two terms. When no semantic detections are available (i.e., $\mathcal{B}_t = \emptyset$), the semantic gain G_s is set to zero and the combined information gain reduces to the geometric information gain G_g.

Candidate viewpoints are uniformly sampled on a hemisphere of radius R_s centered at the workspace center. The hemisphere radius R_s is chosen to ensure the workspace remains visible while satisfying kinematic constraints.

Denoting the set of candidate viewpoints at t by $\mathcal{X}_t$, the next best view is selected as:

$$\xi_{t+1}^* = \underset{\xi \in \mathcal{X}_t}{\arg\max}\ G(\xi) \tag{3.6}$$

3.2.3 Semantic Grasp Filtering

For each camera view ξ_t, the VGN generates a set of grasp candidates $\mathcal{G} = \{g_1, g_2, ..., g_m\}$, where m is the number of grasp candidates and $g_i = (\mathbf{t}_i, \mathbf{R}_i) \in SE(3)$. For each grasp g_i, the VGN also computes a quality score denoted by q_i. Then, among grasp candidates in $\mathcal{G}$ predicted by the VGN, we execute the highest-quality grasp g^* whose position lies within the semantic bounding box $\mathcal{B}_t$ of the target object:

$$g^* = g_k \text{ s.t. } k = \underset{i \in \{1,...,m\}}{\arg\max}\, q_i, \forall q_i > \tau_q \text{ and } \forall \mathbf{t}_i \in \mathcal{B}_t \tag{3.7}$$

where τ_q is the quality threshold and $\mathbf{t}_i \in \mathbb{R}^3$ is the position vector associated with g_i. This spatial constraint ensures grasps are placed on the detected target object rather than on occluders or background surfaces. If no valid grasp satisfying these constraints exists, the system selects the next best view to continue exploration. Similar to Breyer et al. [9], we implement the stability window T to avoid premature grasp execution on noisy or transient detections. Specifically, we require a valid grasp candidate to be detected in at least T views before committing to execution, ensuring the selected grasp is supported by sufficient multi-view evidence.

3.3 Experiments

The proposed method has been validated through simulation and experiments. For both cases, vision inference was performed on a remote server equipped with the RTX 4090 GPU from NVIDIA. The simulation was run on a laptop with an Intel i5 CPU, while real-world experiments were conducted on a desktop PC with an Intel i7-8700T.

3.3.1 Experimental Setup and Performance Metrics

Figure 3.5 shows our experimental platform, which consists of a 7-DoF Franka Emika Panda robotic arm equipped with a parallel-jaw gripper and an eye-in-hand RGB-D camera (Intel RealSense D435i). We evaluate our semantic active grasping approach in both

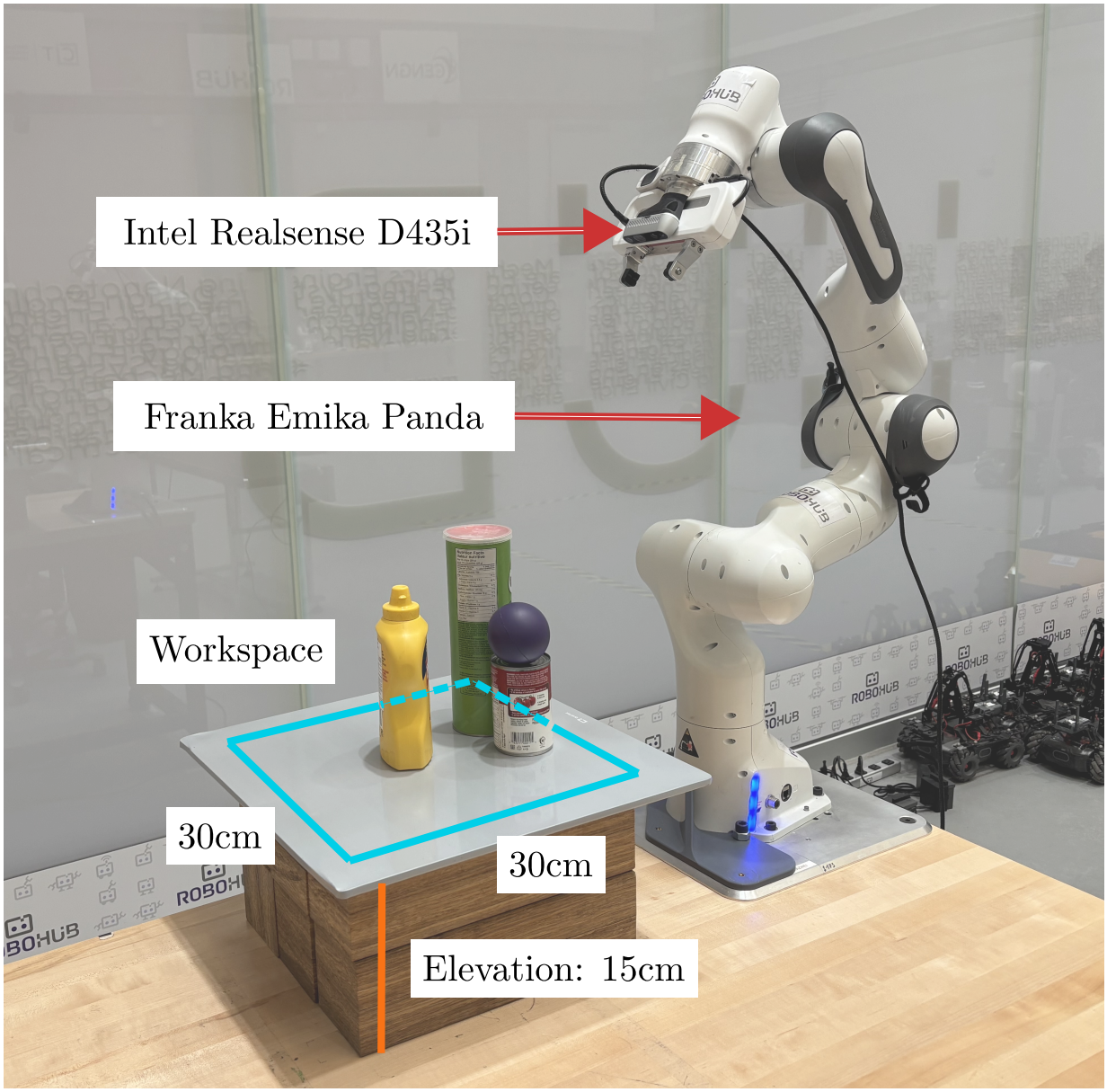


Figure 3.5: Experimental setup. The robot observes a 30 cm × 30 cm workspace containing a target object and occluding objects.



(a) Real-world setup (b) Simulation environment

Figure 3.6: Arrangement of objects for (a) experiment and (b) simulation environments. Top row: scene setup. Bottom row: initial camera view.

simulation and real-world settings (Fig. 3.6), using the PyBullet physics engine [12] for simulation. In simulation, we use ground-truth object instance segmentation masks from the simulator. The semantic segmentation masks are fused with depth images to create a semantic-geometric voxel representation. Test objects are selected from the YCB object set [11]. Objects are placed with random orientations and positions sampled within the workspace, as shown in Fig. 3.6(b). Table 3.1 summarizes the implementation parameters used in all experiments.

We evaluate system performance using the following metrics:

(a) **Success Rate (SR)**: The proportion of trials where the target object was successfully grasped.

(b) **Abort Rate (AR):** The proportion of trials where no valid grasp was found before the maximum view limit.

(c) **Grasp Failure Rate (GFR):** The proportion of attempted grasps that failed during

Table 3.1: Implementation Parameters

Parameter	**Symbol**	**Value**
TSDF size		$0.3^3\,\mathrm{m}^3$
Voxel count per side		40
Hemisphere radius	R_s	0.45 m
Detection confidence threshold	τ_p	0.3
Grasp quality threshold	τ_q	0.7
Semantic IG weight	α	0.5
Stability window	T	5
Max views		15
Linear velocity		5 cm/s

execution.

(d) **Views**: The average number of viewpoints selected per trial.

(e) **Search Time (s):** The average total time spent on viewpoint selection and grasp planning before grasp execution

The proposed semantic-aware NBV approach has been evaluated against several baseline methods, all of which are reimplemented in our framework with the same perception and grasp evaluation pipeline for fair comparison:

(a) **Initial-view**: Single fixed initial viewpoint without view exploration.

(b) **Top-view**: Single fixed overhead viewpoint without view exploration.

(c) **Random**: Selects viewpoints uniformly at random from the view sphere.

(d) **Breyer et al.** [9]: Geometric NBV using rear-side voxel counting as IG on a predefined target bounding box. The target bounding box is provided at the start of each trial.

(e) **GS-NBV** [50]: A semantic-aware NBV baseline that uses Shannon entropy of voxel class distributions as information gain, reimplemented within our perception and grasping pipeline for fair comparison.

GS-NBV requires prior target localization to define the target region for target-driven planning. If the target is detected in the initial view, we initialize this region from the first target detection. If no initial target detection is available, no target region can be established and the trial is counted as a failure.

To evaluate robustness to heavy occlusion, the target object is surrounded by cluttered objects with at least 70% occlusion. We conduct 100 trials per method with varying object poses and clutter configurations.

3.3.2 Simulation Results

Performance Evaluation

Table 3.2 presents simulation results over 100 trials per method in heavily occluded scenes. Our method achieves the highest success rate (84%), outperforming all baselines while maintaining a moderate number of viewpoints and execution time. Single-view baselines struggle under heavy occlusion. The Initial-view baseline fails to detect valid grasps when the target is not visible from the fixed viewpoint, resulting in a high abort rate (70%) as no grasp candidate ever falls within the target region. Top-view achieves 81% success, partly attributed to VGN's preference for top-down grasps [15], but remains limited by incomplete scene reconstruction from a single observation. These results demonstrate the necessity of active viewpoint exploration in cluttered environments. Random exploration improves over Initial-view baseline but requires more viewpoints on average (10.59), as it lacks target awareness when selecting views.

Breyer et al. achieves 74% success rate with fewer viewpoints. However, their method requires a predefined target bounding box provided at the start of each trial, assuming

Table 3.2: Performance of simulation results

Method	**SR (%) ↑**	**AR (%) ↓**	**GFR (%) ↓**	**Views ↓**	**Time (s) ↓**
Initial-view	27	70	10.0 (3/30)	1	1.55 ± 0.11
Top-view	81	10	10.0 (9/90)	1	9.71 ± 0.89
Random	72	14	16.3 (14/86)	10.59 ± 4.41	9.8 ± 3.97
Breyer et al. [9]	74	6	21.3 (20/94)	7.87 ± 4.15	7.54 ± 3.84
GS-NBV [50]	38	56	13.6 (6/44)	3.49 ± 3.49	4.15 ± 3.16
Ours	84	6	10.6 (10/94)	9.74 ± 4.00	9.20 ± 3.43

prior knowledge of object location. In contrast, our method uses semantic segmentation to identify and localize the target during execution, without any prior location information. The semantic TSDF incrementally accumulates detections across views, dynamically constructing a target bounding box that guides exploration toward uncovered regions of the target object. The higher average number of views (9.74 vs. 7.87) reflects the additional exploration needed to first detect and localize the target before a grasp can be computed. Furthermore, semantic grasp filtering ensures candidates are restricted to the target object, avoiding grasps on occluders. GS-NBV achieves 38% success rate, as it requires prior target information to initialize, which is not always obtainable in occluded environments (56% abort rate). In contrast, our method continues exploring even if the target is not initially detected.

Initial-view and Top-view exhibit lower GFR (10.0% each) as a grasp is only attempted when a high-confidence grasp candidate is already visible from the fixed pose. More difficult configurations are aborted rather than attempted (70% and 10% AR, respectively). Our method instead continues exploring through occluded and difficult configurations that these baselines abort, resulting in a broader and more challenging set of attempted trials. Among the active-exploration baselines, our method achieves the lowest GFR (10.6% vs. 16.3%, 21.3%, and 13.6% for Random, Breyer et al., and GS-NBV, respectively), suggesting that the semantic-aware viewpoint selection more effectively resolves occlusions on the target's surfaces before a grasp is attempted.

Ablation Study

Table 3.3: Ablation study

Method	**SR (%) ↑**	**AR (%) ↓**	**GFR (%) ↓**	**Views ↓**	**Time (s) ↓**
Ours ($\alpha = 0.0$)	80	9	12.1 (11/91)	9.90 ± 4.13	9.80 ± 3.96
Ours ($\alpha = 0.5$)	**84**	**6**	**10.6 (10/94)**	9.74 ± 4.00	9.49 ± 3.67
Ours ($\alpha = 0.9$)	82	**6**	12.8 (12/94)	9.64 ± 4.05	9.56 ± 3.89
Ours ($\alpha = 1.0$)	79	7	15.1 (14/93)	**9.58 ± 4.02**	**9.28 ± 3.69**
Ours ($T = 1$)	76	7	18.3 (17/93)	**6.00 ± 4.73**	**5.85 ± 4.21**
Ours ($T = 5$)	**84**	**6**	**10.6 (10/94)**	9.74 ± 4.00	9.20 ± 3.43

Table 3.3 presents the ablation study to investigate the effects of the semantic IG weight α and the stability window T. The stability window T has a clear impact on performance. A

smaller window ($T = 1$) allows immediate execution after a single detection, which reduces search time and requires fewer views but increases grasp failure rate (18.3% vs. 10.6% at $T = 5$). Increasing to $T = 5$ improves success rate and reduces grasp failure by ensuring grasp candidates are confirmed across multiple views before execution.

The semantic IG weight α controls the balance between geometric and semantic information gains. Increasing α from 0 to 0.5 improves success rate from 80% to 84% and reduces grasp failure rate from 12.1% to 10.6%, suggesting that biasing viewpoint selection toward target object regions helps reveal occluded surfaces once the object is detected. However, setting $\alpha = 1.0$ (pure semantic IG) degrades performance on both metrics (79% SR, 15.1% GFR), as the geometric term provides a necessary exploration signal when semantic detections are sparse in early views. Based on these results, we set $\alpha = 0.5$ and $T = 5$ for all experiments.

Semantic Update Strategy Analysis

To illustrate the robustness of the proposed confidence-weighted accumulation strategy against early false positives, Table 3.4 compares the per-class score evolution across three update strategies under a representative scenario: a false positive detection at $p = 0.9$ followed by correct detections at $p = 0.6$. The proposed method correctly recovers and assigns the target class by $t = 3$, as the accumulated correct detections outweigh the single

Table 3.4: Per-class score evolution under a false positive at $p = 0.9$ followed by correct detections at $p = 0.6$. T and F denote the target and false positive classes, respectively. Bold indicates first correct target class assignment.

View	**Ours**		**Pairwise [10]**	**Bayesian [40]**	
	$\bar{p}_t(v, \mathrm{F})$	$\bar{p}_t(v, \mathrm{T})$	$p_t(v)$	$P(\mathrm{F})$	$P(\mathrm{T})$
$t = 1$	0.90	0.00	0.90	0.90	0.10
$t = 2$	0.45	0.30	0.81	0.857	0.143
$t = 3$	0.30	**0.40**	0.73	0.800	0.200
$t = 4$	0.23	0.45	0.66	0.727	0.273
$t = 5$	0.18	0.48	0.59	0.632	0.368
$t = 6$	0.15	0.50	**0.54**	0.513	0.487
$t = 7$	0.13	0.52	0.57	0.419	**0.581**
Recovery	**t = 3**		$t = 6$	$t = 7$	

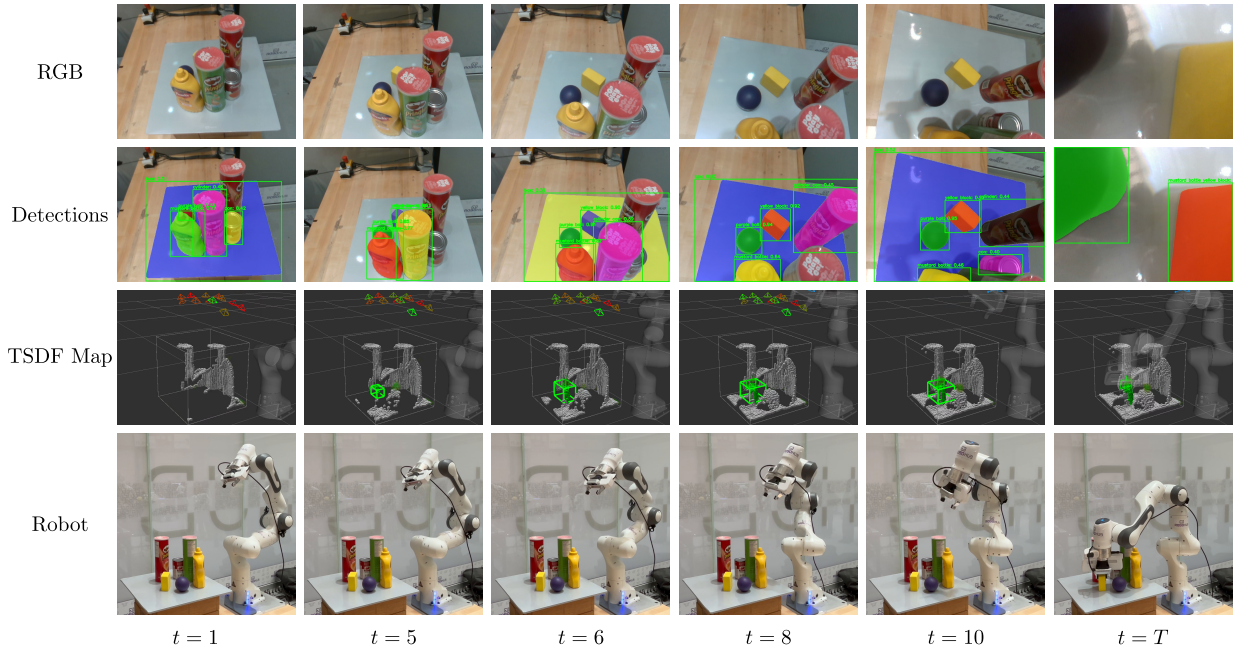


Figure 3.7: Sequence of RGB observations, semantic detections, TSDF map updates, and robot motion across multiple viewpoints during semantic-aware NBV planning. The last column ($t = t_f$) shows the final grasp execution.

false positive. The pairwise update strategy of Burusa et al. [10] relies on a fixed 10% mismatch penalty to overcome the false positive, requiring until $t = 6$ to recover. The recursive Bayesian formulation [40] accumulates evidence multiplicatively, so an initial high-confidence false positive establishes a strong prior that subsequent lower-confidence correct detections cannot quickly overcome, requiring until $t = 7$ to recover.

Furthermore, if an additional high-confidence false positive occurs after recovery, the pairwise strategy of Burusa et al. flips back to the false class with a single observation, and the Bayesian formulation accumulates an additional large likelihood factor, whereas the proposed method absorbs the false positive into the full observation history, preserving the accumulated evidence from correct detections.

While high-confidence false positives were rare in our YCB experiments due to the reliability of GroundingDINO on common objects, this update strategy provides robustness for more challenging open-vocabulary settings with novel objects or visually ambiguous distractors.

Table 3.5: Performance of experimental results

Method	**SR ↑**	**AR ↓**	**GFR ↓**	**Views ↓**	**Time (s) ↓**
No initial observation					
Initial-view	0/10	10/10	–	1.00	—
Top-view	8/10	0/10	2/10	1.00	—
Random	7/10	0/10	3/10	12.40 ± 1.07	12.45 ± 2.38
Breyer et al. [9]	9/10	0/10	1/10	9.50 ± 0.71	9.80 ± 1.35
GSNBV [50]	0/10	10/10	–	1.00	—
Ours	10/10	0/10	0/10	10.60 ± 1.26	10.45 ± 1.00
Partial observation					
Initial-view	3/10	4/10	3/6	1.00	—
Top-view	8/10	0/10	2/10	1.00	—
Random	5/10	0/10	5/10	5.70 ± 1.49	5.51 ± 1.22
Breyer et al. [9]	9/10	0/10	1/10	6.36 ± 1.75	6.55 ± 1.83
GS-NBV [50]	5/10	3/10	2/7	4.30 ± 2.50	5.47 ± 2.52
Ours	9/10	0/10	1/10	6.82 ± 1.94	6.31 ± 1.75

3.3.3 Experimental Results

Figure 3.7 shows snapshots of a representative robot arm trajectory from the experimental trial. As shown in the sequence of TSDF maps, the target object was not visible from the initial view when $t = 1$ but progressively detected as the semantic-aware NBV algorithm guides the end-effector for better views in an optimal way. Table 3.5 presents results from 10 trials with varying occlusion and clutter configurations under two conditions: no initial observation and partial observation. Under no initial observation, the initial-view baseline and GS-NBV fail entirely (0/10 SR) as both require an initial target detection. Top-view achieves reasonable success by capturing a broader scene perspective, while Random NBV reaches a moderate success rate with higher grasp failure and more views required. Our method achieves the highest success rate (10/10 SR), compared to Breyer et al. (9/10 SR) though more views are required to first detect and localize the target before a grasp can be computed (10.60 vs. 9.50). Under partial observation, our method matches Breyer et al. (9/10) with similar views required. GS-NBV achieves 5/10 SR, as the target may not always be detectable even from a partially visible initial view.

In contrast to GS-NBV [50], which requires prior target localization to initialize plan-

ning, our method continues exploring until the target is detected. Unlike Breyer et al. [9], which requires a predefined target bounding box, our approach uses object detection to identify and localize target objects, making it applicable to cluttered scenarios where object locations are unknown. These results validate that our approach generalizes to real-world settings while relaxing the assumption of known object location.

3.4 Conclusion

We presented a semantic active grasping approach that integrates semantic segmentation with NBV planning to grasp target objects in heavily occluded environments without prior knowledge of object location. By maintaining a semantic TSDF that accumulates semantic detections across views, our method guides viewpoint selection toward regions most likely to reveal graspable surfaces of the target object. Simulation and experiments demonstrate that our approach outperforms single-view, random, and geometric NBV baselines in simulation, and matches geometric NBV performance in real-world trials, without requiring predefined object location information.

A current limitation is the dependence on accurate semantic segmentation, as failures in the visual grounding pipeline directly impact grasp success. Future work will investigate the approach in dynamic environments with moving obstacles, and extend to mobile manipulator platforms where target objects may be distributed across larger, unstructured spaces.

Chapter 4

Learning Next-Best-View Policy from Vision-Language Model Preferences

4.1 Introduction

Chapter 3 proposed a semantic-aware NBV framework that integrates geometric and semantic information gain for active grasping under occlusion. More broadly, existing NBV methods for grasping typically focus on one or two signals at a time: geometric approaches reduce volumetric uncertainty [9], semantic-aware methods incorporate object labels into occlusion reasoning [15], and learned approaches predict future grasp quality or affordance from the current observation [38, 60]. While each captures a useful signal, these approaches do not jointly reason about visibility, object semantics, and grasp feasibility. This is needed to determine whether the next selected view is actually useful for grasping a specific target. Joint reasoning over these factors is necessary to determine whether a candidate view will actually expose graspable regions of the target, and is difficult to capture with handcrafted rules or models trained on individual signals. While recent works leverage VLMs for semantic viewpoint reasoning [56], these methods require rendering candidate view images and querying the VLM at every inference step, introducing significant computational overhead that limits real-time deployment on robotic systems.

In this chapter, we propose a lightweight view scoring network that distills VLM viewpoint preferences offline for next-best-view selection, enabling semantic-aware NBV without querying a VLM at runtime. During an offline data collection stage, a VLM evaluates candidate camera views for grasping a target object, reasoning about target visibility, occlusion, and grasp feasibility from 2D images, and ranks them from most to least informative.

These VLM rankings are then used as supervision to train a lightweight network that predicts the best viewpoint from the current observation and candidate pose parameters in a single forward pass (see Fig. 4.1). We evaluate the approach in simulation and show that the distilled model approximates VLM viewpoint preferences while being computationally efficient, making it practical for robotic grasping.

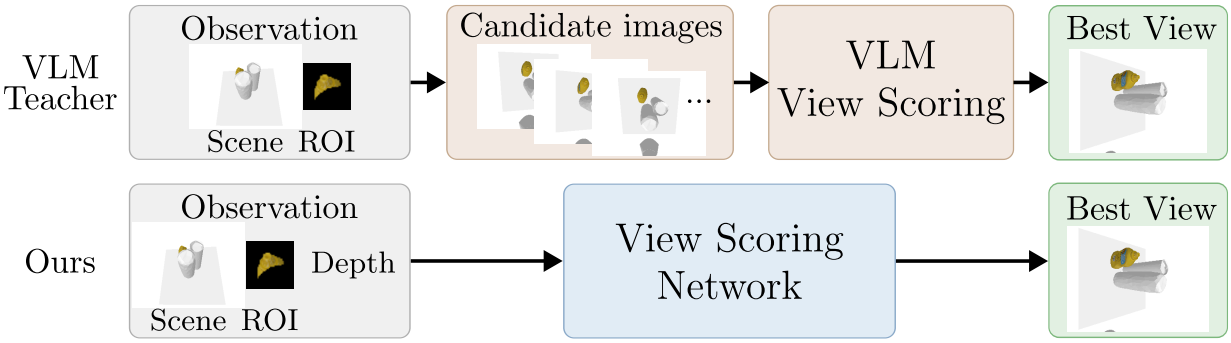


Figure 4.1: During data collection, the VLM teacher ranks candidate views based on target visibility. At inference, our distilled ViewScorer selects the next-best view from a single observation in a single forward pass.

The remainder of this chapter is organized as follows. Section 4.2 describes the proposed methodology. Section 4.3 presents experimental results. Section 4.4 concludes the chapter.

4.2 Methodology

The proposed viewpoint scoring policy for active grasping in cluttered scenes is trained from VLM preferences over candidate viewpoints augmented with predicted grasp poses. By distilling VLM grasp reasoning offline, our system selects informative viewpoints at inference without VLM queries or candidate rendering.

Our framework consists of three modules: (i) a target grounding module that produces a target mask from the language query and current RGB image using Grounding DINO [35] and SAM [29], (ii) a grasp proposal module that fuses RGB-D observations into a TSDF volume and queries VGN [8] for 6-DoF grasp candidates, and (iii) a viewpoint scoring policy that selects the next informative viewpoint. The perception and grasp modules are pre-trained and fixed.

We train the policy in three stages, illustrated in Fig. 4.1: (1) offline data collection in simulation, where we capture candidate viewpoints across diverse cluttered scenes, (2) VLM-based labeling, where the VLM ranks candidates augmented with predicted grasp poses, and (3) training a lightweight viewpoint scorer on these rankings.

4.2.1 Data Collection

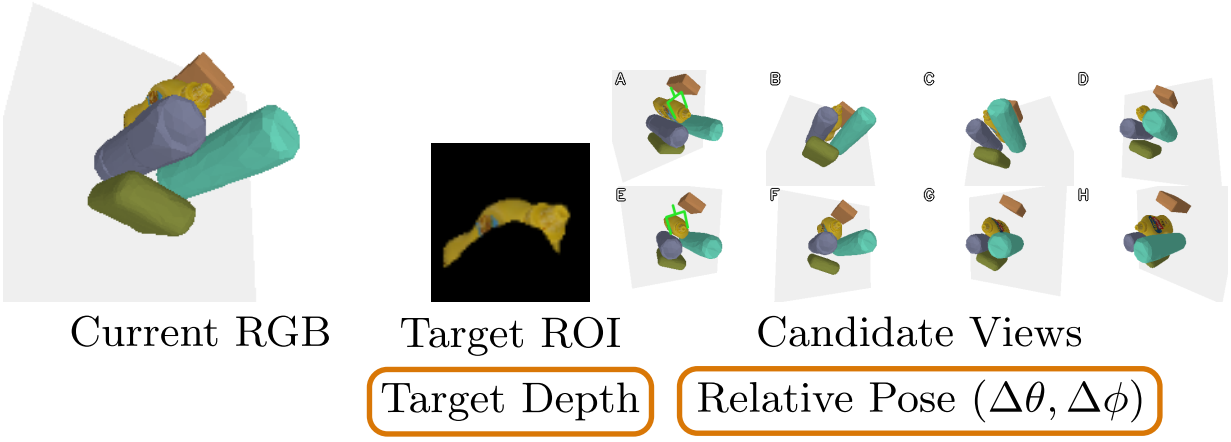


Figure 4.2: Stage 1: Data Collection. The robot observes cluttered scenes from a randomized initial viewpoint, rendering the current RGB image, target ROI, and candidate view images.

As shown in Fig. 4.2, we collect training data in PyBullet [12] simulation using cluttered tabletop scenes with partially occluded target objects from the YCB dataset [11]. For each scene, the robot observes the environment from a randomized initial viewpoint. Scenes are filtered to ensure the target object is at least 30% occluded from the initial viewpoint, focusing training on challenging partial occlusion scenarios. We define occlusion ratio as the fraction of the target's projected mask blocked by other scene objects, computed by comparing visible target pixel counts with and without other objects present. From this viewpoint, we render: (i) the current RGB image, (ii) a target region of interest (ROI) crop containing only the target object, and (iii) K candidate RGB images from sampled viewpoints. For each scene, we sample K candidate viewpoints by drawing two random elevation angles (low and high) and four random azimuth angles (one per 90° quadrant), giving diverse coverage of the hemisphere. The depth of the target object is estimated as the mean depth over the visible target pixels from the initial view.



Figure 4.3: Example candidate viewpoint with the predicted grasp proposal overlaid in green, computed by fusing the candidate's depth into a TSDF and querying VGN [8].

For each candidate view, we record the relative pose offset from the current viewpoint and compute a grasp proposal by fusing the candidate's depth alone into a TSDF and querying VGN [8], recording whether a valid grasp is found and its predicted quality score. This isolated computation evaluates each candidate independently of information already available from the current observation. We additionally compute the occlusion ratio of the target object for each candidate. Together, these provide simulator-derived ground truth used to validate VLM teacher labels (Section 4.3.1). The predicted grasp proposal is overlaid on each candidate viewpoint's RGB image (Fig. 4.3), shown to the VLM teacher during labeling.

4.2.2 VLM Labeling

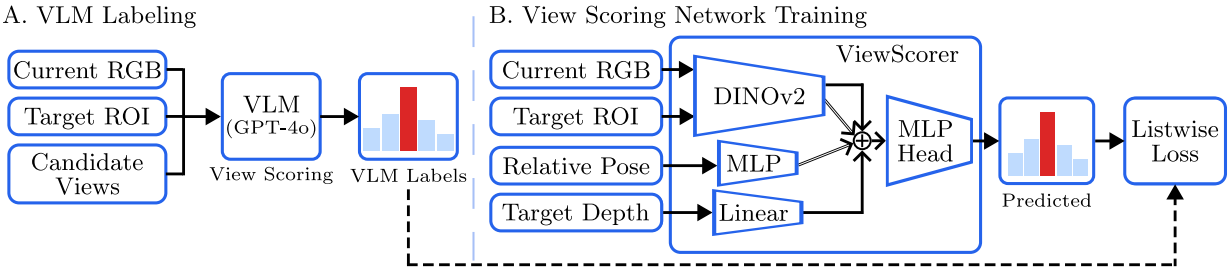


Figure 4.4: Stage 2: VLM Labeling and Network Training. GPT-4o ranks candidate viewpoints based on target visibility, generating soft labels used to train the view scoring network.

Rather than reasoning over 3D volumetric representations, we leverage a VLM to score candidate viewpoints directly from 2D rendered observations, as illustrated in Fig. 4.4.

Specifically, we use GPT-4o [43] for offline annotation. For each scene, the VLM receives the current RGB image, a crop of the target object, and one candidate next-view image overlaid with the predicted grasp pose. The VLM assigns each candidate a score from 0 to 10 based on how well the view reveals the target object for grasping, considering occlusion, visibility, and graspable surface exposure.

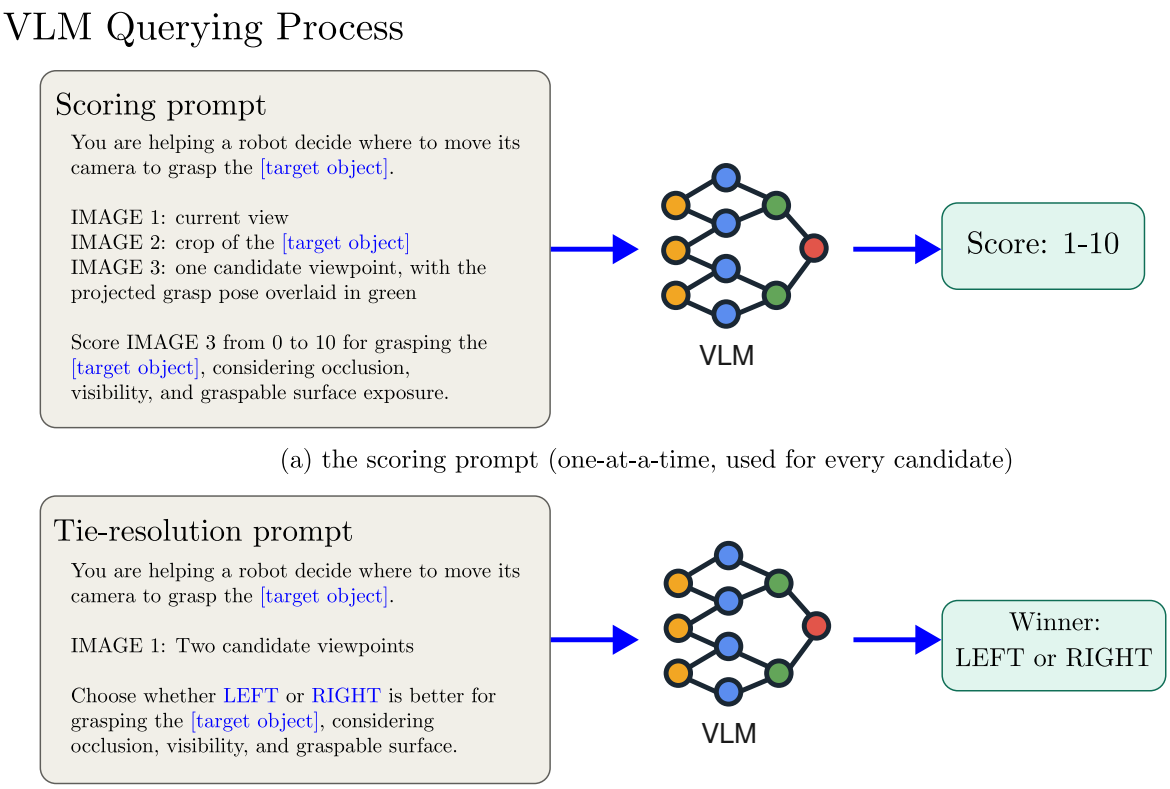
Since VLM scores often contain ties among the highest-scored candidates, we resolve top-score ties using additional pairwise VLM comparisons between tied views (Fig. 4.5). Position bias has been commonly observed in both LLMs [33, 61] and VLMs [52, 1], where the order of presented options affects model preferences. To mitigate this, each pair is queried twice, once in each left-right order. If both queries agree, the winning candidate receives a full vote. If the queries disagree, the comparison is treated as ambiguous and contributes no relative preference between the two candidates.

Pairwise vote totals are used to redistribute probability mass among tied candidates when constructing soft training labels, giving candidates with more pairwise wins proportionally higher weight. Then these scores are converted to soft labels for training.

Raw VLM scores are often closely clustered, so a direct softmax over scores yields near-uniform labels, while naive temperature sharpening can overstate confidence in small, potentially noisy score differences. Sharpening may also disregard genuine ties, concentrating probability mass on a single candidate even when multiple candidates share the same highest score. We instead convert scores to soft labels using a margin-based scheme. Candidates are grouped into a top group (the highest-scoring candidates), a near-top group (scores close to the highest scores), and the remainder. Each group is allocated a fixed share of probability mass (e.g. 0.70, 0.27, 0.03), with pairwise vote totals used to break ties within the top group. Each near-top candidate's share is capped below the smallest per-candidate top-group share, with any excess redistributed to the top group. This ensures no near-top candidate receives more mass than an individual top-scoring candidate. This method preserves confident labels when the VLM shows a genuine, large preference while remaining appropriately uncertain when scores are closely tied.

4.2.3 View Scoring Network

We train PrefNBV, a lightweight network, to approximate the VLM's viewpoint preferences, enabling real-time next-best-view selection without querying the VLM at inference. We parameterize a camera viewpoint as $\mathbf{p} = (\theta, \phi, r, \mathbf{d})$, where θ and ϕ are the elevation and azimuth on a hemisphere centered at the workspace centroid $\mathbf{c}$, r is the radius, and $\mathbf{d}$ is the camera's look-at direction. In our setup, all candidate viewpoints share a fixed



(a) the scoring prompt (one-at-a-time, used for every candidate)

(b) the pairwise tie-resolution prompt (used only when scores tie)

Figure 4.5: Overview of the VLM labeling prompts. (a) The scoring prompt is applied to every candidate view, producing a score from 0 to 10. (b) When multiple candidates share the highest score, the pairwise tie-resolution prompt selects a winner between the tied views.

radius $r = r_0$ and are oriented to always look at $\mathbf{c}$, so $\mathbf{d}$ is fully constrained. A viewpoint is therefore fully specified by its angular coordinates (θ, ϕ), which we use to encode each pose. At step t, we define a candidate's pose offset relative to the current viewpoint (θ_t, ϕ_t) as

$$(\Delta\theta_i, \Delta\phi_i) = (\theta_i - \theta_t, \phi_i - \phi_t). \tag{4.1}$$

The model takes as input the current RGB image and ROI crop, candidate pose offsets $(\Delta\theta_i, \Delta\phi_i)$, and estimated target depth. The RGB image and ROI crop are encoded via a shared, frozen DINOv2 backbone [44], pose offsets via sinusoidal features and an MLP, and depth via a separate MLP. We compare DINOv2 against several alternative backbones in Section 4.3.4, where it achieves the strongest visual context sensitivity among tested

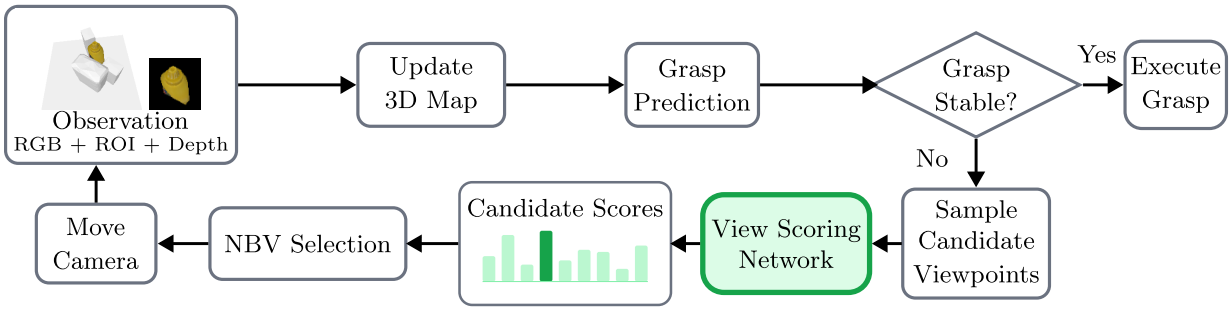


Figure 4.6: Stage 3: Deployment. The trained view scoring network scores candidate viewpoints from the current observation and pose parameters, selecting the next best view without VLM queries.

configurations. These features are concatenated and passed through an MLP head to produce a score for each candidate.

4.2.4 Training

Candidate-view images and per-candidate grasp proposals are available to the VLM teacher during data collection but unavailable at inference without moving the camera. We treat both as privileged information. The VLM rankings, converted to soft labels $\tilde{y}_i$, encode the teacher's viewpoint preferences. The student is trained to reproduce these rankings from candidate pose offsets and the current observation, without access to candidate images.

The network is trained with a listwise softmax cross-entropy loss over VLM-derived soft labels:

$$\mathcal{L} = -\sum_{i=1}^{K} \tilde{y}_i \log \frac{\exp(f_\theta(\xi_i))}{\sum_{j=1}^{K} \exp(f_\theta(\xi_j))} \tag{4.2}$$

where $f_\theta(\xi_i)$ is the predicted score for candidate viewpoint ξ_i, $\tilde{y}_i$ is the soft label for view i, and K is the number of candidates. This listwise formulation captures relative preferences across all candidates simultaneously. We freeze the pretrained backbone and train only the pose, depth, and fusion head parameters.

4.2.5 Inference

At each step t, candidate viewpoints $\xi \in \mathcal{X}_t$ are generated deterministically on a fixed view hemisphere using their azimuth and elevation angles, a fixed radius r_0, and a look-

at constraint toward the workspace center. Candidates are filtered by inverse-kinematic reachability prior to scoring. At inference, as shown in Fig. 4.6, PrefNBV scores each reachable candidate using the current observation and the candidate's relative angular offset $(\Delta\theta, \Delta\phi)$:

$$\xi_t^* = \arg\max_{\xi \in \mathcal{X}_t^{\text{reach}}} f_\theta(I_t, R_t, d_t, \Delta\theta, \Delta\phi), \tag{4.3}$$

where I_t is the current RGB image, R_t is the target ROI crop, d_t is the target depth feature, and $\mathcal{X}_t^{\text{reach}} \subseteq \mathcal{X}_t$ is the reachable candidate subset. The robot moves to ξ_t^*, and the TSDF is updated with the new observation, accumulating across all views observed so far in the episode. This process repeats until a stopping criterion is met, after which VGN queries the accumulated TSDF to produce the final grasp.

4.3 Experiments

4.3.1 Teacher Label Validity

We validate VLM teacher labels against simulator-computed ground truth: occlusion ratio and grasp quality of the candidate selected by each method. Occlusion ratio is computed to measure how much each method improves visibility of the target relative to the initial view. Grasp quality is the confidence score of the best grasp proposal returned by VGN [8], computed from each candidate's view alone (Section 4.2.1) to isolate its individual contribution. This measures whether the selected view yields a viable grasp of the target, independent of visual occlusion.

Table 4.1: Validity of VLM teacher labels against simulator-computed ground truth.

Method	**Occlusion Ratio (%) ↓**	**Grasp Quality ↑**
Initial view	50.7	0.802
Random candidate	13.0	0.908
Semantic IG (single-step)	4.3	0.922
VLM (no grasp overlay)	**2.6**	0.926
VLM (grasp overlay, ours)	3.2	**0.932**
Least-occluded (oracle)	**0.4**	0.939
Best-grasp (oracle)	4.0	**0.960**

Across methods, the VLM teacher substantially reduces occlusion and improves grasp quality relative to random and semantic IG baselines (Table 4.1). We additionally report two oracle baselines with privileged access to ground-truth occlusion and grasp quality (least-occluded and best-grasp selection), representing the upper bound achievable on this candidate set. Both VLM conditions approach these oracle bounds (0.4% occlusion, 0.960 grasp quality). Also, they outperform random selection by a wide margin (13.0% occlusion) and improve on Semantic IG (4.3% occlusion, 0.922 grasp quality) on both metrics, despite IG's access to volumetric occupancy information the VLM does not have. The two VLM conditions show a small but consistent trade-off: the no-overlay condition achieves lower occlusion (2.6% vs. 3.2%), while the grasp-overlay condition achieves higher grasp quality (0.932 vs. 0.926), consistent with each condition's available information. Grasp overlays provide an explicit candidate grasp proposal, shifting VLM judgment toward grasp-relevant considerations, while the no-overlay condition relies purely on visual occlusion and object visibility.

4.3.2 Position Bias in VLM Labeling

Table 4.2: Positional bias in grid-based and pairwise VLM querying.

Diagnostic	**Value**
Grid: most-favored slot win rate (vs. 12.5% uniform)	46.0%
Pairwise: right-image win rate (vs. 50% uniform)	70.7%

As discussed in Section 4.2.2, position bias affects model preferences independent of content. We measured this at multiple stages of our labeling pipeline (Table 4.2).

In an initial grid-based approach (Fig. 4.7), one grid slot was selected as best in 46.0% of scenes (vs. 12.5% uniform). In pairwise tie-resolution comparisons, the right-hand image was preferred in 70.7% of queries. To correct for this, each pair is queried with the original and reversed order. 44.6% of comparisons disagreed on the winner, contributing no relative preference between the pair.

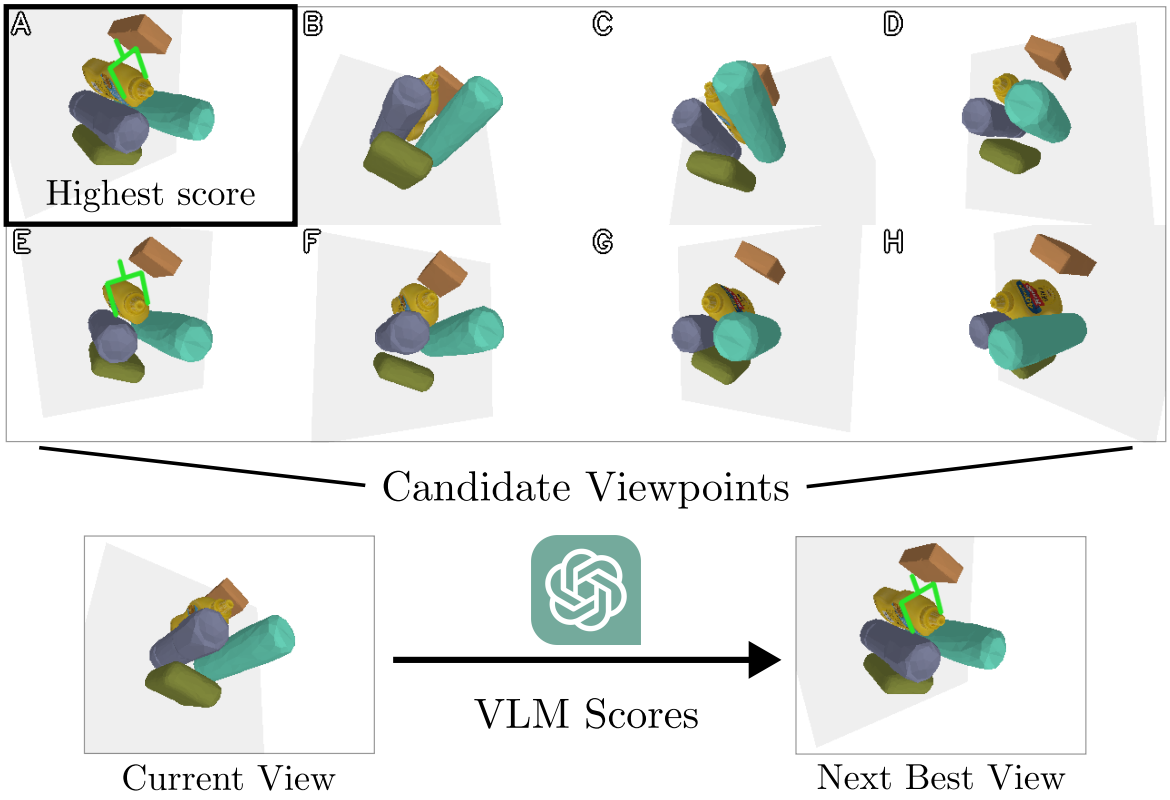


Figure 4.7: The initial grid-based labeling approach, in which the VLM selects the highest-scoring candidate from a grid of viewpoints shown together. This approach was later replaced by one-at-a-time candidate scoring due to the position bias observed here.

4.3.3 Backbone Comparison and the Pose-Only Baseline

Table 4.3: Ranking accuracy of vision-conditioned models versus a pose-only baseline.

Model	Top1	TopGroup	Top3
Pose-only	0.227 ± 0.042	0.547 ± 0.031	0.533 ± 0.042
MobileNetV4-small	0.180 ± 0.035	0.547 ± 0.012	0.473 ± 0.061
MobileNetV4-medium	0.273 ± 0.061	0.513 ± 0.031	0.473 ± 0.076
DINOv2-small	0.247 ± 0.031	0.540 ± 0.035	0.533 ± 0.061
DINOv2-base	0.200 ± 0.028	0.520 ± 0.028	0.480 ± 0.028

We compare PrefNBV trained with several candidate visual backbones against a pose-only baseline that receives no visual input at all to test how much of the distilled ranking signal is explained by pose alone. Backbone size ranges from 3.19M parameters (MobileNetV4-small) to 86.29M (DINOv2-base). All models are trained on the same 500 scene dataset with identical labels, loss, and target construction, varying only the image encoder. Table 4.3 reports Top1, TopGroup, and Top3 accuracy over 3 random seeds. TopGroup credits a prediction as correct if it matches any candidate within the group of scores tied for maximum in a given scene (Section 4.2.2), rather than requiring an exact match to a single winner.

Among vision-conditioned backbones, DINOv2-small achieves the highest TopGroup and Top3 accuracy, and leads MobileNetV4-small on Top1, though MobileNetV4-medium achieves slightly higher Top1 despite lower TopGroup and Top3 accuracy. The pose-only baseline matches DINOv2-small on both TopGroup and Top3 (0.547 vs. 0.540; 0.533 vs. 0.533). Its strong performance indicates that candidate pose alone carries substantial predictive signal for the VLM's view preferences in this setting.

4.3.4 Context-Reliance Ablation

Table 4.4: Context sensitivity under input shuffling

Model	Scene ↓	ROI ↓
DINOv2-small (frozen)	**0.440**	**0.540**
MobileNetV4-medium (frozen)	1.000	1.000
MobileNetV4-small (frozen)	0.960	0.940

To test whether the vision-conditioned backbones genuinely use visual content, we replace each input with the corresponding input from a randomly paired scene and measure how often the model's top-ranked candidate changes as a result. If a model's top prediction rarely changes under this test, it is not meaningfully using that input's content.

Table 4.4 shows a contrast between backbones. Both MobileNetV4 variants rarely change their prediction when inputs are replaced (0–6%), indicating these models learned to disregard visual input despite having access to it during training. DINOv2-small changes its prediction in the majority of scenes under scene shuffling (56.0%) and nearly half under ROI shuffling (46.0%), indicating it does extract and condition on visual content.

4.3.5 Real-World System Demonstration

To test the pipeline on physical hardware, we deployed the trained model on a 7-DoF Franka Emika Panda robotic arm with an eye-in-hand Intel RealSense D435i camera, using a cluttered tabletop scene. Figure 4.8 shows the scene from both the external and wrist-mounted cameras, before and after view selection. After the policy is run, the selected view clearly reveals the target object, enabling a successful grasp.

4.4 Conclusion

We presented PrefNBV, a next-best-view approach that distills GPT-4o viewpoint preferences into a lightweight view scoring network, replacing the ray casting required by the semantic NBV framework of Chapter 3 with a single network forward pass. Building a reliable distillation pipeline required addressing VLM position bias in candidate labeling, designing a margin-based soft-label scheme that avoids both overly flat and overly confident training targets, and validating VLM teacher labels against simulator-computed occlusion and grasp quality. Our final model, DINOv2-small, was selected based on evidence that it genuinely conditions its predictions on visual context. Directions for future work are discussed in Chapter 5.

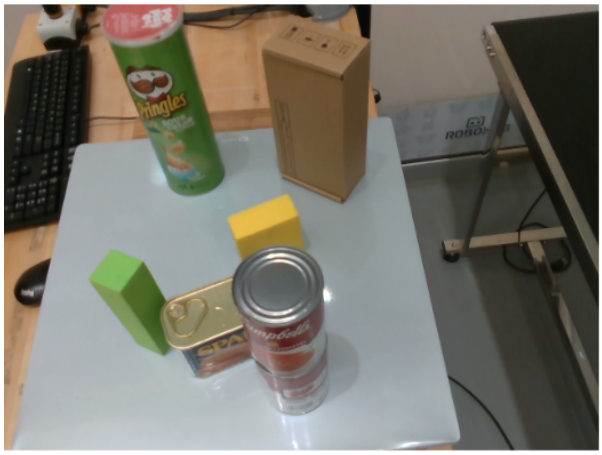
A limitation of our approach is that the VLM teacher provides offline preferences rather than task-grounded robot feedback. Although grasp overlays and pairwise tie resolution reduce labeling ambiguity, the teacher remains an imperfect proxy for physical grasp success. Future work should combine VLM-derived preferences with online robot outcomes.



(a) Initial View



(b) Next Best View



(c) Initial View (Wrist Camera)



(d) Next Best View (Wrist Camera)

Figure 4.8: Real-world pipeline verification. (a, b) External camera views of the scene before and after the NBV policy selects a new viewpoint. (c, d) Corresponding wrist-mounted camera views. The selected view reveals the target object, enabling a successful grasp.

Chapter 5

Conclusion and Future Work

This thesis addressed the problem of robotic grasping under occlusion through two complementary approaches to next-best-view (NBV) planning for active perception.

5.1 Conclusion

The first approach, presented in Chapter 3, proposed a semantic-aware NBV framework that integrates geometric and semantic information gain for active grasping in cluttered scenes. By maintaining a unified volumetric representation that accumulates semantic detections across viewpoints, the framework autonomously explores, localizes, and grasps a target object under heavy occlusion without any prior knowledge of its location. The approach achieved 84% success rate in simulation and 10/10 and 9/10 successful grasps under full and partial occlusion conditions respectively in real-world experiments, outperforming geometric and random NBV baselines while requiring no predefined object location.

The second approach, presented in Chapter 4, proposed a learned NBV policy that distills VLM viewpoint preferences offline into a lightweight view scoring network. By moving VLM reasoning offline, the approach enables view selection at inference without querying a VLM or performing ray casting, replacing per-candidate ray casting with a single network forward pass. Chapter 4 evaluates this approach against information-gain-based NBV and analyzes the extent to which the learned scorer relies on visual context versus candidate geometry.

Together, these two approaches demonstrate that semantic volumetric reasoning and learned VLM preferences are effective and complementary strategies for active perception

in robotic grasping. The former provides interpretable geometric-semantic exploration without requiring training data, and the latter offers a learned alternative for view selection that operates directly from 2D observations without ray casting.

5.2 Future Work

For the semantic-aware NBV framework presented in Chapter 3, future work will investigate the approach in dynamic environments with moving obstacles, and extend to mobile manipulator platforms where target objects may be distributed across larger, unstructured spaces.

For the VLM-distilled NBV approach presented in Chapter 4, the current view scoring network selects a single next view without considering previously visited viewpoints. Future work includes extending to multi-step view planning by marking previously visited viewpoints directly in the candidate view grid presented to the VLM, enabling the VLM to select the next best view from unvisited candidates only. An alternative direction is to move beyond scoring a fixed set of candidate viewpoints toward directly regressing a continuous next-view pose, which would remove the dependence on any particular discrete candidate sampling scheme .

A promising broader direction is to extend the proposed active perception framework from viewpoint selection alone toward interactive manipulation in cluttered environments. In the current framework, the robot improves target observability primarily by moving the camera. However, in more challenging scenarios, the target may remain occluded or physically inaccessible even after multiple viewpoint changes, requiring the robot to manipulate surrounding objects before attempting the final grasp. A vision-language model could serve as a high-level reasoning module for this extension by taking a natural-language instruction, the current RGB-D observation, and the available manipulation skills to determine the most appropriate next action. For example, the model could infer that an apple occluding a target banana should first be moved aside before grasping the target. The selected action could then be executed using the existing perception, grasp detection, and motion-planning modules, enabling closed-loop reasoning over viewpoint motion, obstacle removal, and grasp execution. Key challenges include grounding vision-language model outputs into geometrically feasible robot actions, distinguishing true physical obstruction from image-space occlusion, and maintaining reliable long-horizon execution through replanning and recovery.

References

[1] Md. Atabuzzaman, Ali Asgarov, and Chris Thomas. Benchmarking and mitigating MCQA selection bias of large vision-language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33548–33562, Suzhou, China, November 2025. Association for Computational Linguistics.

[2] Hao-Shu Fang and Chenxi Wang and Hongjie Fang and Minghao Gou and Jirong Liu and Hengxu Yan and Wenhai Liu and Yichen Xie and Cewu Lu. AnyGrasp: Robust and Efficient Grasp Perception in Spatial and Temporal Domains. *arXiv preprint arXiv:2212.08333*, 2023.

[3] Justin Yu and Yide Shentu and Di Wu and Pieter Abbeel and Ken Goldberg and Philipp Wu. EgoMI: Learning Active Vision and Whole-Body Manipulation from Egocentric Human Demonstrations. *arXiv preprint arXiv:2511.00153*, 2026.

[4] Qiyuan Zeng and Chengmeng Li and Jude St. John and Zhongyi Zhou and Junjie Wen and Guorui Feng and Yichen Zhu and Yi Xu. ActiveUMI: Robotic Manipulation with Active Perception from Robot-Free Human Demonstrations. *arXiv preprint arXiv:2510.01607*, 2025.

[5] Yongxi Huang and Zhuohang Wang and Wenjing Tang and Cewu Lu and Panpan Cai. I-Perceive: A Foundation Model for Active Perception with Language Instructions. *arXiv preprint arXiv:2603.00600*, 2026.

[6] Zhenyang Liu and Yongchong Gu and Yikai Wang and Xiangyang Xue and Yanwei Fu. ActiveVLA: Injecting Active Perception into Vision-Language-Action Models for Precise 3D Robotic Manipulation. *arXiv preprint arXiv:2601.08325*, 2026.

[7] Jeannette Bohg, Antonio Morales, Tamim Asfour, and Danica Kragic. Data-driven grasp synthesis—a survey. *IEEE Transactions on Robotics*, 30(2):289–309, 2014.

[8] Michel Breyer, Jen Jen Chung, Lionel Ott, Roland Siegwart, and Juan Nieto. Volumetric grasping network: Real-time 6 dof grasp detection in clutter. In *Proceedings of the 2020 Conference on Robot Learning (CoRL)*, pages 1602–1611, 2021.

[9] Michel Breyer, Lionel Ott, Roland Siegwart, and Jen Jen Chung. Closed-loop next-best-view planning for target-driven grasping. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1411–1416, 2022.

[10] Akshay K Burusa, Joost Scholten, Xin Wang, David Rapado-Rincón, Eldert J van Henten, and Gert Kootstra. Semantics-aware next-best-view planning for efficient search and detection of task-relevant plant parts. *Biosystems Engineering*, 248:1–14, 2024.

[11] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M. Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *IEEE International Conference on Advanced Robotics (ICAR)*, pages 510–517, 2015.

[12] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. http://pybullet.org, 2016.

[13] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, pages 303–312, 1996.

[14] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Seminal graphics papers: pushing the boundaries, volume 2*, pages 121–130. ACM, 2023.

[15] Yixiang Dai, Siang Chen, Kaiqin Yang, Dingchang Hu, Pengwei Xie, Guosheng Li, Yuan Shen, and Guijin Wang. Active-perceptive language-oriented grasp policy for heavily cluttered scenes. *IEEE Robotics and Automation Letters*, 10(11):11094–11101, 2025.

[16] Nils Dengler, Jesper Mücke, Rohit Menon, and Maren Bennewitz. Efficient manipulation-enhanced semantic mapping with uncertainty-informed action selection. In *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pages 768–775, 2025.

[17] Hongyang Du, Zongxia Li, Dawei Liu, Runhao Li, Haoyuan Song, Qingyu Zhang, Yubo Wang, Jingcheng Ni, Shihang Gui, Congchao Dong, and Tao Hu. A cookbook of 3d

vision: Data, learning paradigms, and application. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 8376–8390, June 2026.

[18] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11441–11450, 2020.

[19] Udita Ghosh, Dripta S. Raychaudhuri, Jiachen Li, Konstantinos Karydis, and Amit Roy-Chowdhury. Preference vlm: Leveraging vlms for scalable preference-based reinforcement learning, 2025. arXiv preprint arXiv:2502.01616.

[20] Michel Goossens, Frank Mittelbach, and Alexander Samarin. *The LaTeX Companion*. Addison-Wesley, Reading, Massachusetts, 1994.

[21] Margarita Grinvald, Fadri Furrer, Tonci Novkovic, Jen Jen Chung, Cesar Cadena, Roland Siegwart, and Juan Nieto. Volumetric instance-aware semantic mapping and 3d object discovery. *IEEE Robotics and Automation Letters*, 4(3):3037–3044, 2019.

[22] Huy Ha, Pete Florence, and Shuran Song. Scaling up and distilling down: Language-guided robot skill acquisition. In *Proceedings of the 2023 Conference on Robot Learning*, 2023.

[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[24] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network, 2015.

[25] Stefan Isler, Reza Sabzevari, Jeffrey Delmerico, and Davide Scaramuzza. An information gain formulation for active volumetric 3d reconstruction. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3477–3484, 2016.

[26] Zhenyu Jiang, Yifeng Zhu, Maxwell Svetlik, Kuan Fang, and Yuke Zhu. Synergies between affordance and geometry: 6-dof grasp detection via implicit representations. In *Proceedings of Robotics: Science and Systems (RSS)*, 2021.

[27] Tobias Jülg, Wolfram Burgard, and Florian Walter. Refined Policy Distillation: From VLA generalists to RL experts. In *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 11677–11684, 2025.

[28] Sebastian A. Kay, Simon Julier, and Vijay M. Pawar. Semantically informed next best view planning for autonomous aerial 3d reconstruction. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3125–3130, 2021.

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3992–4003, 2023.

[30] Donald Knuth. *The TEXbook*. Addison-Wesley, Reading, Massachusetts, 1986.

[31] Cagatay Koc and Sanem Sariel. Object-aware interactive perception for tabletop scene exploration. *Robotics and Autonomous Systems*, 175:104674, 2024.

[32] Leslie Lamport. *LATEX — A Document Preparation System*. Addison-Wesley, Reading, Massachusetts, second edition, 1994.

[33] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 26874–26901. PMLR, 21–27 Jul 2024.

[34] Sheng Liu, Zhe Li, Weiheng Wang, Han Sun, Heng Zhang, Hongpeng Chen, Yusen Qin, Arash Ajoudani, and Yizhao Wang. Activepose: Active 6d object pose estimation and tracking for robotic manipulation. *arXiv preprint arXiv:2509.11364*, 2025.

[35] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024.

[36] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations (ICLR)*, 2016.

[37] Tung M. Luu, Donghoon Lee, Younghwan Lee, and Chang D. Yoo. Policy learning from large vision-language model feedback without reward modeling. In *2025*

IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 11685–11692, 2025.

[38] Haoxiang Ma, Modi Shi, Boyang Gao, and Di Huang. Active perception for grasp detection via neural graspness field. *Advances in Neural Information Processing Systems*, 37:38122–38141, 2024.

[39] Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio Ojea, and Ken Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. In *Proceedings of Robotics: Science and Systems (RSS)*, 2017.

[40] John McCormac, Ankur Handa, Andrew Davison, and Stefan Leutenegger. Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4628–4635, 2017.

[41] A.T. Miller, S. Knoop, H.I. Christensen, and P.K. Allen. Automatic grasp planning using shape primitives. In *2003 IEEE International Conference on Robotics and Automation (ICRA)*, volume 2, pages 1824–1829, 2003.

[42] Douglas Morrison, Peter Corke, and Jürgen Leitner. Multi-view picking: Next-best-view reaching for improved grasping in clutter. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8762–8768, 2019.

[43] OpenAI. Gpt-4o system card, 2024.

[44] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features Without Supervision. *Transactions on Machine Learning Research*, 2024.

[45] Yixuan Qian, Xinyi Zhu, Ondrej Biza, Shengxiang Jiang, Linfeng Zhao, Haojie Huang, Yixin Qi, and Robert Platt. ThinkGrasp: A vision-language system for strategic part grasping in clutter. In *8th Annual Conference on Robot Learning*, pages 3568–3586. PMLR, 2025.

[46] Danfeng Qin, Chas Leichner, Manolis Delakis, Marco Fornoni, Shixin Luo, Fan Yang, Weijun Wang, Colby Banbury, Chengxi Ye, Berkin Akin, Vaibhav Aggarwal, Tenghui Zhu, Daniele Moro, and Andrew Howard. MobileNetV4: Universal models for the mobile ecosystem. In *18th European Conference on Computer Vision (ECCV)*, pages 78–96. Springer, 2024.

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

[48] Henry Schaub, Christian Wolff, Maximilian Hoh, and Alfred Schöttl. Probabilistic closed-loop active grasping. *IEEE Robotics and Automation Letters*, 9(4):3964–3971, 2024.

[49] Yitian Shi, Di Wen, Guanqi Chen, Edgar Welte, Sheng Liu, Kunyu Peng, Rainer Stiefelhagen, and Rania Rayyes. Viso-grasp: Vision-language informed spatial object-centric 6-dof active view planning and grasping in clutter and invisibility. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 14931–14938, 2025.

[50] Xiao'Ao Song and Konstantinos Karydis. Gs-nbv: a geometry-based, semantics-aware viewpoint planning algorithm for avocado harvesting under occlusions. In *2025 IEEE 21st International Conference on Automation Science and Engineering (CASE)*, pages 690–695, 2025.

[51] Yingbo Tang, Shuaike Zhang, Xiaoshuai Hao, Pengwei Wang, Jianlong Wu, Zhongyuan Wang, and Shanghang Zhang. AffordGrasp: In-context affordance reasoning for open-vocabulary task-oriented grasping in clutter. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9433–9439, 2025.

[52] Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Identifying and mitigating position bias of multi-image vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10599–10609, June 2025.

[53] Edmond Tong, Advaith Balaji, Anthony Opipari, Stanley Lewis, Zhen Zeng, and Odest Chadwicke Jenkins. OVAL-Grasp: Open-vocabulary affordance localization for

task oriented grasping. In *19th International Symposium on Experimental Robotics (ISER)*, 2025.

[54] Georgios Tziafas and Hamidreza Kasaei. Towards open-world grasping with large vision-language models. In *8th Annual Conference on Robot Learning*, pages 3304–3332. PMLR, 2025.

[55] Sreyas Venkataraman, Yufei Wang, Ziyu Wang, Navin Sriram Ravie, Zackory Erickson, and David Held. Real-world offline reinforcement learning from vision language model feedback. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13452–13459, 2025.

[56] Shenglin Wang, Mingtong Dai, Jingxuan Su, Lingbo Liu, Chunjie Chen, Xinyu Wu, and Liang Lin. Graspview: Active perception scoring and best-view optimization for robotic grasping in cluttered environments. *arXiv preprint arXiv:2511.04199*, 2025.

[57] Yufei Wang, Zhanyi Sun, Jesse Zhang, Zhou Xian, Erdem Biyik, David Held, and Zackory Erickson. Rl-vlm-f: Reinforcement learning from vision language foundation model feedback. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pages 51484–51501. PMLR, 2024.

[58] Yan Xia, Ran Ding, Ziyuan Qin, Guanqi Zhan, Kaichen Zhou, Long Yang, Hao Dong, and Daniel Cremers. Targo: Benchmarking target-driven object grasping under occlusions. *arXiv preprint arXiv:2407.06168*, 2024.

[59] Yuchen Xiao, Sammie Katt, Andreas ten Pas, Shengjian Chen, and Christopher Amato. Online planning for target object search in clutter under partial observability. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8241–8247, 2019.

[60] Xuechao Zhang, Dong Wang, Sun Han, Weichuang Li, Bin Zhao, Zhigang Wang, Xiaoming Duan, Chongrong Fang, Xuelong Li, and Jianping He. Affordance-driven next-best-view planning for robotic grasping. In *Proceedings of The 7th Conference on Robot Learning*, pages 2849–2862, 2023.

[61] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc., 2023.