AI

MDPI

Article

Constraint-Aware Multimodal Bin Grasping and Singulation in Cluttered Environments

Muhammad Arsalan [1], Abid Imran [1,2,*], Saud Ahmed Khan [1], Ikram Ullah [3], Kamran Aqeel [4] and Muhammad Farooq Siddique [5,6,*]

[1] Faculty of Mechanical Engineering, Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi, Swabi 23640, Khyber Pakhtunkhwa, Pakistan; m.arsalan@giki.edu.pk (M.A.); saudve1@gmail.com (S.A.K.)
[2] Department of Robotics and Automation Engineering, Atlantic Technological University, Galway Campus, Dublin Road, H91 T8NW Galway, Ireland
[3] Department of Computer Software Engineering, University of Engineering and Technology, Charsadda Road, Mardan 23200, Khyber Pakhtunkhwa, Pakistan
[4] Department of Computer Systems Engineering, University of Computing Sciences and Information Technology, Swat 19200, Khyber Pakhtunkhwa, Pakistan; kamran.aqeel@ueas.edu.pk
[5] Department of Mechanical Engineering, University of Engineering and Technology, Charsadda Road, Mardan 23200, Khyber Pakhtunkhwa, Pakistan
[6] Department of Computer Science, Khalifa University, Abu Dhabi 127788, United Arab Emirates
* Correspondence: abid_mecha@yahoo.com (A.I.); farooqsiddique.mech@uetmardan.edu.pk (M.F.S.)

Abstract

Rigid container walls invalidate the open-workspace assumptions of tabletop manipulation, leaving three coupled gaps: unconstrained trajectories can be wall-unsafe, single-point grasp checks reject accessible offset positions, and occlusion-removal actions can ignore wall proximity. This paper treats bin wall geometry as a first-class constraint. A multimodal framework couples YOLO11 instance segmentation with a radial-basis-function support vector machine (SVM) and geometric rules to select among five modes—direct grasping, height-based grasping, two-dimensional singulation, and partial and full occlusion removal based on target visibility, relative height, grasp feasibility and wall proximity. Every trajectory passes a common wall-constraint validator before reaching the controller. Experiments used a UR5e with an overhead RGB-D camera. YOLOv11m-Segmentation model reached 0.992 box and 0.991 mask mAP@0.5 with an average inference time of 0.4350 ± 0.11 s. Task success was 100.0%, 93.3%, and 86.66% at low, medium, and high clutter (30 trials per condition). At high clutter, the ablation increased TSR from 66.66% for the single-point wall-unaware baseline to 83.33% with multi-point grasp selection and 86.66% with wall-aware singulation. The proposed system achieved a 90% Task Success Rate (TSR), compared with Target-Oriented VPG, which achieved 60% TSR, and GraspNet with bin-wall constraints, which achieved 40% TSR with 10 trials in a high-clutter environment. Across the 90 proposed-system trials, six task failures were observed, but none involved wall contact. Because container awareness sits in the planning layer rather than the end-effector, the method retrofits to installed cells using standard two-finger hardware.

Keywords: bin picking; robotic grasping; object singulation; occlusion handling; constraint-aware motion planning; instance segmentation; industrial manipulation

Check for updates

Academic Editor: Miguel Angel Cazorla

Received: 25 August 2026
Revised: 7 October 2026
Accepted: 19 October 2026
Published: 3 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Robotic bin picking, the retrieval of individual items from an unordered collection held inside a rigid container, is among the most economically consequential unsolved problems

in industrial manipulation. In discrete manufacturing it underpins automated part feeding, machine tending and kitting, where components arrive in bulk totes rather than in fixture, pre-oriented presentations. In logistics and e-commerce order fulfilment it underpins piece picking from storage totes, mixed-SKU de-palletizing and parcel induction. In materials recovery it underpins the separation of mixed waste streams into recoverable fractions [1,2]. What unites these settings is that the container, rather than the open table, is the native presentation format for the objects to be handled. Bogue [3] observed that bin picking has the potential to eliminate repetitive manual part-handling across large parts of the manufacturing and logistics sectors, but notes that deployed systems remain complex to install and commission. That gap between demonstrated capability and routine deployment has motivated a decade of sustained research effort, from the Amazon Picking and Robotics Challenges [4–6] through to purpose-built industrial benchmarks and datasets [7,8].

The significance of the container is that it removes assumptions that are often made in the manipulation literature. A robot on a flat tabletop can move laterally in any direction from its current position to move away from an obstacle and can come towards a target from any lateral direction that is convenient. When operating inside a bin, these assumptions no longer hold. The workspace is enclosed on all sides by rigid walls, so that a trajectory that is freely generated can cause the end-effector to enter the container. Under gravity, vibration and previous disturbance, objects are forced to move towards the boundaries, and the target may become pressed against a wall, without any clearance for the gripper fingers. However, occasionally objects will not be pushed off easily, since the direction in which the object would be exposed may be the same direction as the direction the wall is obstructing. The results of not taking these restrictions into account are not just lower success rates; the failure of a pick will stop the cell dead, and manual intervention will be required; in a production environment, there is a much bigger expense than a failed pick. These are real challenges to deployment [3,9] but are not necessarily explicitly displayed in the manipulation literature.

Although perception-driven manipulation has progressed substantially, important challenges remain in cluttered bin environments. Deep convolutional detectors can detect objects at real-time speed with cluttered background in natural environments [10,11], and the RGB-D pose estimation approaches can accurately estimate the pose of objects if their full 3D models are provided [12–14]. However, in bin picking, such model-based approaches are not always feasible, due to the volatility of the item populations and the lack of CAD models. Simplified grasp representations, especially 4-DOF top-down grasping, have therefore remain widely used. They are easily computed and represent the end-effectors in use in the industry; the parallel-jaw and suction end-effector [15–17]. The proposed system is designed for horizontal bin-picking scenarios using a 4-DOF top-down grasping policy. It assumes that the target object has a sufficiently accessible grasp region from the upper workspace and does not address substantially inclined containers, deep undercuts, or configurations requiring non-top-down approach directions.

Singulation is one of the key sub-problems: objects need to be moved to space them apart so that the object to be grasped can be reached. This is because classical pushing models which can give analytical guarantees rely on assumptions about the objects geometry and the friction which do not hold in real bins [18,19]. In cluttered environments, they have been largely replaced by learning-based policies [20–23]. Another challenge is occlusion handling, which allows for localizing partially occluded targets but may not be able to safely grasp them, and fully occluded targets must be uncovered via a series of exposing actions whose order is not necessarily optimized for a limited container geometry [24–26].

A similar limitation arises in both singulation and occlusion handling. Methods of singulation that fail to model bin boundaries yield pushes that either may collide with

a wall or move an object inadequately against a wall. Feasibility checks were performed at only one point on the target and evaluated whether the grasp would be successful if positioned a certain distance away. Wall-oblivious occlusion-removal algorithms may place the removed object back into the target area or displace it into a corner, adding to the clutter in the scene. These are all individually small; the combined effect of all of them creates a significant difference between tabletop and in-bin performance and they cannot be optimized separately, since an optimized grasp decision, push direction, and de-stacking direction can all be limited by the same geometry.

In this paper, an adaptive multimodal approach to bin picking is proposed, where bin wall geometry is considered as a first-class constraint both in the selection of the actions and in the generation of the picking trajectory. The five manipulation modes are direct grasping, height-based grasping, two-dimensional singulation, partial-occlusion removal, and full-occlusion removal, which are selected based on the visibility of the target, grasping feasibility, and proximity of the target to the workspace and the relative height between the target and the workspace. All of the trajectories generated by all of the modes are passed through a common-trajectory validator that checks the trajectory for collisions with the walls before it is sent to the robot controller. The novelty therefore lies in the coupling of grasp selection, singulation and occlusion recovery under one container-aware decision loop, rather than in the use of any individual perception or learning model.

The main contributions of this work are:

- A constraint-aware multimodal decision framework that jointly reasons over target visibility, relative height, grasp feasibility and wall proximity to select among five manipulation modes.
- A unified wall-constraint trajectory validator that checks every generated waypoint against the container boundary and invokes a wall-safe fallback when the candidate trajectory violates the admissible region.
- A nonlinear RBF-SVM two-dimensional singulation strategy that generates target-centered separation boundaries, combined with a wall-safe linear fallback when the nonlinear path is inadmissible.
- A three-point grasp-feasibility scheme that evaluates the centroid and two principal-axis offsets, reducing dependence on a single centroid candidate while retaining low computational cost.
- Physical validation on a UR5e platform across three graded clutter levels, with 90 trials used to characterize system-level performance and motion demand.

The remainder of this paper is organized as follows. Section 2 reviews related work on industrial bin picking, visual perception, singulation, and occlusion handling. Section 3 presents the proposed method together with the experimental setup and evaluation metrics. Section 4 reports the results. Section 5 discusses their implications and the limitations of the approach, and Section 6 concludes.

2. Related Work

2.1. Industrial Bin Picking: Applications, Deployment, and Benchmarks

Bin picking is a distinctive field of robotics: on the one hand it is already a well-established product category for commercial use, and, on the other hand, it is still an active research problem. Bogue [3] discussed both aspects of the field, with examples of robotic gripping, machine vision and specialist software, provided by robot manufacturers, system integrators, and vision specialists, but complexity of installation and commissioning remains a barrier to adoption. It has a broad application range. In the automotive and general manufacturing industry, fasteners, castings, and machined parts are delivered to

assembly lines via bins and are fed by standardized totes. In warehousing, the rise in e-commerce orders has become a key target for automation with piece picking from storage totes being a prime candidate [27,28]. Robots are also used in materials recovery facilities to separate mixed waste streams with extreme clutter and object variability [1,29], where the variety of objects to be separated is effectively unbounded by construction, and CAD based approaches are not available.

The most prominent common testbeds in the field were the Amazon Picking Challenge and the Amazon Robotics Challenge. At the first event, twenty-six teams participated in a survey that Correll et al. [4] conducted to determine which factors influence performance as a function of the type of mechanism chosen, and how the other factors affect performance. Correll et al. found that the type of mechanism chosen has a significant effect on execution speed, while other factors that do not depend on the mechanism, such as systems integration and a combination of reactive control and deliberative planning, are also factors. From the winning entry, Eppner et al. [5] reached a similar conclusion that the construction of a robotic system requires a balance between modularity & integration, generality & assumption, computation & embodiment, and planning & feedback. Corbato et al. [6] had similar results with the 2016 winning system and Morrison et al. [30] demonstrated with a low-cost Cartesian manipulator that it is better to design a system to the task than to make a more general one. Fujita et al. [31] then scrutinized the competition entries and selected those based on a common set of performance metrics and determined which of these ones mattered. What is common in this work is that regardless of the type of container system design (shelf bin or tote), the geometry of the container dominates object recognition more than it does in the algorithms that followed, and that the geometry of the container is the most commonly abstracted away aspect of the system.

Benchmark datasets have followed the same industrial orientation. Kleeberger et al. [7] released a large-scale 6D pose dataset generated by iteratively filling a bin with parts possessing various symmetries, explicitly targeting industrial bin picking rather than tabletop scenes. Gilles et al. [8] introduced MetaGraspNet, comprising 217k synthetic RGB-D images across 82 article types together with a real subset graded into five difficulty levels, with annotations for a modal perception, manipulation ordering and both parallel-jaw and vacuum grasps. Back et al. [32] provided a recent large-scale real-world dataset for perception and grasping in cluttered scenes. These resources have accelerated perception research considerably, but they are principally annotated for recognition and grasp quality rather than for the feasibility of a *motion* within the container, which is the aspect this paper addresses.

A related industrial thread concerns entanglement, where parts such as wire harnesses, tubes and hooked components interlock inside the bin. Leão et al. [33] addressed tube entanglement detection and resolution; Moosmann et al. [34,35] applied deep reinforcement learning and transfer learning to separate entangled workpieces in random bin picking; and Zhang et al. [36] learned efficient policies for picking entangled wire harnesses. This literature is instructive because it takes the container seriously as a physical constraint, since separation motions must remain inside the bin, but it treats entanglement rather than wall proximity as the operative difficulty, and the two require different geometric reasoning.

2.2. Visual Perception for Bin Grasping

Before convolutional networks became practical for real-time use, bin object recognition relied on template matching and handcrafted descriptors. The shift from engineered to learned features raised accuracy substantially, particularly for objects in varied orientations or under partial occlusion. Two-stage detectors demonstrated that high accuracy need not preclude deploy-ability, and single-stage detectors subsequently closed much of the

gap [10,11]. Cordeiro et al. [9] reviewed perception methods developed for bin picking since 2009 and organized them into object localization, pose estimation, and grasp prediction, noting that boundary-limited environments such as bins and totes present distinct difficulties from unbounded scenes.

The YOLO family has proved well suited to manipulation pipelines. YOLOv11m-seg in particular introduces refinements to the decoupled segmentation head that benefit thin-profile and irregularly shaped objects, both common in bins. Alongside trained detectors, foundation segmentation models have added zero-shot capability for objects outside a detector's training distribution [37,38], and open-vocabulary and language-conditioned formulations extend this further [39,40].

A separate line of work improves the depth channel itself, which matters directly for the height-based reasoning used in this paper. Yang et al. [12] fused active stereo depth maps probabilistically across viewpoints for bin picking, and Fu et al. [41] demonstrated a non-stop multi-view active vision scheme that substantially reduces task time in a factory setting. Li et al. [13] addressed the sim-to-real gap for industrial bin-picking recognition and localization. These contributions improve the quality of the perceptual input; they do not address what the planner is permitted to do with it once the container boundary is taken into account.

2.3. Object Singulation and Push Planning

Early analytical models of pushing were developed under quasi-static assumptions with known surface friction. They give rise to predictions that can be interpreted and that are reasonably valid in controlled experiments but break down in the real bin due to non-uniform contact between particles [18,19]. For cluttered scenes, they have been largely replaced by learning-based methods, especially those that operate directly on the visual input [20–23,42–44]. Such systems jointly learn when to push and when to grasp without the need for hand-specified heuristics. The inspirational push-grasp synergy framework developed by Zeng et al. [20] was followed by a significant line of research that went beyond the general clearance task to the one of picking up a given target object [45–48].

In particular, it was noted that an attempt to fit a decision boundary which is not linear around the target object causes an outward push or separation of the surrounding objects from the target object, which is not a push across the target as the linear decision boundary would do. This formulation was developed and validated without a container boundary check on open tabletops. The same classification is used, with the addition of an explicit boundary-enforcement stage: The generated trajectory is tested pointwise against the four wall coordinates and any trajectory resulting in a clearance margin violation is replaced prior to output to the arm controller.

2.4. Grasping Under Occlusion

When a target is partially occluded, the occluding object must be displaced before a grasp can proceed, and the removal trajectory must be designed so that the displaced object does not land back on the target or wedge against the container side [17,49]. Under complete occlusion the robot must infer the target's likely location from indirect evidence, namely the relative heights of visible objects and the structure implied by their segmentation masks and then act to reveal it. Mechanical search formulations address exactly this problem of multi-step retrieval of an occluded target [24,50]. Recent solutions include hierarchical stacking relationship prediction [51], graph-based deep reinforcement learning for fully occluded targets [25], pre-grasp manipulation policies for dense clutter [26], and goal-oriented non-prehensile pushing [52].

These works establish that occlusion is tractable, but they share a common assumption: the removal direction is unconstrained. On an open surface an occluding object can be pushed arbitrarily far in any direction. Inside a bin, the safe set of removal directions is a function of where the occluding object sits relative to the walls, and a strategy that does not compute that set will periodically select a direction that either collides with the boundary or leaves the displaced object in a worse configuration than before.

2.5. Bin-Specific Challenges and Hardware Approaches

A distinct strand of work accepts the container constraint and addresses it through hardware. Corner-Grasp [53] redesigns the gripper head to reach into container corners, and TetraGrip [54] employs multiple suction contacts to tolerate variation in surface type. Three-fingered adaptive grippers and shape-adaptive end-effectors extend the range of tractable bin contents further. Human–robot collaboration has also been proposed as a route to improved bin-picking throughput [55]. Two-stage formulations that separate localization from grasp selection improve reliability for small objects [56].

These are genuine contributions to real problems. They are, however, capital-intensive and carry integration cost, since they require modification or replacement of the end-effector on an installed cell. The position taken in this paper is that a software-side approach capable of handling the most prevalent wall-proximity situations is more readily deployable and more easily retrofitted to existing installations using standard two-finger hardware.

2.6. Summary and Research Gap

The literature comparison in Table 1 highlights several research gaps in existing approaches to cluttered manipulation and bin picking. First, previous studies address individual aspects such as recognition, grasping, singulation, or occluded-object retrieval, but these capabilities are generally not integrated within a single adaptive framework. Second, although singulation and occlusion handling have been extensively investigated, explicit bin-wall constraints are not reported in the reviewed methods, including both bin-specific and clutter-manipulation approaches. Third, existing methods generally lack a common geometric validation stage that evaluates manipulation trajectories against the container boundary before execution.

Table 1. Literature review summarized and comparison with existing work. ST = singulation trajectories, NST = non-singulation Trajectories, PO = partial Occlusion, CO = complete Occlusion.

Ref.	Environment/ Application	ST/NST	PO/CO	Explicit Bin-Wall Constraint	Limitation
Gilles et al. (2022) [8]	Cluttered bin picking	X	X	X	No unified bin-wall-constrained trajectory validation.
Li et al. (2022) [13]	Industrial bin picking	X	X	X	Focuses on recognition/ localization rather than constrained manipulation.
Cao et al. (2023) [56]	Bin picking of small objects	X	X	X	Does not jointly address singulation, occlusion, and wall-constrained trajectories.

Table 1. *Cont.*

Ref.	Environment/ Application	ST/NST	PO/CO	Explicit Bin-Wall Constraint	Limitation
Ong Son et al. (2025) [53]	Cluttered/ container manipulation	X	X	X	Uses gripper adaptation rather than software-based wall constraint.
Torrado et al. (2025) [54]	Cluttered manipulation/ bin-type scenarios	X	X	X	Hardware-centric approach without unified wall-aware trajectory validation.
Dogar and Srinivasa (2011) [18]	Cluttered manipulation/ open workspace	X	X	X	Developed for open-workspace push-grasping, not bounded bins.
Sarantopoulos et al. (2020) [21]	Cluttered object singulation	✓ Deep-Q-learning singulation	X	X	Singulation does not explicitly enforce bin-wall clearance.
Kiatos et al. (2022) [22]	Dense clutter	X	X	X	Push-grasp learning lacks a unified container-wall constraint.
Yang et al. (2020) [45]	Cluttered manipulation/ target grasping	X	X	X	Addresses occluded grasping without bin-wall-constrained recovery.
Huang et al. (2022) [46]	Cluttered object retrieval	X	X	X	Object retrieval lacks a common bin-wall geometric constraint.
Murali et al. (2020) [49]	Target-driven manipulation in clutter	X	✓ Partial occlusion	X	Does not integrate wall-aware singulation and occlusion recovery.
Proposed method	Target object grasping in dense cluttered bin environment	✓ RBF-SVM 2D singulation + wall-safe fallback	✓ Partial + complete occlusion removal	✓ wall-clearance validator for every trajectory	Integrates direct/height-based grasping, multi-point grasp feasibility, singulation, partial/ full occlusion removal, and common container-wall validation within one adaptive decision loop.

These observations identify a specific research gap: existing approaches do not jointly integrate direct/depth-based grasping, multi-point grasp feasibility, 2D singulation, partial and complete occlusion removal, and explicit bin-wall safety within a unified decision framework. The proposed method addresses this gap by treating the bin wall as a shared geometric constraint and applying a common wall clearance validator to every generated trajectory before it is passed to the robot controller. This provides a unified mechanism for adapting manipulation actions to the physical boundaries of dense cluttered bin environments.

3. Materials and Methods

3.1. System Overview

The overall architecture is shown in Figure 1. The design follows three principles. First, every component reasons explicitly about the container boundary rather than treating the workspace as unbounded. Second, action selection is driven by the measured state of the bin at each cycle rather than by a fixed sequence. Third, all five action modes converge on a

single trajectory validation stage before any command reaches the robot controller, so that wall safety is enforced structurally rather than being implemented within each mode.

At the start of every control cycle an Intel RealSense D457 depth camera (RealSense, Inc., Cupertino, CA, USA) captures a registered RGB-D frame. The color image is passed to a YOLOv11m-seg network, which returns per-instance bounding boxes, class labels and segmentation masks for the sixteen object classes on which it was trained. The adaptive action selector then evaluates the bin state and dispatches to one of five modes. The trajectory produced by the selected mode is validated against the wall-clearance constraint, corrected if necessary, and executed.

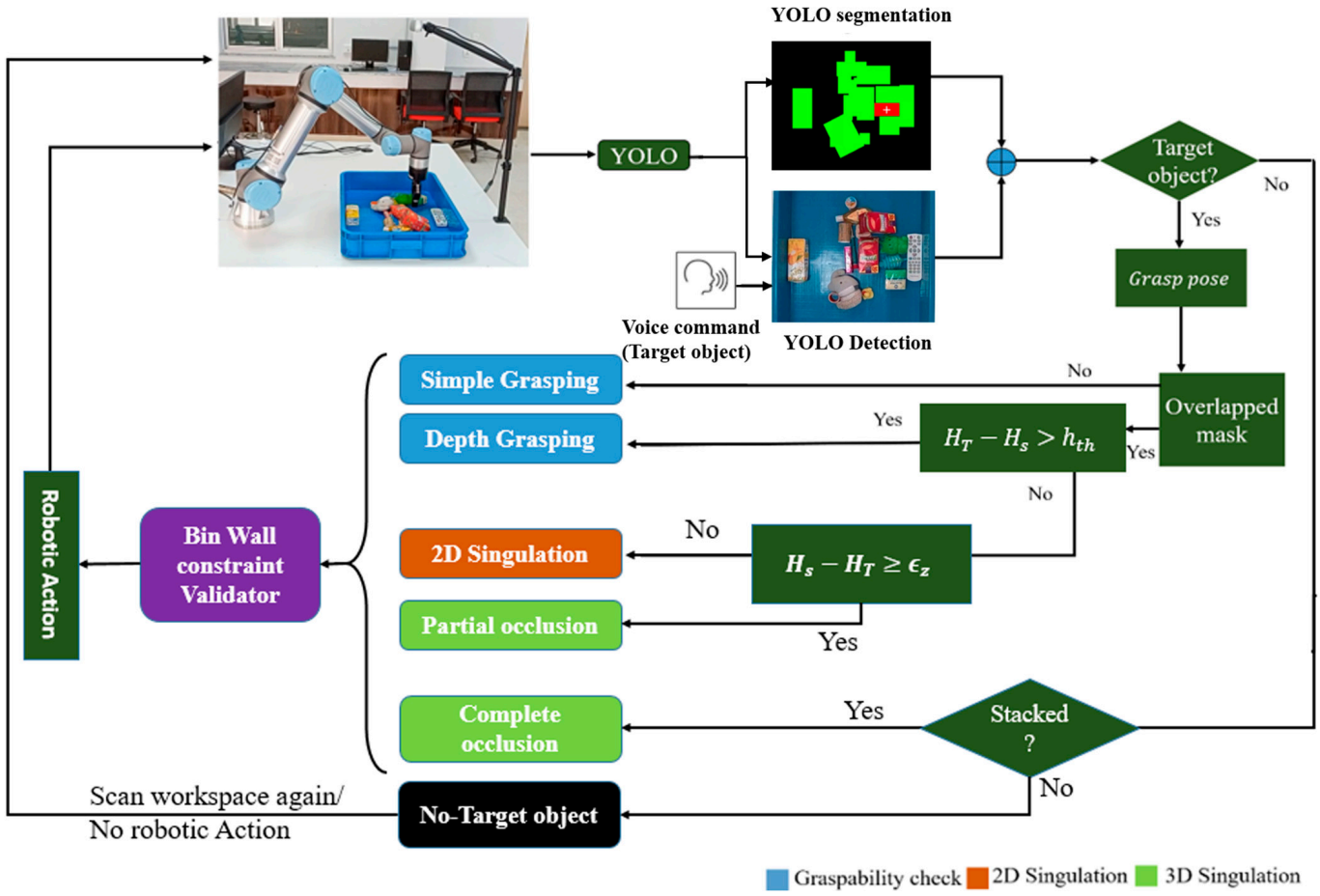


Figure 1. Overall system architecture, covering direct grasping, height-based grasping, two-dimensional singulation and occlusion removal, with wall-constraint validation applied to every generated trajectory.

3.2. Problem Formulation and Notation

The bin is modelled as a rectangular container whose interior footprint is

$$\mathcal{B} = \{(x,y) \in \mathbb{R}^2 : x_{\min} \leq x \leq x_{\max}, y_{\min} \leq y \leq y_{\max}\} \quad (1)$$

To keep the end-effector clear of the walls, a reduced admissible region is defined by contracting $\mathcal{B}$ by a clearance margin δ:

$$\mathcal{B} = \{(x,y) \in \mathcal{B} : x_{\min} + \delta \leq x \leq x_{\max} - \delta, y_{\min} + \delta \leq y \leq y_{\max} - \delta\} \quad (2)$$

with $\delta = 30$ mm throughout this work. The value was selected as the smallest margin that reliably accommodated the finger width of the gripper together with the residual calibration error of the hand-eye transform.

The perception module returns a set of instance masks at each cycle, partitioned into the target mask M_t and the set of surrounding object masks $M_o = \{M_1,...,M_n\}$. It is convenient to define their union as

$$U_o = \bigcup_{i=1}^{n} M_i \quad (3)$$

The thresholds were selected according to the geometric characteristics of the gripper and tested objects. A 30 mm wall-clearance threshold was used to provide sufficient separation between the gripper and bin wall and prevent unintended contact. A 10 mm height-difference threshold was selected to ensure reliable top-down grasping; when the difference is below this value, the object is first singulated to reduce collision risk with surrounding objects. A 5 mm de-stacking offset was selected to provide sufficient displacement of the occluding object and create a graspable region. These values are configuration-dependent: changes in gripper geometry may require adjustment of wall clearance, while changes in bin dimensions alone do not inherently require retuning. The height and de-stacking thresholds primarily depend on the object dimensions.

3.3. Multimodal Perception Module

Instance segmentation. The YOLOv11m-seg model was trained on a custom dataset acquired in bin and tabletop environments under varying illumination. With the segmentation head resized from the 80 COCO classes to the sixteen classes used here, the network comprises 22.35 million parameters and requires 113.0 GFLOPs per forward pass at 640 × 640 input. The medium variant was selected over the small one because the additional capacity measurably improved mask quality on the thin, elongated objects in the set, at an inference cost that still clears the control loop rate by a wide margin. The resulting segmentation masks are the common input to the grasp pose estimator, the three-point feasibility checker and the singulation module.

Dataset split and applicability: The 6422 images were collected across separate recording sessions with different object arrangements and clutter configurations. The test set was obtained from independent recording sessions and did not contain frames from the training sequences, thereby reducing the risk of near-duplicate frame leakage. However, the dataset remains limited to the 16 object classes and the specific bin geometry used in this study. Therefore, the reported results demonstrate applicability within the evaluated setup rather than general industrial deployment.

Target designation. In addition to vision, the second input modality is voice, as the second part of the instruction specifies the object to be retrieved. Commands are recorded via a USB microphone and transcribed by an offline speech recognition system, eliminating network latency and keeping everything contained within the cell. The command vocabulary is deliberately limited: it consists of the 16 trained class names plus a limited number of carrier phrases like “pick the” and “get the”. This limits the grammar and increases the recognition accuracy well beyond that achieved by open dictation and eliminates the effects of an out-of-vocabulary word. The class name recognized is compared to the label set of the segmentation network, and the corresponding instance detected by the segmentation network is taken as the target mask M_t for this cycle. If the spoken class is not present in scene, the system reports that the class is ‘absent’ and waits for a new class to be spoken.

Coordinate transformation. A pixel (u,v) with associated depth d is back-projected into the camera frame using the intrinsic parameters (f_x, f_y, c_x, c_y):

$$\mathbf{p}^c = \left[\frac{(u - c_x)\, d}{f_x}, \frac{(v - c_y)\, d}{f_y}, d \right]^{\top} \quad (4)$$

and mapped into the robot base frame through the hand-eye transform ${}^{b}\mathbf{T}_{c}$:

$$\begin{bmatrix} \mathrm{p}^{b} \\ 1 \end{bmatrix} = {}^{b}\mathrm{T}_{c} \begin{bmatrix} \mathrm{p}^{c} \\ 1 \end{bmatrix}, \qquad {}^{b}\mathrm{T}_{c} = \begin{bmatrix} \mathrm{R} & \mathrm{t} \\ 0^{\top} & 1 \end{bmatrix} \tag{5}$$

Object orientation. The principal axis of the target is recovered from the second-order central moments of its mask. With centroid $(\overline{u}, \overline{v})$, the central moments are

$$\mu_{pq} = \sum_{(u,v)\in M_t} (u - \overline{u})^{p} (v - \overline{v})^{q} \tag{6}$$

and the orientation of the principal axis follows as

$$\theta = \frac{1}{2} \arctan\left(\frac{2\mu_{11}}{\mu_{20} - \mu_{02}} \right), \qquad \hat{\mathbf{u}} = [\cos\theta,\ \sin\theta]^{\top} \tag{7}$$

3.4. Adaptive Action Selection

The selector evaluates four predicates in order at each cycle: whether the target is detected at all; whether a collision-free grasp exists; whether the target stands sufficiently proud of its neighbors; and whether the candidate action would violate the wall clearance. The outcome determines the active mode:

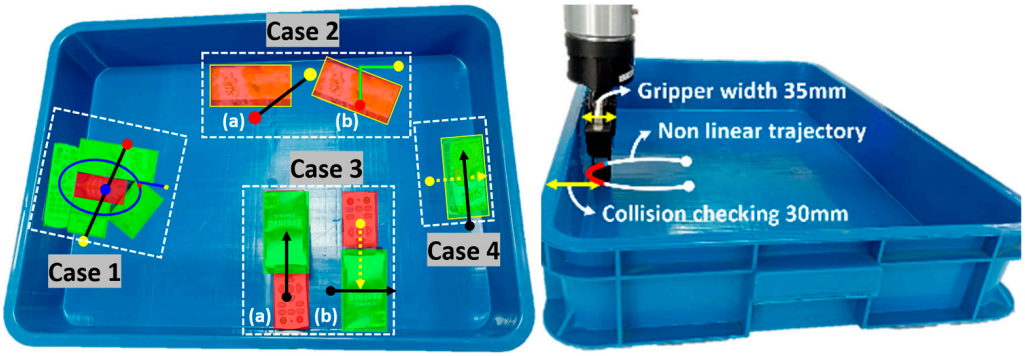
- **Mode 1, Simple Grasping.** The target is visible and at least one of the three candidate positions passes the feasibility test of Section 3.5.
- **Mode 2, Depth grasping.** No candidate is clear, but the target surface stands at least h_{th} above its neighbors, permitting a top-down approach at the target height.
- **Mode 3, 2D singulation.** Target and neighbors are of comparable height and no candidate is clear; a nonlinear push path is generated to displace the neighbors.
- **Mode 4, Partial occlusion.** The target is detected but a stacking object lies partly over it, confirmed by a depth differential.
- **Mode 5, Complete occlusion.** The target is not detected; the highest stack is disturbed and perception is re-run.

The distinction between an absent target, a potentially fully occluded target, and an unreliable detection is handled using object detection and height information. If YOLO11 returns no instance of the commanded class and no evidence of a stacked configuration is observed, the target is declared absent. If no target instance is detected but an object region indicating a height substantially greater than that of an individual object is observed, Mode 5 is activated to perform de-stacking and expose potentially occluded objects. Detections with a posterior confidence below 0.60 are suppressed to reduce false-label interference and are not used to initiate grasping.

As long as the generated trajectories remain within the bin limits, the corresponding decision from the tabletop environment is retained, provided that a minimum clearance of 30 mm from the bin wall is maintained. When a trajectory violates this clearance and risks collision with the bin boundary, an additional linear push trajectory is generated to move the target or occluding object away from the wall. Representative cases are illustrated in Figure 2.

In **Case 1**, non-linear singulation trajectories (blue) that extend beyond the bin boundary are replaced by linear push trajectories (black) to maintain safe operation within the bin. In **Case 2**, when the target object is pressed against the bin wall, the algorithm first applies a corner push toward the bin center (black, **Case 2a**), followed by a center push (green, **Case 2b**) to create sufficient graspable space. In Case 3, for partial occlusion, trajectories directed toward the bin wall (yellow) are modified into parallel linear pushes (black, **Case 3b**), whereas trajectories that originate near the wall and move inward remain unchanged (**Case 3a**) because they do not violate the wall-clearance constraint. In **Case 4**, for

complete occlusion, wall-directed trajectories (yellow) are replaced by linear trajectories parallel to the bin wall (black), allowing the occluding object to be displaced without collision. Figure 2b further illustrates a case in which the robotic gripper approaches the bin boundary. The original non-linear push trajectory is shown in white, with the red-highlighted point indicating the detected unsafe position. A minimum clearance of 30 mm from the bin wall is enforced to prevent collision. Thus, the proposed constraint validator preserves the original grasp or singulation mode whenever it is geometrically safe and activates an additional linear push trajectory only when a potential bin-wall collision is detected.



(**a**) (**b**)

Figure 2. Bin-boundary constraint handling for safe grasping and singulation: (**a**) representative cases of trajectory modification under different object configurations, including boundary violation, wall-contact, partial occlusion, and complete occlusion; (**b**) detection of an unsafe gripper trajectory and application of the 30 mm minimum wall-clearance constraint.

3.5. Three-Point Grasp Feasibility Check

Conventional feasibility checks evaluate only the centroid of the target and test whether the gripper finger rectangles intersect neighboring masks. This single-point test rejects configurations in which the centroid is blocked but one end of the object is clear. The proposed check evaluates three candidate positions distributed along the principal axis (Figure 3a):

$$\mathbf{p}_k = \mathbf{c} + k\,\frac{L}{4}\,\hat{\mathbf{u}}, \qquad k \in \{-1, 0, +1\} \tag{8}$$

where $\mathbf{c}$ is the mask centroid and L the object extent along $\hat{\mathbf{u}}$. At each candidate a rectangular template $\mathcal{R}(\mathbf{p}_k,\theta)$ representing the closed gripper fingers is placed, with dimensions

$$w_R = w_o + 2c_l, \qquad l_R = l_f \tag{9}$$

in which w_o is the measured object width, $c_l = 20$ mm is the lateral clearance allowed on each side, and $l_f = 40$ mm is the finger length. The degree of interference at each candidate is the area of intersection between the template and the surrounding masks,

$$O(\mathbf{p}_k) = |\mathcal{R}(\mathbf{p}_k,\theta) \cap U_o| \tag{10}$$

and the feasible set is those candidates that are simultaneously free of interference and wholly inside the admissible region:

$$\mathcal{F} = \{\mathbf{p}_k : O(\mathbf{p}_k) = 0 \wedge \mathcal{R}(\mathbf{p}_k,\theta) \subset \mathcal{B}_\delta\} \tag{11}$$

If $\mathcal{F} \neq \emptyset$, Mode 1 is selected and the executed grasp is the feasible candidate closest to the centroid, which minimizes the moment arm about the object's center of mass:

$$\mathbf{p}^* = \text{argmin}_{\mathbf{p}_k \in \mathcal{F}} \| \mathbf{p}_k - \mathbf{c} \| \tag{12}$$

The second condition in Equation (11) is what excludes the wall-adjacent grasp candidates that a conventional check would accept.

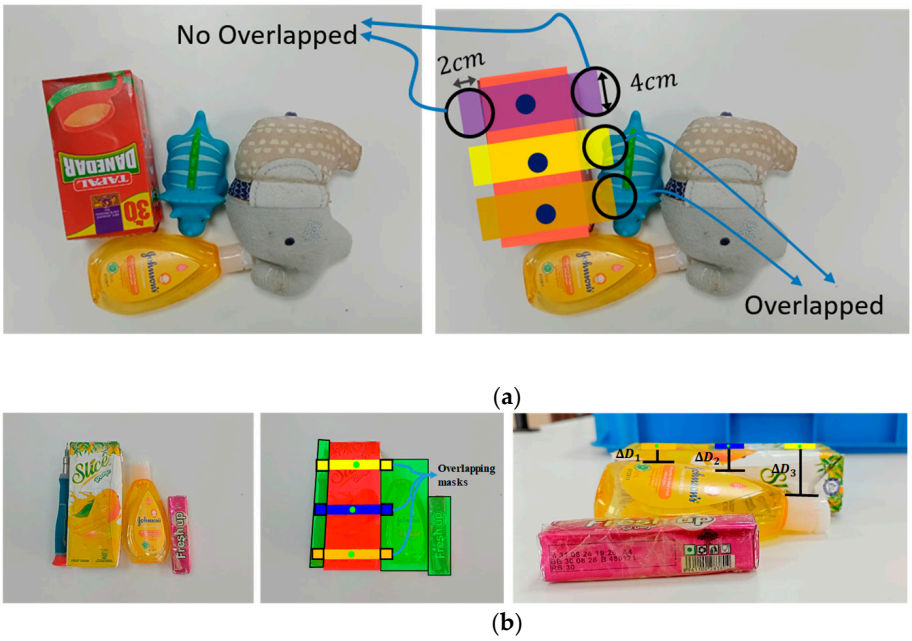


Figure 3. Grasping scenarios. (**a**) Three-point grasp feasibility check, showing the gripper template with 40 mm finger length and 20 mm lateral clearance evaluated at three candidate positions. (**b**) Height-based grasping, showing the height differentials measured at each candidate position.

The 4-DOF policy generates top-down grasp poses and is therefore applicable when the target provides an accessible grasp region from above. Objects located near bin walls are first considered by the wall-aware singulation module and, when necessary, repositioned toward the bin center before grasp execution. However, if the bin wall obstructs the required gripper approach or the object is located beneath a geometrically inaccessible region, a top-down grasp may become infeasible.

3.6. Height-Based Grasping

When $\mathcal{F} = \emptyset$, the system tests whether the target protrudes sufficiently above its surroundings for a top-down approach to clear them. Writing z_t for the target surface height and $\mathcal{N}(\mathbf{p}_k)$ for the neighborhood of candidate k, the height differential is

$$\Delta h(\mathbf{p}_k) = z_t(\mathbf{p}_k) - \max_{(u,v)\in\mathcal{N}(\mathbf{p}_k)} z_o(u,v) \tag{13}$$

Mode 2 is triggered when

$$\max_k \Delta h(\mathbf{p}_k) \geq h_{\text{th}}, \qquad h_{\text{th}} = 10 \text{ mm} \tag{14}$$

Figure 3b shows the height differentials measured at each candidate position, and the grasp is executed at the position of greatest clearance, $\mathbf{p}^* = \text{argmax}_k \Delta h(\mathbf{p}_k)$.

3.7. 2D Singulation

If no candidate is clear and no candidate satisfies the height criterion, a singulation path is generated, as illustrated in Figure 4. A machine learning technique, support vector machine (SVM), can be used for both regression and classification tasks [57]. In classification, SVM can construct nonlinear decision boundaries through appropriate kernel functions. Therefore, an RBF-kernel SVM is employed in this study to generate the nonlinear singulation trajectory. A binary support vector (SVM) machine with a radial basis function

kernel is trained at each singulation cycle on the boundary pixels of the relevant masks. Restricting the training set to boundary pixels rather than all mask pixels reduces the sample count by approximately an order of magnitude, which is what keeps the classification step fast enough for the control loop. The hyper-parameters for the RBF-SVM are regularization constant C = 10, kernel width gamma = 0.01, and class 0 associated with target object mask and class 1 associated with non-target object masks. The training set is

$$\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}, \quad y_i = +1 \text{ if } \mathbf{x}_i \in \partial M_t, \quad y_i = -1 \text{ if } \mathbf{x}_i \in \partial U_o \tag{15}$$

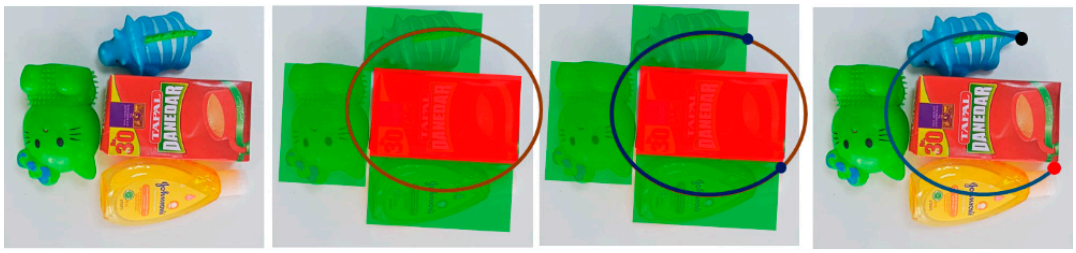
where $\partial(\cdot)$ denotes the boundary pixel set of a mask. The classifier is obtained by solving the dual problem

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K\left(\mathbf{x}_i, \mathbf{x}_j\right) \tag{16}$$

$$\text{subject to} \quad 0 \le \alpha_i \le C, \quad \sum_{i=1}^{N} \alpha_i y_i = 0 \tag{17}$$

with the radial basis function kernel

$$K\left(\mathbf{x}_i, \mathbf{x}_j\right) = \exp\left(-\gamma \| \mathbf{x}_i - \mathbf{x}_j \|^2\right) \tag{18}$$



(**a**) (**b**) (**c**) (**d**)

Figure 4. Nonlinear push trajectory generation for two-dimensional singulation. (**a**) Original RGB frame. (**b**) Decision boundary produced by the nonlinear classifier. (**c**) Overlapping portion of the trajectory, in blue, intersecting the surrounding object masks. (**d**) Final singulation trajectory after removal of the longest non-overlapping segment.

The kernel produces a smooth decision function in image space,

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b \tag{19}$$

whose zero level set envelops the target and forms the candidate singulation path:

$$\mathcal{T}_{\text{svm}} = \{\mathbf{x} \in \Omega : f(\mathbf{x}) = 0\} \tag{20}$$

Interpreted as a robot trajectory, this curve pushes neighboring objects away from the target rather than across it, which is the behavior a straight-line push cannot provide.

Not all of $\mathcal{T}_{\text{svm}}$ is useful. The portions that do not intersect any surrounding mask,

$$\mathcal{T}_{\text{no}} = \{\mathbf{x} \in \mathcal{T}_{\text{svm}} : \mathbf{x} \notin U_o\} = \bigcup_{j=1}^{m} \mathcal{T}_{\text{no}}^{(j)} \tag{21}$$

decompose into m maximal connected segments. Executing the longest of these would displace the target itself rather than its neighbors, so it is removed:

$$\mathcal{T}_{\text{final}} = \mathcal{T}_{\text{svm}} \text{argmax}_j \left| \mathcal{T}_{\text{no}}^{(j)} \right| \tag{22}$$

What remains is the shortest set of boundary segments that contact only the surrounding objects, which is precisely the motion required for singulation. A five-pixel positional

tolerance is applied at the trajectory start point when converting from image to workspace coordinates.

The RBF-SVM uses 2D instance-mask boundary pixels without imposing a convexity constraint, allowing nonlinear contours to represent irregular and non-convex shapes. For thin or elongated objects, the resulting trajectory becomes approximately linear along the dominant object direction, while the soft-margin formulation provides tolerance to minor boundary irregularities and segmentation noise.

3.8. Occlusion Removal

Partial occlusion (Mode 4). The target is detected, but the depth channel indicates a stacking object whose surface lies above the target surface. The mode is triggered when

$$\Delta H_o = H_s - H_t \geq \epsilon_z, \qquad \epsilon_z = 2 \text{ mm} \tag{23}$$

The threshold ϵ_z was determined empirically: below 2 mm the depth noise floor of the sensor produced false triggers at an unacceptable rate.

Full occlusion (Mode 5). When the detector returns no target instance, the target is presumed buried. The algorithm identifies the highest stacked object in the bin and disturbs it, then re-runs perception, repeating until the target appears or all stacked objects have been disturbed. The de-stacking motion is aligned with the *short* axis of the stacking object, which maximizes the displacement achieved per motion:

$$\hat{\mathbf{d}} = \left[\cos\left(\theta_s + \frac{\pi}{2}\right), \sin\left(\theta_s + \frac{\pi}{2}\right)\right]^\top \tag{24}$$

where θ_s is the principal-axis orientation of the stacking object obtained from Equation (7). The push start point is placed at the near end of the short axis, offset below the object surface by $\epsilon_o = 5$ mm, and the end point extends past the far end by a margin ρ:

$$\mathbf{p}_{\text{start}} = \mathbf{c}_s - \frac{l_s}{2}\hat{\mathbf{d}} - \epsilon_o \hat{\mathbf{z}}, \qquad \mathbf{p}_{\text{end}} = \mathbf{c}_s + \left(\frac{l_s}{2} + \rho\right)\hat{\mathbf{d}} \tag{25}$$

If $\hat{\mathbf{d}}$ points toward a wall lying within the clearance zone, the direction is rotated to run parallel to that wall face before execution.

The de-stacking loop operates within a cycle budget of B = 5 iterations, based on the maximum observed stack depth of four objects across the 90 experimental trials. After each de-stacking action, the scene is re-evaluated to detect the target. If the target remains undetected after B iterations, the system terminates the process and reports a task failure.

3.9. Wall-Constraint Trajectory Validator

The validator is the single point through which every trajectory passes, regardless of the mode that generated it. For a waypoint $\mathbf{p} = (p_x, p_y)$, the distance to the nearest wall is

$$d_w(\mathbf{p}) = \min\{p_x - x_{\min}, x_{\max} - p_x, p_y - y_{\min}, y_{\max} - p_y\} \tag{26}$$

and the validation predicate over a trajectory $\mathcal{T}$ is

$$V(\mathcal{T}) = 1 \text{ if } d_w(\mathbf{p}) \geq \delta \ \forall\, \mathbf{p} \in \mathcal{T}, \quad \text{and } V(\mathcal{T}) = 0 \text{ otherwise} \tag{27}$$

When validation fails, the nonlinear path is discarded in its entirety and replaced by a linear push directed toward the bin center $\mathbf{c}_\mathcal{B}$:

$$\mathcal{L}(\mathbf{p}_0, \mathbf{c}_\mathcal{B}) = \{\mathbf{p}_0 + t(\mathbf{c}_\mathcal{B} - \mathbf{p}_0) : t \in [0,1]\} \tag{28}$$

so that the trajectory actually issued to the controller is

$$\mathcal{T}^\dagger = \mathcal{T} \text{ if } V(\mathcal{T}) = 1, \quad \text{and} \quad \mathcal{T}^\dagger = \mathcal{L}(\mathbf{p}_0, \mathbf{c}_\mathcal{B}) \text{ if } V(\mathcal{T}) = 0 \tag{29}$$

Because Equation (29) is applied identically to the output of all five modes, no mode can issue a wall-violating trajectory, and the geometric safety condition does not depend on each mode implementing its own boundary logic correctly.

3.10. Experimental Setup

All experiments were conducted on a UR5e robotic arm (Universal Robots A/S, Odense, Denmark) fitted with a two-finger electric gripper Robotiq gripper (Robotiq Inc., Lévis, QC, Canada). Perception was provided by an Intel RealSense D457 depth camera mounted on a fixed overhead bracket. The bin measured 400 × 300 × 120 mm. Spoken target designations were captured through a USB microphone. System logic, motion control and sensor acquisition were implemented in ROS 2 Humble with Python3.10. The network was trained on an NVIDIA RTX 3060 (12 GB; NVIDIA Corporation, Santa Clara, CA, USA) and deployed on CPU.

The UR5e's integrated force/torque sensing (4 N accuracy) and configurable safety limits were retained, with a conservative force limit below 1 N and reduced motion speed used during linear fallback. This ensured controlled operation and protective stopping without requiring an external force/torque sensor.

3.11. Evaluation Metrics

Two performance metrics are reported. The first metric is the task success rate over a set of N trials, which is:

$$\mathrm{TSR} = \frac{N_s}{N} \times 100\% \tag{30}$$

where N_s is the number of trials in which the target was retrieved without a wall collision or a dropped object. The second matric is the mean motion count is the average number of executed actions per completed task,

$$\overline{m} = \frac{1}{N}\sum_{i=1}^{N} m_i \tag{31}$$

where m_i counts every grasp, push, and de-stacking motion in trial i. Because $N = 30$ per condition, point estimates of TSR carry appreciable uncertainty, so Wilson score intervals at the 95% level are reported alongside them:

$$\hat{p}_{\pm} = \frac{1}{1+\frac{z^2}{N}}\left(\hat{p} + \frac{z^2}{2N} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{N} + \frac{z^2}{4N^2}}\right) \tag{32}$$

with $z = 1.96$. The Wilson interval is preferred to the normal approximation because it remains well behaved as $\hat{p}$ approaches unity, which is the case at the low-clutter condition.

4. Results

4.1. Perception Performance

The segmentation model was trained using Ultralytics 8.4.38 and PyTorch 2.5.1 on a custom dataset of 6422 images (5138 training, 642 validation and 642 test images) captured within the bin and tabletop environment under varying lighting and object configurations. The dataset contains 16 object classes. All selected objects are rigid or semi-rigid, with varying shapes and dimensions. The split was performed at the image level. Training and validation curves are shown in Figure 5. All loss terms decrease steadily without the oscillations or plateaus characteristic of an inappropriate learning rate or batch size. The gap between training and validation curves remains small throughout, indicating that the model is not overfitting despite the continued decrease in training loss. Box and mask mAP@0.5 exceeded 0.99 from approximately epoch 50 onward. Training was configured for

a maximum of 500 epochs, but the validation loss reached its minimum at approximately epoch 100 and training terminated there; Figure 5 therefore shows the complete run rather than a truncated view of it.

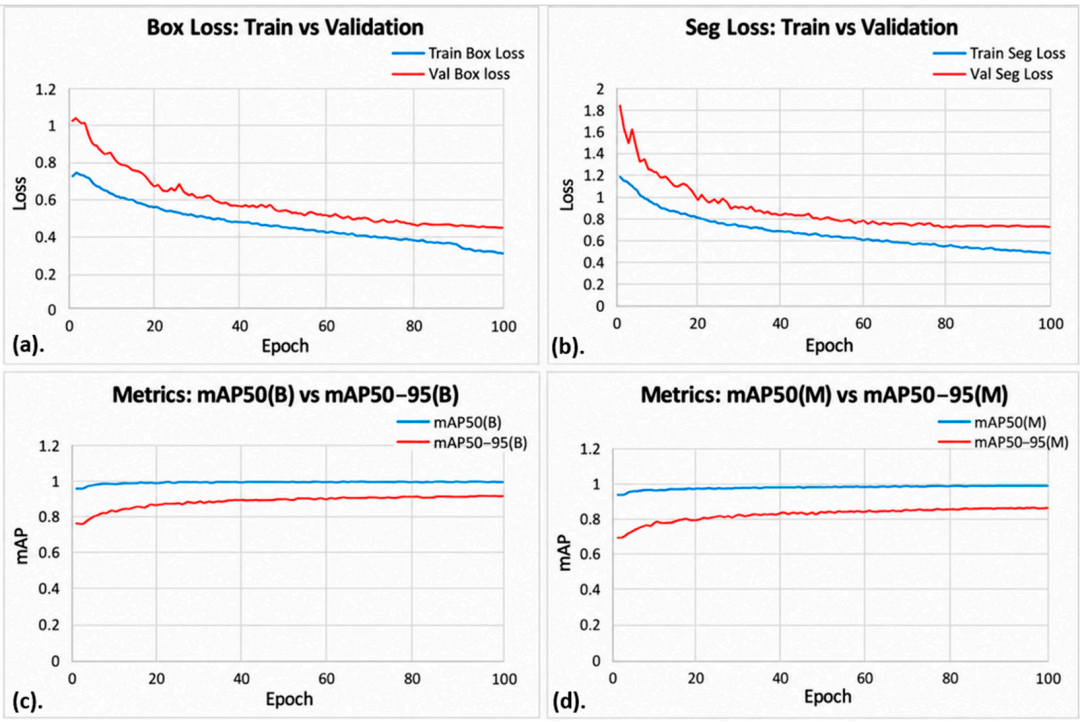


Figure 5. Training and validation curves. (**a**) Bounding box loss. (**b**) Segmentation loss. (**c**) Bounding box mAP. (**d**) Mask mAP.

Table 2 reports per-class segmentation results for all sixteen classes on the held-out test set. Overall box mAP@0.5 reached 0.992 and mask mAP@0.5 reached 0.991, with box mAP@0.5:0.95 of 0.921 and mask mAP@0.5:0.95 of 0.856. The per-class box mAP@0.5 of the lowest two classes was the same, 0.984, for the screwdriver and bandage classes. The thin and long profile of the screwdriver along with the various orientations it can take in a bin make it the most challenging object for most detectors (most of the time, 0.984 is still high enough to make a good grasp decision). The lower mask mAP@0.5:0.95 values for solder paste (0.714) and screwdriver (0.716) indicate greater difficulty in accurately localizing their boundaries at stricter IoU thresholds. The segmentation model reached 0.992 box mAP@0.5 and 0.991 mask mAP@0.5 at 0.4350 ± 0.11 s inference time.

4.2. Qualitative Multi-Mode Behavior

In-bin trials of representative systems are shown in Figure 6 for direct grasping and two-dimensional singulation and occlusion removal. In Figure 6a,b, the toy rhino is detected but is surrounded by non-target objects, so a nonlinear singulation trajectory is generated. The first path would be a push of the gripper against the wall of the bin, so validation will be the substitution of a linear push towards the interior of the bin. In Figure 6c, the target is pushed flat against the bin wall, and two successive linear pushes are created: the first is pushing one corner of the target away from the wall to form clearance; the second pushing the target towards the bin center. Then, there are two grasp candidates, one inside the admissible region $\mathcal{B}_\delta$, and one outside; the latter is rejected by Equation (11), and the one inside is executed. In the case of Figure 6d, the toy cat is surrounded by clutter, but the nonlinear singulation path is completely contained within $\mathcal{B}_\delta$ and thus executes without modification and separates the target in a single motion.

Table 2. Per-class segmentation results on the test set.

Class	Box mAP50	Box mAP50–95	Mask mAP50	Mask mAP50–95
Baby shampoo	0.994	0.931	0.994	0.910
Bandage	0.984	0.928	0.984	0.846
Bubble gum	0.987	0.931	0.987	0.917
Dishwash liquid	0.993	0.963	0.994	0.945
Elephant	0.993	0.943	0.994	0.926
Juice	0.994	0.943	0.994	0.879
Plaster paste	0.991	0.931	0.991	0.861
Plug	0.993	0.945	0.993	0.911
Remote	0.993	0.941	0.993	0.886
Screwdriver	0.984	0.860	0.984	0.716
Solder paste	0.995	0.814	0.993	0.714
Tap	0.995	0.894	0.995	0.739
Tea	0.995	0.961	0.995	0.930
Toy cat	0.995	0.952	0.995	0.895
Toy duck	0.995	0.878	0.995	0.765
Toy rhino	0.995	0.915	0.974	0.853
Overall	0.992	0.921	0.991	0.856

In three of the four trials presented, the initially generated trajectory was indeed geometrically sensible, but not physically safe, and was corrected prior to its arrival at the controller, as illustrated in these cases. Supplementary Video S1 provides an accompanying demonstration of the proposed UR5e bin-manipulation framework, illustrating robot actions for grasping and singulation in cluttered bin environments and the resulting object configurations.

4.3. Performance Across Clutter Levels

The proposed grasping and singulation framework is evaluated for three clutter levels, with 30 trials performed for each clutter level to demonstrate the robustness of proposed framework in more challenging scene conditions. The summary is contained in Table 3.

In a low-clutter environment, there were 4–6 non-target objects around the target object in the bin, all the objects were in contact with the target object, but there was no significant occlusion of the target. The system successfully completed all 30 trials with a task success rate (TSR) of 100.0%. The average number of robot motions was 1.4, meaning that most of the targets could be grasped directly without needing to singulate and/or de-stack them. The success rate of the proposed framework is high, and the number of motions is low, indicating that, when the target is clearly visible and adequate free space is provided around the grasped region, the proposed framework is capable of operating without failure.

Table 3. Task success rate, 95% Wilson confidence Interval (CI), and mean motion count across clutter levels.

Clutter Level	Trials	Successes	TSR (%)	95% Wilson CI for TSR (%)	Mean Motions
Low	30	30	100.0	88.6–100.0	1.4
Medium	30	28	93.3	78.7–98.2	2.7
High	30	26	86.66	70.3–94.7	3.4

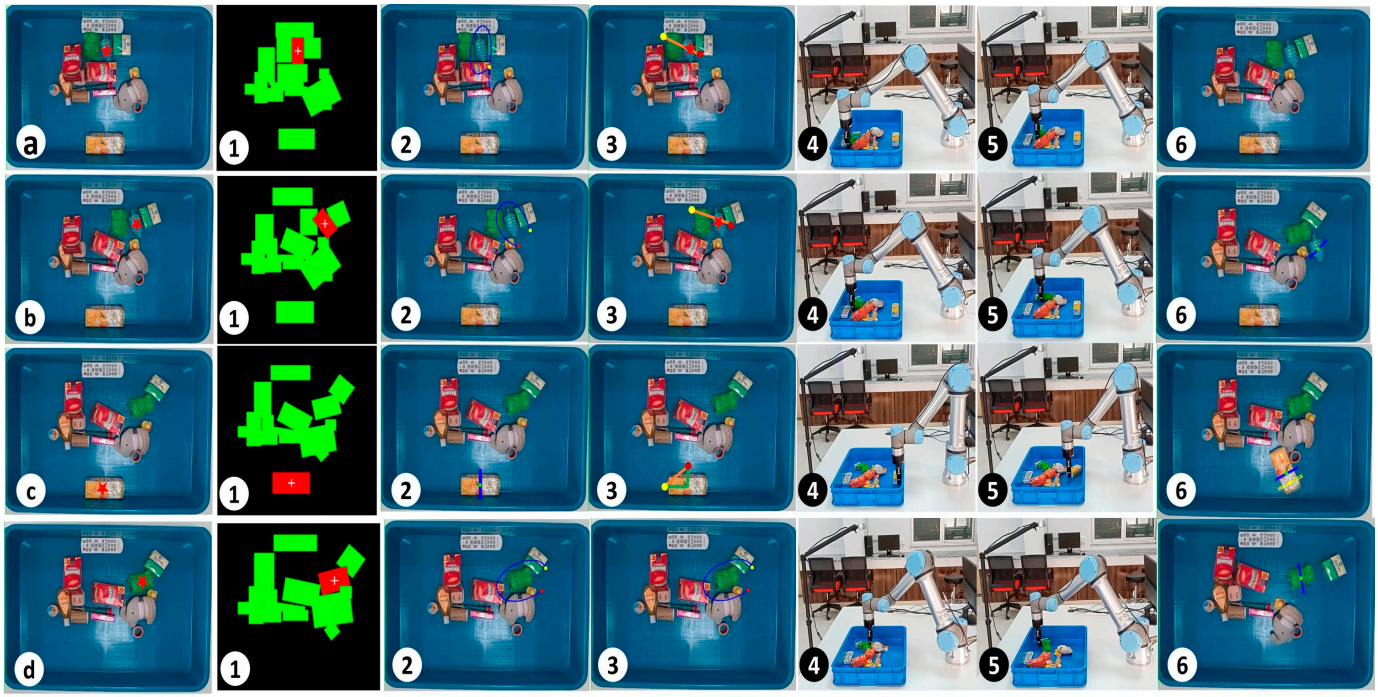


Figure 6. Real-world experiments in randomly cluttered bin environments. Each row is one trial, shown across seven panels: (**a**) initial scene with the target marked; (1) segmentation mask; (2) initially generated decision; (3) decision after wall-constraint validation; (4) robotic action at start; (5) robotic action at end; (6) resulting configuration after execution. Rows (**a**–**d**) show four representative trials.

In the medium-clutter condition, there were 7–9 objects in the vicinity of the target object, while the non-target objects are directly in contact with the target object and there could be partial occlusion The system successfully completed 28 of 30 trials, corresponding to a TSR of 93.3%. The average number of moves went up to 2.7. The decrease in TSR and increase in motion count when compared to the low-clutter condition suggests that more complex scenes had an impact on target accessibility. Sometimes, it was necessary to have the robot move the objects further over or perform further singulation or de-stacking movements to have an object ready for a grasp in the Z axis. The high-clutter condition was the most difficult experimental condition. It contained 10–15 non-target objects around the target object and the target could be partially or completely occluded. In this condition, the system was able to pass 26/30 trials; its TSR is 86.66%. The average number of motions was further raised to 3.4 motions/trial. This reduction in task success is due to the greater level of occlusion, the higher number of objects placed near feasible grasp locations, and the smaller level of clearance around feasible grasp locations. This made it harder to satisfy the conditions for direct grasp execution and led to an increase in the number of intermediate singulation and de-stacking actions. In general, the results show that there is a definite correlation between the level of clutter and system performance. The TSR diminished from 100.0% to 93.3% and eventually to 86.66% and the mean number of motions rose from 1.4 to 2.7, and then to 3.4, respectively. The trend shows that more cluttering makes it harder to successfully grasp an object and also makes manipulation to reveal and isolate the object more difficult.

The system still had an 86.66% TSR in the high-clutter environment, although it did not perform as well as it did in the more benign conditions. The results thus show that the proposed perception, grasp-feasibility evaluation, and singulation approach can deal with increasingly cluttered scenes and accommodate the number of manipulation actions depending on the level of difficulty of the configuration. The increase in motion count

is not only a limitation of the proposed approach; it should be considered as a natural consequence of the additional singulation and de-stacking operations needed in a dense cluttered scene.

To further compare the proposed system with established baseline methods, experiments were conducted in a high-clutter environment using 10 trials. GraspNet [15] with bin-wall constraints achieved a TSR of 40% (4/10 trials) with a mean number of motions of 4.90. In the evaluated scenarios, some GraspNet-generated grasp poses were physically infeasible due to potential interference with neighboring objects or the bin surface, resulting in additional recovery attempts and reduced task success. Target-oriented VPG (Visual Push Grasp) [20] with bin-wall constraints (VPG-WC) achieved a higher TSR of 60% (6/10 trials) and a lower mean number of motions of 3.80. Although VPG can distinguish between pushing and grasping actions, its linear pushing primitive may require multiple iterations to singulated tightly packed or partially occluded targets. A maximum of five manipulation iterations was permitted; trials exceeding this limit were considered unsuccessful.

The proposed method achieved the highest TSR of 90% (9/10 trials) with the lowest mean number of motions (3.30). This improvement is attributed to its multiple feasible grasp-point evaluation, task-specific singulation strategy, nonlinear trajectories, and wall-clearance validation. The quantitative comparison is presented in Table 4.

Table 4. Comparison of proposed method with other standard methods regarding TSR (%) and mean number of motions.

Method	Attempts (Success/Total)	TSR (%)	Mean Number of Motions
GraspNet [15] with bin-wall constraints	4/10	40	4.90
VPG-WC [20]	6/10	60	3.80
Proposed method	9/10	90	3.30

4.4. Ablation Study: Multi-Point Grasp Selection and Wall-Aware Singulation

Table 5 shows a controlled ablation study in which both the two key components of the proposed framework, i.e., multi-point grasp selection and wall-aware singulation, were ablated. The three configurations were tested in low, medium and high clutter conditions with the same experimental protocol. The key measures of evaluation were the task success rate (TSR) and mean number of robot motions. The comparison makes it possible to consider the effect of a single component and compare the proposed method with it.

The first configuration, middle grasp point with wall-unaware operation, represents a simplified baseline in which the grasp is selected using only a single middle grasp point and no explicit consideration of the surrounding bin walls is incorporated into the singulation process. The performance of this configuration under low-, medium-, and high-clutter conditions was 83.33%, 80.0%, and 66.66% TSR, respectively. These mean motion counts were 3.3, 4.1, and 6.2, respectively. The large drop in TSR with an increasing amount of clutter demonstrates that it is less effective to use a single "grasp location" when the object to grasp is partially concealed, or when there are other objects near the grasp location that limit the ability to reach it. In addition, as the number of motions increases, the grasping attempts seem to have failed or were impossible to make, and thus, more manipulation actions are needed to expose the target.

Table 5. Controlled ablation across the three clutter levels, 30 trials per configuration on matched scenes.

Method	Low Clutter (%)		Medium Clutter (%)		High Clutter (%)	
	TSR (%)	Mean Motions	TSR (%)	Mean Motions	TSR (%)	Mean Motions
Middle grasp point, wall unaware	83.33 *(25/30)*	3.3	80.0 *(24/30)*	4.1	66.66 *(20/30)*	6.2
Three grasp points, wall unaware	90.0 *(27/30)*	2.1	86.66 *(26/30)*	3.2	83.33 *(25/30)*	4.2
Three grasp points, wall aware singulation (Proposed method)	100.0 *(30/30)*	1.4	93.3 (28/30)	2.7	86.66 (26/30)	3.4

The second configuration is to have three grasp points, but not use the wall-awareness, bringing in the proposed multi-point grasp selection strategy with the wall-awareness disabled. The TSRs were 90.0%, 86.66%, and 83.33% for low, medium, and high clutter, respectively. Meanwhile, the average motion counts dropped to 2.1, 3.2, and 4.2, respectively. This improvement over the single-point baseline shows the advantages of computing a point in the region of the target as opposed to the geometric center. If a candidate is blocked or is inappropriate, another potential point can be chosen which will further improve the chances of the grasp being executed successfully. The decrease in mean motions also suggests that improved grasp point selection will also decrease the number of reattempts and the need to rearrange objects.

The third setup is the full approach proposed in this work, which is a three-point grasp selection and wall-aware singulation. The best TSR was obtained in this configuration at all clutter levels: 100.0% in low clutter, 93.3% in medium clutter, and 86.66% in high clutter. The mean motion counts were 1.4, 2.7, and 3.4, respectively. The results show that combining wall constraints with multi-point grasp evaluation further improves task reliability and manipulation efficiency.

This difference is more evident when the high clutter condition is considered: The baseline resulted in a TSR of 66.66% and a mean motion of 6.2. As a result of implementing the 3 grasp points, the TSR increased to 83.33%, and the mean motions decreased to 4.2. The further addition of wall-aware singulation increased the TSR further to 86.66%, while the mean motions went down to 3.4. This evolution implies that the two components are dealing with other failure mechanisms. The main advantage for multi-point grasp selection is the increase in the number of possible grasp positions, while the wall-aware singulation reduces the number of failures from a lack of clearance space and an unfavorable object movement towards the walls of the bin.

This is also seen under medium clutter. The single point, wall-unaware configuration, obtained 80.0% TSR, and the three grasp points yielded 86.66% TSR. When the three-point selection and wall-aware singulation were both turned on, TSR was 93.3%. At the same time the mean motion count fell from 4.1 to 3.2 to 2.7 motions. Therefore, this entire technique not only increases the likelihood of successfully completing the task but also decreases the number of manipulations needed to complete the task.

All clutter levels have consistent results with the contribution from each component provided. This shift from 1 possible middle grasp point to 3 candidate grasp points brings in a definite improvement in TSR and a decrease in motion count, highlighting the need to consider multiple grasp locations. The extra improvement achieved with the wall-aware singulation suggests that the feasibility of a grasp is not enough in constrained bin environ-

ments; the selected grasp and then the object motion should also take into consideration the clearance space present from the bin walls.

Table 5 shows that the improvement of the proposed framework cannot be attributed to either component alone; rather, multi-point grasp selection and wall-aware singulation complement each other. The best TSR and lowest mean motion count in all three clutter levels is obtained by the complete configuration. The benefit of the proposed constraints becomes more pronounced as clutter increases and workspace and grasp accessibility become more limited.

4.5. Complete Bin Grasping and Singulation System Latency Analysis

To evaluate the temporal performance of the proposed framework, the complete latency from user command to robotic-action completion was analyzed for each manipulation mode. In contrast to reporting only the segmentation inference time, the proposed analysis considers the complete processing chain, including user voice-command acquisition and offline speech to text conversion, segmentation-model inference, mode selection, bin-wall trajectory validation, and physical robotic execution. The complete system latency is therefore mode-dependent and is expressed as:

$$T_{pipeline} = T_{voice} + T_{seg} + T_{mode} + T_{Bin_wall} + T_{Robotic_action} \quad (33)$$

As shown in Table 6, the calculated mean pipeline latencies are 30.83 s for simple grasping, 29.63 s for depth grasping, 28.79 s for 2D singulation, and 22.13 s for partial/complete occlusion. The bin-wall validator contributes only 4.0 ms on average, indicating that the additional wall-safety check introduces negligible computational overhead relative to the overall manipulation cycle. The latency for 2D singulation and partial or complete occlusion removal may increase when the scene requires multiple manipulation actions. In each mode, only the mode selection and planning, bin-wall validation, and robotic action times differ. The user voice-command acquisition and recognition time and the segmentation-model inference time remain the same across all modes.

4.6. Failure Analysis

Across the 90 trials, six failures were observed: two under medium clutter and four under high clutter. The failures were primarily associated with grasp slippage, insufficient grasp space, ineffective singulation or de-stacking, and segmentation errors under severe occlusion. The failure types are summarized in Table 7. The failure types are summarized in Table 7. The most frequent failures observed were associated with problems of grasping objects. Two failures happened with soft objects when the selected grasp point was near the far end of the object, and the object could not be lifted due to lack of support and slip. Two other types of failure related to thin or low-profile objects, in which there was not enough grasp space or a reliable difference between the object and other objects in the environment.

The former was due to poor singulation/de-stacking, and the latter due to not being able to segment the target in severe clutter. To further verify the contribution of the proposed wall-aware validation, a controlled ablation was conducted by comparing wall-unaware and wall-aware configurations across low, medium, and high clutter levels (Table 8). The middle-grasp-point wall-unaware method resulted in three, five, and eight wall contacts, while the three-point wall-unaware method resulted in two, three, and five contacts for low, medium, and high clutter, respectively. In contrast, the proposed three-point wall-aware singulation method recorded zero wall contacts at all three clutter levels. These results demonstrate that multi-point grasp feasibility alone does not eliminate wall interactions; the explicit wall-aware trajectory validation is required to prevent unsafe trajectories from being executed.

The rest of the failures were mostly due to the inability to perceive material compliance and object geometry, not to the violation of the proposed geometric wall constraint.

Table 6. End-to-end latency analysis of the constraint-aware multimodal bin grasping and singulation in cluttered environments in seconds (s).

Mode	User Voice Command + Voice Recognition	Segmentation Model Inference	Mode Selection and Planning	Bin-Wall Validator	Robotic Action	Complete Pipeline Latency
Simple Grasp	5.49 ± 0.87	0.4350 ± 0.11	0.0009 ± 0.0002	0.0040 ± 0.0013	24.9 ± 2.5	30.8299
Depth Grasp	5.49 ± 0.87	0.4350 ± 0.11	0.0016 ± 0.0004	0.0040 ± 0.0013	23.7 ± 2.8	29.6306
2D Singulation	5.49 ± 0.87	0.4350 ± 0.11	3.5583 ± 0.2947	0.0040 ± 0.0013	19.3 ± 1.9	28.7873
Partial/ Complete Occlusion	5.49 ± 0.87	0.4350 ± 0.11	0.0032 ± 0.0013	0.0040 ± 0.0013	16.2 ± 2.4	22.1272

Table 7. Failure-mode distribution across medium- and high-clutter trials.

Failure Mode	Medium Clutter	High Clutter	Total
Target segmentation failure	0	1	1
Grasp slip/drop due to far-end grasp pose on soft object	1	1	2
Wall contact	0	0	0
Very thin/low-profile object	1	1	2
Ineffective singulation/de-stacking	0	1	1
Total	2	4	6

Table 8. Wall contact counts for controlled ablation across the three clutter levels.

Methods	Low Clutter	Medium Clutter	High Clutter
Middle grasp point, wall unaware	3	5	8
Three grasp points, wall unaware	2	3	5
Three grasp points, wall aware singulation (Proposed method)	0	0	0

4.7. Sensitivity Analysis of Empirical Parameters

A parameter-specific sensitivity analysis was conducted by varying each empirical parameter under the manipulation operation it directly affects, while keeping the remaining parameters unchanged, and five trials were performed for each parameter shown in Table 9. For wall clearance (δ), wall-adjacent grasping and pushing were evaluated at 20, 30, and 40 mm. At 20 mm, gripper wall collisions occurred at each trial, while 30 and 40 mm achieved 100% task success with 3.2 and 4.0 mean motions, respectively.

Thus, 30 mm provided the best balance between clearance and efficiency. For the height threshold (h_{th}), height-dependent grasping and transfer were evaluated at 5, 10, and 15 mm. The 5 mm setting achieved 100% grasp success but only 60% placement success due to object slippage. At 10 mm, both grasping and placement achieved 100%, whereas 15 mm reduced grasp success to 60% because of surrounding-object interference. Therefore, 10 mm was selected. For the de-stacking offset (ϵ_o), values of 3, 5, and 8 mm produced de-stacking success rates of 100%, 100%, and 60%, with 3.4, 2.8, and 4.2 mean motions, respectively.

Hence, 5 mm was selected. Finally, varying the cycle budget (C_b) from 3 to 5 and 7 iterations increased task success from 60% to 80% and 100%, respectively, while allowing progressively more manipulation motions. Thus, ($C_b = 5$) was selected as a practical compromise between recovery capability and manipulation effort. Overall, ($\delta = 30$) mm, ($h_{th} = 10$) mm, ($\epsilon_o = 5$) mm, and ($C_b = 5$) provided the most suitable operating points for the evaluated setup. These values are hardware- and task-dependent and may require recalibration for substantially different grippers, bin geometries, or sensing configurations.

Table 9. Quantitative sensitivity results for the empirical parameters.

Parameter	Setting	Success/Accuracy (%)	Mean Motions
Wall clearance ($\boldsymbol{\delta}$)	20 mm	Unsuccessful due to wall collision	—
	30 mm	100 TSR	3.2
	40 mm	100 TSR	4.0
Height threshold ($\boldsymbol{h_{th}}$)	5 mm	100 grasp/60 placement	—
	10 mm	100 grasp/100 placement	—
	15 mm	60 grasp/100 placement	—
De-stacking offset ($\boldsymbol{\epsilon_o}$)	3 mm	100 de-stacking	3.4
	5 mm	100 de-stacking	2.8
	8 mm	60 de-stacking	4.2
Cycle budget ($\boldsymbol{C_b}$)	3	60 TSR	3.0
	5	80 TSR	5.0
	7	100 TSR	7.0

5. Discussion

The experimental results demonstrate that the proposed framework maintains effective grasping and singulation performance as scene complexity increases, although increased clutter leads to reduced task success and a higher number of manipulation actions. The ablation study provides further insight into the contribution of the two principal mechanisms. Under high-clutter conditions, the single-point, wall-unaware baseline achieved a TSR of 66.66% with 6.2 mean motions. Introducing multi-point grasp selection increased the TSR to 83.33% and reduced the mean motion count to 4.2, suggesting that considering multiple candidate grasp locations improves the probability of identifying a feasible grasp without excessive object manipulation. Adding wall-aware singulation further increased the TSR to 86.66% while reducing the mean motions to 3.4. This improvement indicates that grasp-point availability and wall-aware object displacement address different aspects of the manipulation problem: the former improves grasp feasibility, whereas the latter helps maintain a feasible configuration when object accessibility is constrained by the bin boundary.

The wall-constraint validator operates at the planning level and checks whether planned actions and, where applicable, object displacement maintain sufficient clearance from the bin wall. Across 90 trials, no wall contacts were observed when the wall-aware validator was enabled. The remaining failures were mainly associated with grasp slippage, incomplete singulation, and segmentation errors under severe occlusion. These results indicate that further improvements should focus on grasp reliability, perception, and manipulation in dense scenes.

Existing methods such as VPG, GraspNet, and Dex-Net demonstrate effective grasping and push-grasp manipulation under their respective experimental conditions. However, their objectives and evaluation settings differ from the present wall-constrained bin-picking problem. In contrast, the proposed framework integrates grasp selection, occlusion-aware singulation, and explicit wall-constraint validation within a cluttered bin environment. Direct quantitative comparison is therefore difficult because of differences in hardware, sensing, objects, and protocols.

A key practical feature is that container awareness is implemented in the planning layer rather than through specialized end-effector hardware. The generated grasping or singulation trajectory can therefore be transferred to the robot execution layer, with robot-specific kinematics and control adapted to the selected robotic arm. This separation facilitates potential integration with different robotic platforms using standard end-effectors.

Future directions: First, the three-point grasp formulation may be less effective for highly irregular, concave, or deformable objects. Second, the current framework assumes a rectangular container with known wall coordinates and has not been evaluated with irregular or compliant containers or deeply interlocked objects. Third, soft, thin, and low-profile objects may require grasp ability models that explicitly consider material properties and object thickness.

Fourth, although the fallback trajectories maintain bin-wall clearance, inter-object collision and multi-object jamming are not explicitly modeled. Occupancy-aware collision checking could improve performance in highly dense scenes. Fifth, the current 4-DOF top-down grasp configuration may restrict manipulation of objects with significant inclination or in deeper bins; future work should investigate full 6-DOF grasp poses while jointly considering wall and inter-object constraints. Finally, the perception module uses a closed-set YOLO11 model trained on 16 object classes. Evaluation with unseen objects, broader object geometries, and alternative open-set or open-vocabulary perception models remains an important direction for future work.

6. Conclusions

This paper presented a constraint-aware multimodal framework for bin picking in which grasp feasibility, target accessibility, occlusion recovery and container geometry are resolved within a single adaptive decision loop. The architectural contribution is the common wall-constraint validator: because every trajectory passes through it regardless of the mode that generated it, wall safety is enforced structurally rather than depending on each of the five modes implementing boundary logic correctly. Two supporting mechanisms were introduced alongside it, namely a three-point grasp-feasibility check that evaluates the centroid and two principal-axis offsets, and a nonlinear singulation strategy that generates target-centered separation boundaries with a wall-safe linear fallback.

Across 90 physical trials on a UR5e cell, task success was 100.0% at low clutter (30/30), 93.3% at medium clutter (28/30) and 86.66% at high clutter (26/30), with the mean motion count rising from 1.4 to 3.4 as clutter increased. Six task failures were observed across the 90 trials; importantly, none involved wall contact. The segmentation model reached 0.992 box mAP@0.5 and 0.991 mask mAP@0.5 at 0.4350 $\pm$ 0.11 s inference time. The qualitative trials show the validator substituting a wall-safe path in cases where the initially generated trajectory was geometrically sensible but would have brought the end-effector into contact with the container.

Three directions follow from these limitations. First, a learned grasp-point predictor could replace the fixed three-point heuristic, generating candidates from local object geometry and available clearance rather than from a fixed parameterization. Second, replacing the closed class-name vocabulary with free-form natural-language target specification would

remove the requirement that the operator name a trained class. Third, the current perception module uses a closed-set YOLO11 detector trained on 16 predefined object classes. Consequently, the present evaluation does not establish performance on unseen object categories. Extending the perception module to open-set or open-vocabulary recognition is an important direction for future work. This could be achieved by incorporating foundation models such as the Segment Anything Model (SAM) with language-conditioned prompts or open-vocabulary detectors such as YOLO-World. These alternatives are compatible with the proposed downstream planning framework because it primarily requires the target object's class information and segmentation mask. Evaluation with unseen objects and diverse object categories is therefore identified as future work to further assess the generalizability of the proposed framework.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ai7110455/s1. Supplementary Video S1: Demonstration of the proposed constraint-aware robotic bin-grasping and singulation framework using the UR5e robotic system in cluttered bin environments.

Author Contributions: Conceptualization, M.A., I.U., K.A., M.F.S., and A.I.; methodology, M.A.; software, M.A., I.U., K.A., and M.F.S., validation, M.A., S.A.K., I.U., K.A., M.F.S., and A.I.; formal analysis, M.A.; investigation, M.A. and S.A.K.; resources, A.I.; data curation, M.A., I.U., K.A., M.F.S., and S.A.K.; writing—original draft preparation, M.A.; writing—review and editing, S.A.K., M.F.S., and A.I.; visualization, M.A. and M.F.S.; supervision, A.I.; M.F.S., project administration, M.F.S. and A.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding authors.

Acknowledgments: The authors confirm that in this study, ChatGPT 6.0 was used for assistance with the correction of grammatical mistakes in the manuscript. No AI tools were used for data generation, figure preparation, or interpretation of the experimental results, and all scientific content remains the original work of the authors. This research was supported by the Higher Education Commission (HEC) Pakistan under NPRU project #20-14813.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CAD	Computer-Aided Design
DOF	Degrees of Freedom
GFLOPs	Giga Floating-Point Operations per Second
IoU	Intersection over Union
mAP	mean Average Precision
RBF	Radial Basis Function
RGB-D	Red–Green–Blue and Depth
ROS	Robot Operating System
SVM	Support Vector Machine
TSR	Task Success Rate
YOLO	You Only Look Once

References

1. Kiyokawa, T.; Katayama, H.; Tatsuta, Y.; Takamatsu, J.; Ogasawara, T. Robotic waste sorter with agile manipulation and quickly trainable detector. *IEEE Access* **2021**, *9*, 124616–124631. [CrossRef]

2. Bashkirova, D.; Abdelfattah, M.; Zhu, Z.; Akl, J.; Alladkani, F.; Hu, P.; Ablavsky, V.; Calli, B.; Bargal, S.A.; Saenko, K. ZeroWaste dataset: Towards deformable object segmentation in cluttered scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2022.
3. Bogue, R. Bin picking: A review of recent developments. *Ind. Robot* **2023**, *50*, 873–877. [CrossRef]
4. Correll, N.; Bekris, K.E.; Berenson, D.; Brock, O.; Causo, A.; Hauser, K.; Okada, K.; Rodriguez, A.; Romano, J.M.; Wurman, P.R. Analysis and observations from the first Amazon Picking Challenge. *IEEE Trans. Autom. Sci. Eng.* **2018**, *15*, 172–188. [CrossRef]
5. Eppner, C.; Höfer, S.; Jonschkowski, R.; Martín-Martín, R.; Sieverling, A.; Wall, V.; Brock, O. Lessons from the Amazon Picking Challenge: Four aspects of building robotic systems. In Proceedings of the Robotics: Science and Systems (RSS), Ann Arbor, MI, USA, 18–22 June 2016.
6. Corbato, C.H.; Bharatheesha, M.; van Egmond, J.; Ju, J.; Wisse, M. Integrating different levels of automation: Lessons from winning the Amazon Robotics Challenge 2016. *IEEE Trans. Ind. Inform.* **2018**, *14*, 4916–4926. [CrossRef]
7. Kleeberger, K.; Landgraf, C.; Huber, M.F. Large-scale 6D object pose estimation dataset for industrial bin-picking. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: New York, NY, USA, 2019.
8. Gilles, M.; Chen, Y.; Winter, T.R.; Zeng, E.Z.; Wong, A. MetaGraspNet: A large-scale benchmark dataset for scene-aware ambidextrous bin picking via physics-based metaverse synthesis. In *Proceedings of the IEEE International Conference on Automation Science and Engineering (CASE)*; IEEE: New York, NY, USA, 2022; pp. 220–227.
9. Cordeiro, A.; Rocha, L.F.; Boaventura-Cunha, J.; Figueiredo, D.; Souza, J.P. A review of visual perception for robotic bin-picking. *Robot. Auton. Syst.* **2025**, *196*, 105236. [CrossRef]
10. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
11. Cheng, T.; Song, L.; Ge, Y.; Liu, W.; Wang, X.; Shan, Y. YOLO-World: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 16901–16911.
12. Yang, J.; Li, D.; Waslander, S.L. Probabilistic multi-view fusion of active stereo depth maps for robotic bin-picking. *IEEE Robot. Autom. Lett.* **2021**, *6*, 4472–4479. [CrossRef]
13. Li, X.; Cao, R.; Feng, Y.; Chen, K.; Yang, B.; Fu, C.-W.; Li, Y.; Dou, Q.; Liu, Y.-H.; Heng, P.-A. A sim-to-real object recognition and localization framework for industrial robotic bin picking. *IEEE Robot. Autom. Lett.* **2022**, *7*, 3961–3968. [CrossRef]
14. Sundermeyer, M.; Mousavian, A.; Triebel, R.; Fox, D. Contact-GraspNet: Efficient 6-DoF grasp generation in cluttered scenes. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*; IEEE: New York, NY, USA, 2021; pp. 13438–13444.
15. Fang, H.-S.; Wang, C.; Gou, M.; Lu, C. GraspNet-1Billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2020; pp. 11444–11453.
16. Breyer, M.; Chung, J.J.; Ott, L.; Siegwart, R.; Nieto, J. Volumetric grasping network: Real-time 6 DOF grasp detection in clutter. In *Proceedings of the Conference on Robot Learning (CoRL)*; PMLR: New York, NY, USA, 2021; Volume 155, pp. 1602–1611.
17. Cao, Y.; Liao, B.; Li, Y.; Zhao, Z.; Li, X. Efficient heatmap-guided 6-DoF grasp detection in cluttered scenes. *IEEE Robot. Autom. Lett.* **2023**, *8*, 4895–4902. [CrossRef]
18. Dogar, M.; Srinivasa, S. A framework for push-grasping in clutter. In *Proceedings of Robotics: Science and Systems (RSS)*; IEEE: New York, NY, USA, 2011.
19. Hermans, T.; Rehg, J.M.; Bobick, A. Guided pushing for object singulation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: New York, NY, USA, 2012; pp. 4783–4790.
20. Zeng, A.; Song, S.; Welker, S.; Lee, J.; Rodriguez, A.; Funkhouser, T. Learning synergies between pushing and grasping with self-supervised deep reinforcement learning. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: New York, NY, USA, 2018; pp. 4238–4245.
21. Sarantopoulos, I.; Kiatos, M.; Doulgeri, Z.; Malassiotis, S. Split deep Q-learning for robust object singulation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*; IEEE: New York, NY, USA, 2020; pp. 6225–6231.
22. Arsalan, M.; Imran, A.; Shah, S.S.; Malaika, S.; Iqbal, Z.A. Deep Learning Based Autonomous Robotic Grasping and Sorting System for Industry 4.0. In *Proceedings of the 2024 International Conference on Robotics and Automation in Industry (ICRAI)*; IEEE: New York, NY, USA, 2024; pp. 1–6.
23. Wang, Y.; Kasaei, H. Learning dual-arm push and grasp synergy in dense clutter. *IEEE Robot. Autom. Lett.* **2025**, *10*, 5154–5161. [CrossRef]
24. Danielczuk, M.; Kurenkov, A.; Balakrishna, A.; Matl, M.; Wang, D.; Martín-Martín, R.; Garg, A.; Savarese, S.; Goldberg, K. Mechanical search: Multi-step retrieval of a target object occluded by clutter. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*; IEEE: New York, NY, USA, 2019; pp. 1614–1621.
25. Zuo, G.; Liu, Y.; Zhang, X.; Chen, H. A graph-based deep reinforcement learning approach to grasping fully occluded objects. *Cogn. Comput.* **2023**, *15*, 36–49. [CrossRef]
26. Kiatos, M.; Sarantopoulos, I.; Doulgeri, Z.; Malassiotis, S. Learning a pre-grasp manipulation policy to effectively retrieve a target in dense clutter. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: New York, NY, USA, 2024; pp. 7543–7549.

27. Bogue, R. Growth in e-commerce boosts innovation in the warehouse robot market. *Ind. Robot* **2016**, *43*, 583–587. [CrossRef]
28. Bogue, R. Warehouse robot market boosted by Covid pandemic and technological innovations. *Ind. Robot* **2022**, *49*, 181–186. [CrossRef]
29. Kujala, J.V.; Lukka, T.J.; Holopainen, H. Classifying and sorting cluttered piles of unknown objects with robots: A learning approach. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: New York, NY, USA, 2016; pp. 971–978.
30. Morrison, D.; Tow, A.W.; McTaggart, M.; Smith, R.; Kelly-Boxall, N.; Wade-McCue, S.; Erskine, J.; Grinover, R.; Gurman, A.; Hunn, T.; et al. Cartman: The low-cost Cartesian manipulator that won the Amazon Robotics Challenge. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*; IEEE: New York, NY, USA, 2018.
31. Fujita, M.; Domae, Y.; Noda, A.; Garcia Ricardez, G.A.; Nagatani, T.; Zeng, A.; Song, S.; Rodriguez, A.; Causo, A.; Chen, I.-M.; et al. What are the important technologies for bin picking? Technology analysis of robots in competitions based on a set of performance metrics. *Adv. Robot.* **2020**, *34*, 560–574. [CrossRef]
32. Back, S.; Lee, J.; Kim, K.; Rho, H.; Lee, G.; Kang, R.; Lee, S.; Noh, S.; Lee, Y.; Lee, T. Graspclutter6d: A large-scale real-world dataset for robust perception and grasping in cluttered scenes. *IEEE Robot. Autom. Lett.* **2025**, *10*, 10498–10505. [CrossRef]
33. Leão, G.; Costa, C.M.; Sousa, A.; Veiga, G. Detecting and solving tube entanglement in bin picking operations. *Appl. Sci.* **2020**, *10*, 2264. [CrossRef]
34. Moosmann, M.; Kulig, M.; Spenrath, F.; Mönnig, M.; Roggendorf, S.; Petrovic, O.; Bormann, R.; Huber, M.F. Separating entangled workpieces in random bin picking using deep reinforcement learning. *Procedia CIRP* **2021**, *104*, 881–886. [CrossRef]
35. Moosmann, M.; Spenrath, F.; Rosport, J.; Melzer, P.; Kraus, W.; Bormann, R.; Huber, M.F. Transfer learning for machine learning-based detection and separation of entanglements in bin-picking applications. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: New York, NY, USA, 2022; pp. 1123–1130.
36. Zhang, X.; Domae, Y.; Wan, W.; Harada, K. Learning efficient policies for picking entangled wire harnesses: An approach to industrial bin picking. *IEEE Robot. Autom. Lett.* **2023**, *8*, 73–80. [CrossRef]
37. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y.; et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2023; pp. 4015–4026.
38. Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. SAM 2: Segment anything in images and videos. *arXiv* **2024**, arXiv:2408.00714.
39. Li, M.; Zhao, Q.; Lyu, S.; Wang, C.; Ma, Y.; Cheng, G.; Yang, C. OVGNet: A unified visual-linguistic framework for open-vocabulary robotic grasping. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: New York, NY, USA, 2024; pp. 7507–7513.
40. Tziafas, G.; Kasaei, H. Towards open-world grasping with large vision-language models. In Proceedings of the Conference on Robot Learning (CoRL), Munich, Germany, 6–9 November 2024.
41. Fu, X.; Miao, L.; Ohnishi, Y.; Hasegawa, Y.; Suwa, M. A low-cost, high-speed, and robust bin picking system for factory automation enabled by a non-stop, multi-view, and active vision scheme. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: New York, NY, USA, 2024; pp. 11566–11573.
42. Li, E.; Feng, H.; Zhang, S.; Fu, Y. Learning target-oriented push-grasping synergy in clutter with action space decoupling. *IEEE Robot. Autom. Lett.* **2022**, *7*, 11966–11973. [CrossRef]
43. Arsalan, M.; Imran, A.; Arif, A.; Khan, K.; Siddiqi, M.R. Deep learning-based assistive robotic grasping system for visually impaired individuals. *IEEE Access* **2026**, *14*, 60205–60228. [CrossRef]
44. Wu, L.; Chen, X.; Zhang, Z.; Wang, Y.; Li, J. Efficient push-grasping for multiple target objects in clutter environments. *Front. Neurorobot.* **2023**, *17*, 1188468. [CrossRef] [PubMed]
45. Yang, Y.; Liang, H.; Choi, C. A deep learning approach to grasping the invisible. *IEEE Robot. Autom. Lett.* **2020**, *5*, 2232–2239. [CrossRef]
46. Huang, H.; Danielczuk, M.; Kim, C.M.; Ichnowski, J.; Angelova, A.; Ichter, B.; Goldberg, K. Visual foresight trees for object retrieval from clutter with nonprehensile rearrangement. *IEEE Robot. Autom. Lett.* **2022**, *7*, 1919–1926. [CrossRef]
47. Du, W.; Yang, Z.; Zhang, Y.; Chen, X.; Wang, J. Multi-teachers distillation for robust target-oriented push-grasping. *IEEE Robot. Autom. Lett.* **2024**, *9*, 10303–10310.
48. Xu, L.; Liu, Z.; Gui, Z.; Guo, J.; Jiang, Z.; Zhang, T.; Xu, Z.; Gao, C.; Shao, L. DexSinGrasp: Learning a unified policy for dexterous object singulation and grasping in cluttered environments. *IEEE Robot. Autom. Lett.* **2025**, *11*, 1346–1353. [CrossRef]
49. Murali, A.; Mousavian, A.; Eppner, C.; Paxton, C.; Fox, D. 6-DOF grasping for target-driven object manipulation in clutter. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*; IEEE: New York, NY, USA, 2020; pp. 6232–6238.
50. Huang, H.; Dharmarajan, V.; Adebola, S.; Shivakumar, K.; Danielczuk, M.; Ichnowski, J.; Goldberg, K. Mechanical search on shelves with efficient stacking and destacking of objects. *Springer Proc. Adv. Robot.* **2023**, *27*, 205–221. [CrossRef]

51. Wu, Z.; Liang, Y.; Liu, X.; Li, J. Prioritized planning for target-oriented manipulation via hierarchical stacking relationship prediction. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: New York, NY, USA, 2023; pp. 4873–4880.
52. Dengler, N.; Großklaus, D.; Bennewitz, M. Learning goal-oriented non-prehensile pushing in cluttered scenes. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: New York, NY, USA, 2022.
53. Son, Y.G.; Um, S.; Hong, J.; Bui, T.H.; Choi, H.R. Corner-grasp: Multi-action grasp detection and active gripper adaptation for grasping in cluttered environments. *arXiv* **2025**, arXiv:2504.01861.
54. Torrado, P.; Levin, J.; Grotz, M.; Smith, J. TetraGrip: Sensor-Driven Multi-Suction Reactive Object Manipulation in Cluttered Scenes. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: New York, NY, USA, 2025; pp. 6534–6540.
55. Boschetti, G.; Sinico, T.; Trevisani, A. Improving robotic bin-picking performances through human–robot collaboration. *Appl. Sci.* **2023**, *13*, 5429. [CrossRef]
56. Cao, H.; Yang, Z.; Zhang, Y.; Wang, J. Two-stage grasping: A new bin picking framework for small objects. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*; IEEE: New York, NY, USA, 2023; pp. 2584–2590.
57. Khan, A.H.; Umar, F.; Ayoub, U.; Khan, M.U.R.; Haneef, S.; Siddique, M.F. A Heterogeneous Multi-Output Stacked Learning Framework for Mechanical Property Prediction of FDM-Printed ASA: Experimental Validation. *Polymers* **2026**, *18*, 2100. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.