machines

MDPI

Article

Material-Aware First-Grasp Target Preselection for Mixed Rigid-Deformable Technical-Waste Mock-Ups

Yongzhuo Liu [1,*], Jiangmei Zhang [1,2,*], Haolin Liu [1,2] and Yongfa Mi [1]

[1] School of Information Engineering, Southwest University of Science and Technology, Mianyang 621010, China
[2] Caca Innovation Center of Nuclear Environmental Safety Technology, Mianyang 621010, China
* Correspondence: liuyongzhuo427@mails.swust.edu.cn (Y.L.); zjm@swust.edu.cn (J.Z.); Tel.: +86-15982126252 (Y.L.)

Abstract

Robotic sorting of mixed rigid–deformable technical waste requires the selection of a graspable first target while minimizing disturbance to surrounding objects. This paper proposes a lightweight material-conditioned target-preselection method for cluttered RGB-D scenes. Object instances are detected using YOLO26n and segmented using SAM2-B, while their material categories are predicted using DPB-CNN. Depth-consistency filtering is subsequently applied to refine the candidate masks. Each candidate is evaluated according to geometric accessibility and local overlap, while visually inferred rigid-over-deformable relations are used to penalize or exclude deformable objects constrained by rigid-like objects. AnyGrasp generates a 6-DoF grasp pose using dense target points together with a downsampled workspace cloud retained for collision checking. Experiments were conducted on a UR5 platform using 40 predefined layouts covering four representative rigid–deformable interaction patterns, with three trials per method for each layout. The proposed method achieved a target selection accuracy (TSA) of 86.7%, a first-grasp success rate (FSR) of 80.8%, and a severe-disturbance rates (SDR) of 8.3%. After Holm correction, TSA was significantly higher than for both baselines, and SDR was significantly lower than for Native-AnyGrasp. The observed FSR improvements did not reach the corrected significance threshold. These results support improved first-grasp decision-making and reduced disturbance relative to Native-AnyGrasp in the evaluated technical-waste mock-up scenarios.

Keywords: robotic waste sorting; cluttered grasping; material-aware perception; target preselection; rigid–deformable interaction; nuclear technical-waste mock-ups

Academic Editors: Quang Minh Ta and Tan Chen

Received: 2 September 2026
Revised: 23 September 2026
Accepted: 28 September 2026
Published: 29 September 2026

1. Introduction

The safe and efficient sorting and disposal of technical waste generated during the operation and decommissioning of nuclear power plants (NPPs) is a critical challenge in the nuclear industry. To minimize human exposure to hazardous and potentially radioactive environments, robotic manipulators have attracted increasing interest for reducing direct personnel involvement in waste-sorting operations [1–5]. However, unlike standard industrial bin-picking tasks that deal with homogeneous or well-defined objects, robotic grasping in NPP scenarios faces exceptional challenges due to the highly unstructured, densely cluttered, and physically heterogeneous nature of the waste.

Specifically, NPP technical waste typically comprises a complex mixture of rigid tools (e.g., wrenches, screwdrivers) and highly deformable materials (e.g., protective clothing, rubber gloves, masks). In such dense clutter, items are frequently stacked, occluded, or

even entangled. Executing a grasp without considering spatial hierarchy and material properties may drag, rotate, or topple neighboring items, thereby increasing scene uncertainty and complicating subsequent manipulation. Therefore, selecting a reasonable first-grasp target—strictly following the principle of "minimal environmental disturbance"—is as crucial as the grasp execution itself.

Current robotic grasping methodologies generally fall into two categories: heuristic-based approaches and data-driven generative models. Heuristic methods typically rely on geometric rules, such as prioritizing the highest point along the *Z*-axis or the largest visible area. While computationally efficient, these rules often fail when handling deformable items, as the highest point of a crumpled protective suit may just be an unstable fold. On the other hand, recent end-to-end data-driven models, such as AnyGrasp [6] and Dex-Net [7], excel at proposing 6-DoF grasp poses based on local geometric features and achieve impressive zero-shot success rates on singulated objects. Related work also includes point-cloud grasp detection, GraspNet, Contact-GraspNet and graspness-based methods, which broaden the available approaches to grasp synthesis in clutter [8–12]. However, a locally high-quality grasp does not necessarily account for the material-dependent extraction risks considered in this study. Executing such a grasp can still move neighboring objects when the selected target is constrained by the surrounding arrangement.

1.1. Robotic Grasping in Clutter

Robotic grasping in densely cluttered environments has been extensively studied. Traditional analytical methods focus on modeling precise contact forces and kinematics, which often struggle to scale to unstructured industrial scenes. Data-driven grasp-generation methods such as Dex-Net, AnyGrasp, GraspNet, Contact-GraspNet, and related approaches [6–12] use learned models to detect or evaluate candidate grasps. These methods provide effective grasp proposals, while the extraction risk of a chosen object remains relevant in mixed rigid–deformable clutter. A confident grasp proposal can still correspond to an object whose removal disturbs neighboring items. The present method uses AnyGrasp for grasp-pose generation and adds a preceding decision layer to select the first-grasp target.

1.2. Material-Aware and Deformable Object Manipulation

Object properties affect the grasping strategy, particularly when a parallel-jaw gripper handles both rigid tools and deformable protective items. The present study considers clutter containing both types of object. A deformable region may be visually exposed while remaining constrained by a neighboring rigid tool; selecting that region can disturb other objects during extraction. Target preselection therefore considers material-related attributes together with geometric exposure and local object relations. The aim is to distinguish a locally graspable surface from an object that can be removed without pulling or toppling its neighbors. The proposed framework uses these attributes to adjust candidate priorities and suppress deformable targets that are visually inferred to be constrained by upper rigid-like objects.

1.3. Spatial Reasoning and Safe Target Selection

To overcome the limitations of blindly executing the highest-score grasp, recent studies have begun incorporating scene understanding into task planning. Visual pushing and grasping methods can reduce clutter, but in nuclear technical-waste scenarios unnecessary pushing motions are undesirable because they may aggravate environmental disturbance. Other approaches exploit semantic segmentation, scene graphs, or reasoning over object relations [13–21], while our study focuses on a compact target-scoring layer: candidate masks are first refined using depth cues, and target choice is then made by a

compact material-aware scoring model that explicitly penalizes overlap, visually inferred constraints, and risky rigid–deformable interactions before forwarding the selected target to the AnyGrasp-based grasp executor used in this study.

To address these limitations, this paper proposes a lightweight material-aware target preselection framework that decouples the high-level decision of which object should be grasped first from the low-level 6-DoF grasp-pose generation stage. Instead of directly executing the locally highest-confidence grasp, the robot first reasons over candidate instances extracted from RGB-D observations. Candidate masks are refined using depth cues, and each candidate is then evaluated through an interpretable scoring model that combines visual confidence, depth-based exposure, boundary richness, material priors, overlap suppression, and pressed-object inhibition. Deformable targets visually inferred to be constrained by upper rigid-like objects are explicitly penalized or blocked, after which the selected target is forwarded to AnyGrasp for grasp-pose generation. Target-driven manipulation and scene-relation reasoning already connect target choice with grasp execution [16–21]. The specific contribution here is a compact material-dependent scoring and hard-blocking rule for choosing an acceptable first object in mixed rigid–deformable clutter. The main contributions of this paper are summarized as follows:

1. Material-Conditioned First-Target Selection: We propose a practical framework for robotic sorting in mixed rigid–deformable clutter, using material-dependent scoring and blocking rules before low-level grasp-pose generation.
2. Lightweight Material-Aware Scoring: We introduce an interpretable target scoring mechanism that integrates detection confidence, depth-based exposure, boundary richness, material priors, overlap suppression, and pressed-object penalties.
3. Rigid-over-Deformable Blocking Strategy: We design a suppression strategy that explicitly discourages selecting deformable targets when they are visually inferred to be constrained by upper rigid-like objects, thereby reducing unsafe first-grasp choices and severe scene disturbances.
4. Robotic Mock-Up Evaluation: Experiments on a UR5-based sorting platform demonstrate that the proposed target-preselection layer improves first-grasp target selection and reduces severe scene disturbance in the evaluated mixed rigid–deformable layouts.

Section 2 describes the proposed method, Section 3 presents the experiments and results, Section 4 discusses the findings and limitations, and Section 5 concludes the paper.

2. Materials and Methods

2.1. Methodology

The overall pipeline is shown in Figure 1. The proposed framework decouples high-level target decision-making from low-level grasp-pose generation. Given an aligned RGB-D observation of a cluttered workspace, the system first extracts a set of candidate instances, then refines these candidates using depth-guided filtering, evaluates their grasping priority through a lightweight material-aware scoring function, and finally forwards the selected target to an AnyGrasp 6-DoF grasp generator. The overall objective is not to directly maximize local grasp confidence, but to identify a reasonable first-grasp target under mixed rigid–deformable clutter. Here, lightweight refers to the rule-based scoring layer, not to a measured runtime advantage of the complete perception and grasping pipeline.

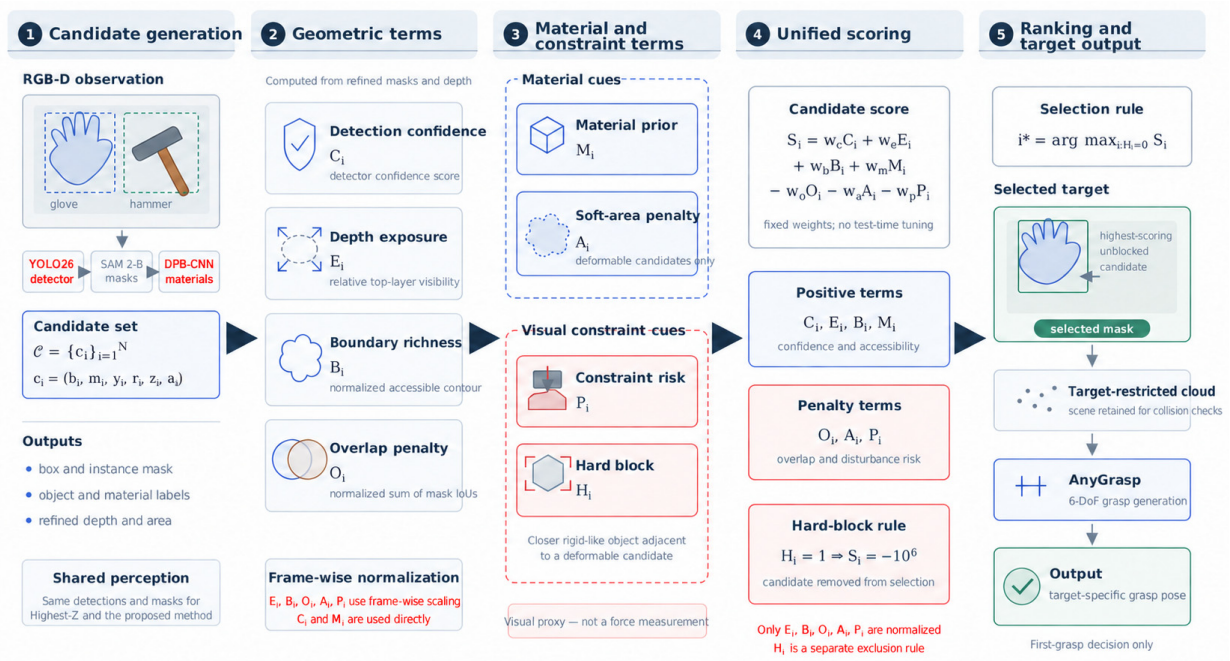

Figure 1. Overall pipeline of the proposed lightweight material-aware target preselection framework. Here i* denotes the selected index and c_{i^*} the selected candidate. Frame-wise normalization applies to E_i, B_i, O_i, A_i and P_i; C_i and M_i are used directly, and H_i is a separate exclusion rule.

2.2. Candidate Perception and Instance Proposal

Let the RGB image be denoted by **I** and the aligned depth image by **D**. From the RGB-D input, the perception module generates a set of candidate instances

$$\mathbf{C} = \{\mathbf{c_1}, \mathbf{c_2}, \ldots, \mathbf{c_N}\} \tag{1}$$

Each candidate c_i is represented as

$$\mathbf{c_i} = (\mathbf{b_i}, \mathbf{m_i}, \mathbf{y_i}, \mathbf{r_i}, \mathbf{z_i}, \mathbf{a_i}) \tag{2}$$

where b_i is the 2D detection box, m_i is the binary instance mask, y_i is the semantic class label, r_i is the predicted material label, z_i is the representative depth of the instance, and a_i is the visible mask area. N is the number of candidate instances; c_i denotes a candidate, whereas C_i in the scoring function denotes detection confidence.

2.3. Depth-Guided Candidate Refinement

In cluttered scenes, raw segmentation masks often contain mixed regions from multiple stacked objects. This effect is particularly problematic when a large deformable object overlaps with a rigid object or when two deformable objects touch each other. To reduce such cross-layer contamination, each candidate mask is refined using depth information.

For a candidate $\mathbf{c_i}$, let the valid depth set inside the mask be

$$\boldsymbol{Z_i} = \{\boldsymbol{D(p)} \mid \boldsymbol{p} \in \boldsymbol{m_i}, \boldsymbol{D(p)} > 0\} \tag{3}$$

The front-layer reference is the 20th percentile of positive depths within the original mask. With depth expressed in millimeters, pixels are retained when $0 < D(p) \leq z_{ref} + 18$ mm. The binary mask then undergoes one opening and one closing operation with a 3×3-pixel square kernel. The largest 8-connected component is retained if it contains at least 120 pixels; otherwise the component-filter input is kept. If fewer than 120 valid pixels

are available, or the depth-filtered or final mask contains fewer than 120 pixels, refinement returns the original mask.

The representative depth z_i used for scoring is the median positive depth within the resulting mask. A candidate with no positive depth samples is skipped. This separates the front-layer reference used for mask filtering from the median used in the exposure term.

2.4. Lightweight Material-Aware Target Preselection

After candidate refinement, the framework assigns each candidate a priority score that reflects both graspability and scene safety. Rather than building a complex global graphical model, w_e use a lightweight and interpretable scoring formulation tailored to mixed rigid–deformable clutter.

For each candidate $\mathbf{c_i}$, the final score is defined as

$$\mathbf{S_i} = \mathbf{w_c C_i} + \mathbf{w_e E_i} + \mathbf{w_b B_i} + \mathbf{w_m M_i} - \mathbf{w_o O_i} - \mathbf{w_a A_i} - \mathbf{w_p P_i} \quad (4)$$

Here, C_i denotes detection confidence, E_i denotes depth-based exposure, B_i describes local boundary richness, M_i is a class-aware and material-aware prior, O_i measures overlap with surrounding candidates, A_i penalizes large but poorly graspable deformable areas, P_i estimates the visually inferred risk of the candidate being pressed by another object, and H_i is a hard blocking term that is activated when a deformable target is significantly constrained by an upper rigid object. C_i is used directly as detector confidence, and M_i is obtained from the fixed lookup. E_i, B_i, O_i, A_i and P_i use per-frame normalization; H_i is a separate hard-blocking indicator and has no weight. The weights and thresholds in Table 1 were engineering settings established during system development. No independent pilot set was used; local sensitivity is reported in Section 3.7. For a candidate satisfying the hard-blocking condition $H_i = 1$, its score is set to -10^6 before ranking; otherwise, S_i is computed using Equation (4).

$$N(x_i) = \frac{x_i - min(x)}{max(x) - min(x)}, \quad max(x) - min(x) \geq 10^{-6}$$

$$N(x_i) = 0.5, \quad otherwise$$

$$E_i = 1 - N(z_i), \; B_i = N(\frac{L_i}{\sqrt{a_i}})$$

$$O_i = N(\Sigma_{j \neq i} IoU(m_i, m_j)), \; A_i = s_i N(a_i), \; P_i = N(\rho_i)$$

L_i is the sum of closed external contour lengths, s_i is one for a deformable candidate and zero otherwise, and ρ_i is the raw pressed-object risk. Area normalization uses all candidates before applying s_i. IoU denotes intersection over union.

Candidate objects were detected using the YOLO26n object-detection model [22], with the task checkpoint best.pt from the yolo26n_det_exp13 run. The detector dataset contains 700 training, 200 validation and 100 test images. Training was performed using Ultralytics v8.4.0 on an NVIDIA GeForce RTX 4090 GPU. Training was configured for 100 epochs with a batch size of 64. AdamW was used with an initial learning rate of 0.00125. The remaining training settings were left at their implementation defaults, and no additional hyperparameter tuning was performed for the detector. Its detection head directly predicts bounding boxes, class labels and confidence scores. Both training and inference used an image-size setting of 640 pixels. In all experiments, the detection confidence threshold was set to 0.80, followed by class-agnostic non-maximum suppression with an IoU threshold of 0.65. The retained boxes serve as spatial prompts for SAM2-B [23], which generates an instance mask per candidate using its default pretrained configuration without task-

specific adjustment. No text prompt is used. The cutoffs in Table 1 are separate offline filtering conditions.

The material attribute (r_i) was predicted from the cropped RGB region using DPB-CNN [24]. The ten-class DPB-CNN model used in the reported experiments assigns each candidate to one of ten predefined material categories. No material label was manually assigned during testing. DPB-CNN combines Bayesian and deterministic CNN branches through uncertainty-aware adaptive fusion. The OCC-10 dataset contains 1000 images split into 700 training and 300 test images, with the classes fabric, foliage, glass, leather, metal, paper, plastic, stone, water and wood. Each branch uses three 3×3 convolutional layers with 32, 64 and 128 channels. The model study uses Adam with a learning rate of 10^{-5}, KL annealing, and 10 Monte Carlo samples during training and 30 for uncertainty estimation. The source manuscript remains under review. Images from the 40 grasp layouts were not used for model training, fine-tuning or model selection. The material-prior score (M_i) was then obtained using a fixed mapping function ($M_i = f(y_i,r_i)$), which combines the detected object class and the predicted material category. The perception settings were shared by Highest-Z and Ours. Native-AnyGrasp generates grasps from workspace geometry without candidate-level preselection.

M_i first uses case-insensitive object-class matching: hammer/tool/screwdriver and helmet/hardhat/hard_hat receive 0.85, shoe 0.75, glove 0.42, and mask/cloth/fabric 0.35, in that order. If no class rule matches, the ten material outputs use metal 0.85, plastic/wood/leather 0.80, fabric 0.35, and foliage/glass/paper/stone/water 0.55. These are fixed engineering priorities, not probabilities of grasp success.

The deformable flag s_i is activated by class keywords glove/mask/cloth/fabric/rag/garment or, among the ten material labels, fabric. The rigid-like flag g_i is activated by class keywords hammer/screwdriver/tool/shoe/helmet/hardhat/hard_hat or material labels wood/metal/plastic/leather. Class matching is case-insensitive; the flags describe heuristic extraction-risk categories rather than measured stiffness and need not be mutually exclusive.

Table 1. Definition and implementation parameters of the target-scoring terms.

Term	Definition	Weight/Rule
C_i	Detector confidence used directly	$w_c = 0.18$
E_i	$1 - N(z_i)$, where z_i is median refined-mask depth	$w_e = 0.18$
B_i	$N(\frac{L_i}{\sqrt{a_i}})$, where L_i is contour length and a_i is refined mask area	$w_b = 0.16$
M_i	Class-first material lookup described in the text	$w_m = 0.16$
O_i	N(Σ of mask IoUs with other candidates)	$w_o = 0.07$
A_i	s_i $N(a_i)$	$w_a = 0.05$
P_i	$N(\rho_i)$, with raw risk ρ_i defined in the text	$w_p = 0.20$
H_i	$s_i = 1$, $G_i = 1$ and $\rho_i > 0.10$	No weight; score = -10^6

A key failure mode in mixed clutter arises when the highest visible part of a deformable item is selected even though the item is physically constrained by an upper rigid object. To address this issue, local contact and depth relations are estimated between nearby masks. If another candidate lies above the current candidate in the contact region and exhibits rigid-like properties, a strong penalty is applied. When this suppression exceeds a predefined threshold, the deformable candidate is directly blocked from being selected as the first-grasp target. In this way, the system explicitly avoids unsafe first-grasp choices such as pulling out a glove or cloth region from beneath a rigid tool.

For candidates i and j, define $Q_{ij} = m_i \cap \text{dilate}(m_j,K)$ and Q_{ji} analogously. A valid near-contact pair requires at least 30 pixels and 20 positive-depth samples in each region,

using a 9 × 9-pixel kernel K. Let d_{ij} and d_{ji} be the median valid depths in the respective regions. Candidate j is above i when d_{ji} + 6 mm < d_{ij}. For the set J_i of such neighbors,

$$\rho_i = min\left(1, \Sigma_{j \in J_i} \alpha_{ij} \beta_{ij} min\left(\left|Q_{ij}\right| / a_i, 1\right)\right)$$

α_{ij} is 1.35 if $s_i = 1$ and $g_j = 1$, otherwise 1.20 if $s_i = 1$ and $s_j = 0$, and 1.0 otherwise. β_{ij} is 1.10 when $a_j < a_i$ and 1.0 otherwise. Let G_i indicate the presence of a rigid-like neighbor in J_i. Then $H_i = 1$ only if $s_i = 1$, $G_i = 1$ and $\rho_i > 0.10$; otherwise $H_i = 0$. Empty masks or missing valid depth produce zero risk and no rigid-occluder flag. Blocked candidates receive a score of -10^6 and are excluded from selection; no target is returned if none is eligible.

2.5. Decoupled 6-DoF Grasp Execution

Once the optimal candidate is determined,

$$i^* = argmax_{i:H_i=0} S_i, \; selected\ candidate = c_{i*} \tag{5}$$

The refined target mask identifies the localized target point cloud. Target points define the grasp-search bounds and the target-affiliation constraint; they do not replace all surrounding geometry used for collision checking.

AnyGrasp [6] receives dense target points together with a downsampled workspace cloud. Candidate generation is restricted using target-derived bounds, and subsequent filtering retains target-affiliated grasps. Surrounding scene points provide collision context during grasp generation. This is distinct from a guarantee that all non-target objects are represented in the motion-planning collision model. Only AnyGrasp is evaluated as the grasp generator in this study.

3. Results

The proposed target-preselection framework was evaluated in real-robot experiments on mixed rigid-deformable cluttered scenes, focusing on first-target choice, grasp success, and severe disturbance. The experiments were designed to answer three questions:

1. whether the proposed method can select a safer and more reasonable first-grasp target than heuristic or direct grasping baselines;
2. whether the proposed preselection strategy can improve grasp execution performance when coupled with a generic 6-DoF grasp generator;
3. whether the pressed-object inhibition and material-aware terms contribute to reducing severe scene disturbances.

Unlike standard bin-picking benchmarks that mainly evaluate local grasp-pose quality, the focus of this work is on safe target selection before grasp execution. Therefore, the experiments evaluate both target-level decision quality and execution-level grasping performance.

3.1. Experimental Setup

3.1.1. Robotic Platform

The experimental system was built on a UR5 6-DoF robotic manipulator (Universal Robots A/S, Odense, Denmark) equipped with a Robotiq 2F-85 parallel-jaw gripper (Robotiq Inc., Lévis, QC, Canada). Visual perception was provided by an Azure Kinect RGB-D camera (Microsoft Corporation, Redmond, WA, USA) mounted in an eye-to-hand configuration, which provided a stable global view of the workspace. The software system was implemented under Ubuntu 20.04 and ROS Noetic.

The perception module took aligned RGB-D images as input. Object candidates were generated by a detector and refined by an instance segmentation model. The material-

aware target preselection module then evaluated all visible candidates and selected one target for grasping. Dense points from the selected target, together with a downsampled workspace cloud retained for collision checking, were sent to AnyGrasp for 6-DoF grasp-pose generation. Target-derived bounds restricted the grasp search, while scene geometry provided collision context as described in Section 2.5. The generated grasp-pose was finally executed by the robot through the motion planning module.

3.1.2. Test Objects

To simulate mixed technical waste in nuclear power plant maintenance and decommissioning environments, the test objects included both rigid and deformable items. Rigid items included hammers, screwdrivers, shoes, and small hard-shell tools. Deformable items included rubber gloves, work gloves, masks, and cloth-like protective items.

These objects were selected because they exhibit different physical properties and grasping risks. Rigid tools usually provide stable geometric structures for parallel-jaw grasping, but they may cause severe disturbance if pulled from under deformable objects. Deformable items are lightweight and compliant, but they are difficult to grasp reliably when pressed or constrained by upper rigid objects. Therefore, this object set is suitable for evaluating rigid–deformable interaction-aware target selection. All test objects were non-radioactive mock-ups.

3.1.3. Cluttered Scene Design

A total of 40 representative cluttered layouts were designed. Instead of randomly piling objects without structure, the layouts were organized according to four typical interaction patterns observed in mixed rigid-deformable sorting tasks:

(A) rigid-over-deformable scenes, where rigid tools are placed on top of gloves or cloth-like objects;
(B) deformable-over-rigid scenes, where gloves or cloth-like objects partially cover rigid objects;
(C) adjacent-contact scenes, where multiple objects are close to or lightly touching each other without strong vertical stacking;
(D) multi-layer stacking scenes, where several rigid and deformable objects form a more complex layered structure.

Each group contained ten scenes, resulting in 40 fixed test layouts. The same layout designs were used for all methods.

Figure 2 illustrates the diversity of object arrangements, ranging from sparse to densely packed configurations, with varying degrees of occlusion and orientation.

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

Figure 2. Representative cluttered layouts used for evaluating target preselection and grasp execution. (**a**) Rigid-over-deformable; (**b**) deformable-over-rigid; (**c**) adjacent contact; (**d**) multi-layer stacking.

3.1.4. Compared Methods

Three methods were compared.

1. **Highest-Z Baseline**

This heuristic method selects the candidate with the smallest representative depth value, corresponding to the most visually exposed or highest object region from the camera view. The selected candidate is then sent to the grasp generator. This method is simple and fast, but it does not consider material properties or physical constraints.

2. **Native-AnyGrasp Baseline**

This baseline applies AnyGrasp directly to the complete workspace point cloud and selects the highest-scoring grasp that passes the grasp-generation collision check. For target selection accuracy (TSA), projection of the grasp center into the image provides an initial object association. The actual selected object is then checked manually. Missing or overlapping predicted masks do not automatically imply a TSA error; a manually identified acceptable object is scored as correct, and an unacceptable object as incorrect. This baseline uses local geometry without explicit target-level material or constraint reasoning.

3. **Ours: Material-Aware Target Preselection**

The proposed method first evaluates all candidate instances using the lightweight material-aware scoring function. The scoring function considers visual confidence, depth-based exposure, boundary richness, material priors, overlap suppression, pressed-object inhibition, and rigid-over-deformable blocking. The selected target restricts AnyGrasp candidate generation, while scene geometry is retained for collision context.

Highest-Z provides an exposure-based heuristic baseline, whereas Native-AnyGrasp represents direct geometry-based execution. Highest-Z and Ours share mask refinement, target-point-cloud extraction, retention of downsampled scene geometry for grasp-generation collision checking, and grasp-execution settings. Native-AnyGrasp searches the workspace without the same target restriction; its comparison with Ours therefore includes both target preselection and target-specific point-cloud processing.

Each method was tested on the same 40 layouts. For each layout, each method was repeated three times. Therefore, the target selection experiment contained 360 first-grasp trials in total. After each grasp, the objects were manually restored to the original arrangement. Methods were executed in the fixed order Highest-Z, Native-AnyGrasp and Ours. The records link each layout and repeat to one reference-frame identifier shared across the three methods; 120 reference frames correspond to 360 main trials. All 360 main trial records contain a selected-object identifier and complete TSA, FSR and SDR outcomes and are included in the corresponding denominators.

3.2. Evaluation Metrics

The experiments were evaluated using both decision-level and execution-level metrics.

3.2.1. Target Selection Accuracy

Target selection accuracy (TSA) measures whether the method selects a human-validated reasonable first-grasp target. For each scene, one or more reasonable first-grasp targets were manually annotated according to physical accessibility and expected disturbance risk. The acceptable target set was determined before method comparison and was applied consistently to all three methods.

Two annotators, a doctoral student and a master's student in Control Science and Engineering, independently labeled acceptable targets without viewing the methods' results. Disagreements were adjudicated by the doctoral student, who is also an author of this paper. Across 600 paired object-level decisions in 120 frames, initial agreement was 560/600 (93.3%), with Cohen's $\kappa = 0.8003$. Entire acceptable-target sets agreed in 84/120 frames (70.0%). The paired labels comprise 453 joint rejections, 107 joint acceptances, 17 pairs

labeled (0,1) and 23 labeled (1,0) by the two recorded annotator IDs, respectively. Trial-level selected-object identifiers are compared with the final acceptable-target set for the corresponding reference frame.

$$TSA = \frac{N_{target\ hit}}{N_{trials}} \tag{6}$$

where $N_{target\ hit}$ denotes the number of trials in which the selected target belongs to the annotated reasonable target set, and N_{trials} denotes the total number of trials. This metric directly evaluates the quality of high-level target decision-making.

3.2.2. First-Grasp Success Rate

First-grasp success rate (FSR) measures whether the selected target can be successfully grasped and lifted in the first attempt.

$$FSR = \frac{N_{first\ success}}{N_{trials}} \tag{7}$$

A grasp is considered successful if the target object is securely grasped, lifted, and transported away from the cluttered region.

3.2.3. Severe Disturbance Rate

Severe-disturbance rate (SDR) measures the probability that a grasp causes significant unintended movement or collapse of non-target objects.

$$SDR = \frac{N_{severe\ disturbance}}{N_{trials}} \tag{8}$$

A severe disturbance is recorded when a non-target object is dragged together with the selected object, flipped or toppled, dropped from the stack, or causes an evident partial collapse of the surrounding arrangement. The same criteria are applied to all compared methods. A trial with one or more such events is counted once, and no displacement threshold is used. FSR and SDR are recorded separately, so a successful grasp can also cause severe disturbance.

3.2.4. Statistical Analysis

Layouts were treated as the sampling units, with three repeats nested within each layout. Pointwise percentile 95% confidence intervals were obtained from 10,000 stratified layout-cluster-bootstrap resamples (seed 20260924), drawing ten layouts with replacement within each scene category and retaining all methods and repeats together [25].

The six comparisons of Ours with the two baselines on TSA, FSR and SDR used exact two-sided paired layout-level sign-flip tests, with Holm correction at a family-wise significance level of 0.05 [26]. This inference assumes independent layouts and sign-exchangeability under the null. Confidence intervals are pointwise, not multiplicity-adjusted; secondary analyses are descriptive. Fixed method order and manual restoration remain potential sources of bias. Each layout contributes three binary outcomes per method. The paired test uses within-layout differences in the corresponding counts (0–3), computes the exact null distribution over both signs for every nonzero difference, and compares absolute sums. The six resulting p-values are adjusted together. These calculations do not remove possible fixed-order effects.

3.3. Quantitative Results

3.3.1. First-Grasp Target Selection Results

Table 2 reports the first-grasp target selection results over the 40 representative cluttered scenes. The proposed method achieves the highest TSA among all compared methods, supporting improved acceptable-target selection in the evaluated layouts.

Table 2. First-grasp outcomes across 40 layouts, with 120 trials per method. (**a**) Rates with pointwise 95% layout-clustered confidence intervals. (**b**) Paired differences, Ours minus baseline, in percentage points (pp), with pointwise 95% confidence intervals and Holm-adjusted *p*-values.

(a) Main outcomes

Method	TSA ↑ % [95% CI]	FSR ↑ % [95% CI]	SDR ↓ % [95% CI]
Highest-Z	68.3% [59.2, 77.5]	61.7% [53.3, 70.0]	18.3% [10.8, 26.7]
Native-AnyGrasp	60.0% [51.7, 68.3]	70.0% [61.7, 78.3]	27.5% [19.2, 35.8]
Ours	86.7% [80.8, 91.7]	80.8% [74.2, 87.5]	8.3% [4.2, 13.3]

(b) Paired comparisons

Baseline	Outcome	Difference [95% CI] pp	Holm p
Highest-Z	TSA	18.3 [8.3, 28.3]	0.0103
Highest-Z	FSR	19.2 [7.5, 30.8]	0.0539
Highest-Z	SDR	−10.0 [−19.2, −1.7]	0.0954
Native-AnyGrasp	TSA	26.7 [17.5, 35.8]	<0.0001
Native-AnyGrasp	FSR	10.8 [0.0, 21.7]	0.0987
Native-AnyGrasp	SDR	−19.2 [−27.5, −11.6]	0.0007

As shown in Table 2, the Highest-Z baseline achieves a higher TSA and lower SDR than Native-AnyGrasp, indicating that a simple top-layer selection strategy is already useful for reducing risky first-grasp decisions in many cluttered scenes. Its lower observed FSR may reflect that the most exposed region is not always physically graspable, especially for deformable objects such as gloves and cloth-like items.

Native-AnyGrasp has a higher observed FSR than Highest-Z, which may reflect its direct ranking of grasps by local grasp quality. Nevertheless, it produces the highest SDR among the three methods, showing that local geometric grasp confidence alone does not always prevent non-target disturbance in the evaluated layouts.

The proposed method achieves the best observed values across the three metrics. It improves target selection accuracy over Highest-Z while maintaining a much lower severe-disturbance rates than Native-AnyGrasp. These observations support improved first-target choice and lower scene disturbance in the evaluated layouts.

After Holm correction, TSA was significantly higher than for both baselines, and SDR was significantly lower than for Native-AnyGrasp. The observed FSR improvements over both baselines and SDR reduction relative to Highest-Z did not reach the corrected significance threshold (Table 2b).

3.3.2. Results by Scene Type

To further analyze the behavior of different methods, the 40 layouts were grouped into four scene types. The results are shown in Table 3.

Table 3. Target selection accuracy under different scene types.

Scene Type	Highest-Z	Native-AnyGrasp	Ours
Rigid-over-deformable	53.3%	43.3%	90.0%
Deformable-over-rigid	70.0%	60.0%	90.0%
Adjacent contact	86.7%	76.7%	93.3%
Multi-layer stacking	63.3%	60.0%	73.3%
Average	68.3%	60.0%	86.7%

The largest TSA improvement is observed in rigid-over-deformable scenes. In rigid-over-deformable scenes, deformable objects may appear visually prominent while being physically constrained by an upper rigid tool. The Highest-Z baseline tends to select such deformable regions because they are close to the camera. Native-AnyGrasp may also produce a valid-looking local grasp without considering the risk of pulling a constrained object. The proposed method identifies this risky relationship through local contact and depth comparison, and suppresses deformable candidates pressed by rigid objects.

In adjacent-contact scenes, the performance gap among methods is usually smaller because physical constraints are weaker and most candidates are relatively accessible. This result suggests that the proposed method does not over-penalize easy cases, while still providing clear advantages in difficult stacked scenes.

These scene-specific results are descriptive, with ten layouts and 30 trials per method in each category.

3.4. Ablation Study

An ablation study examined the effects of removing individual components from the proposed material-aware target-preselection framework. The complete method consists of detection confidence, depth-based exposure, boundary accessibility, material prior, overlap suppression, soft-area penalty, pressed-object inhibition, hard blocking, and depth-guided candidate refinement. Equations (4) and (5) define scoring and target selection, respectively. The Full Method row reuses the main Ours trials; the seven ablations contain 840 additional trials. Removing the material prior disables w_m but retains material-dependent flags; removing the pressed penalty disables w_p but retains hard blocking, while removing hard blocking retains P_i. The ablations are descriptive component comparisons. Because object-class rules take precedence in M_i and material-dependent flags remain active, this ablation does not isolate the contribution of the DPB-CNN network or remove all material information.

The quantitative ablation results are summarized in Table 4.

Because each variant independently removes a different component from the full method, the rows do not represent a cumulative ablation sequence and are therefore not expected to show a monotonic performance decrease from top to bottom. As shown in Table 4, the full method has the best observed values across the three metrics. Removing the pressed-object penalty causes the largest increase in SDR, from 8.3% to 25.0%, consistent with a contribution from contact- and depth-based suppression in these trials. Removing hard blocking also increases SDR to 19.2%, showing a higher observed disturbance rate when this exclusion rule is removed.

The auxiliary-term ablations also show differences in the observed outcomes. Removing the overlap penalty raises SDR from 8.3% to 16.7%, while removing the soft-area penalty raises SDR to 13.3%. Removing the boundary term lowers FSR from 80.8% to 75.8%. These descriptive comparisons are consistent with contributions from overlap suppression, deformable-area penalization and boundary information within the complete implementation; they do not independently establish the physical cause of each failure.

Table 4. Ablation study of the proposed target-preselection framework. Each variant independently removes only the indicated component from the full method, while all other components remain unchanged (120 trials per variant). Entries are percentages [pointwise 95% layout-clustered confidence interval].

Variant	TSA ↑	FSR ↑	SDR ↓
Full Method	86.7% [80.8, 91.7]	80.8% [74.2, 87.5]	8.3% [4.2, 13.3]
w/o Material Prior M_i	77.5% [68.3, 85.8]	75.0% [68.3, 81.7]	14.2% [7.5, 20.8]
w/o Pressed Penalty P_i	69.2% [62.5, 75.8]	69.2% [60.8, 77.5]	25.0% [17.5, 33.3]
w/o Hard Blocking H_i	75.0% [69.2, 80.8]	72.5% [64.2, 80.0]	19.2% [12.5, 26.7]
w/o Overlap Penalty O_i	80.0% [73.3, 85.8]	75.8% [67.5, 83.3]	16.7% [11.7, 21.7]
w/o Soft-Area Penalty A_i	82.5% [76.7, 87.5]	76.7% [70.0, 83.3]	13.3% [7.5, 20.8]
w/o Boundary Term B_i	81.7% [73.3, 89.2]	75.8% [67.5, 83.3]	12.5% [6.7, 19.2]
w/o Depth Refinement	72.5% [65.0, 79.2]	66.7% [59.2, 74.2]	22.5% [15.8, 30.0]

The variant without depth-guided refinement shows a clear performance drop in both TSA and FSR. This variant disables mask postprocessing, not the exposure term E_i; mask-dependent features and the target point cloud can therefore change. A possible explanation is that raw segmentation masks contain mixed regions from multiple stacked objects, affecting score computation and the target point cloud. Overall, these results describe the effects of individual component removals within the evaluated implementation.

3.5. Qualitative Case Study

Figure 3 compares first-grasp target selection and generated grasp poses for two representative cluttered layouts. Each pair of columns corresponds to one layout, and the rows show Highest-Z, Native-AnyGrasp and Ours.

In the left example, a hammer lies above a rubber glove. Highest-Z selects the exposed glove region, whereas Ours selects the hammer. This difference is consistent with the pressed-object penalty and hard-blocking rule, which discourage selecting a lower deformable candidate when an upper rigid-like neighbor is identified.

In the right example, Highest-Z also selects a deformable candidate, whereas Ours selects a rigid tool. Native-AnyGrasp generates grasps from the workspace point cloud without a separate target-preselection stage, as shown by its grasp-pose visualizations.

These examples illustrate differences in target choice and generated grasp poses. Grasp execution success and severe disturbance are assessed separately using the trial-level results in Section 3.3.

3.6. Perception Evaluation

Perception was evaluated offline using saved predictions for the first reference frame (R1) from each of the 40 layouts, with 200 annotated objects. Predictions were processed in descending confidence order and matched one-to-one to an unmatched object of the same class with the highest box IoU, provided that IoU ≥ 0.50. Table 5 reports the export cutoff of 0.05 and filtering of the same saved predictions at 0.25 and 0.80.

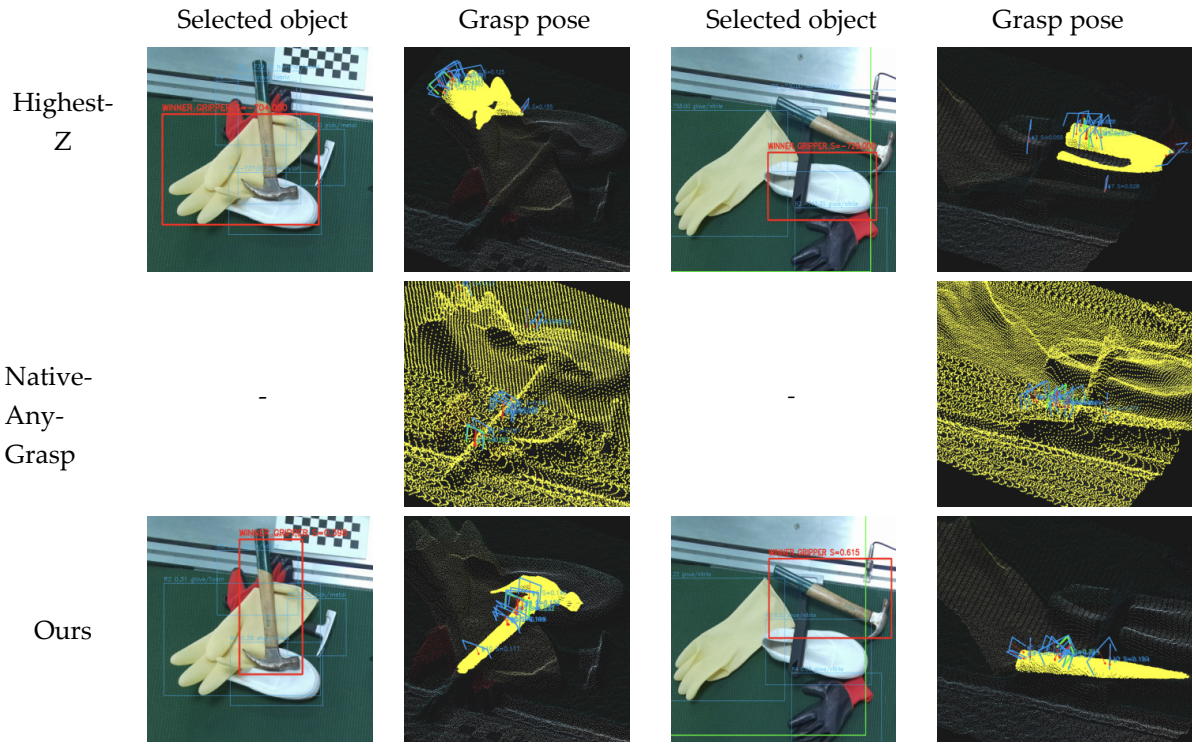

Figure 3. Qualitative comparison of target selection and generated grasp poses in two cluttered layouts. Each pair of columns shows the selected-object overlay and grasp-pose visualization for one layout. The dashes for Native-AnyGrasp indicate that it has no separate target-preselection output; they do not denote grasp failure.

Table 5. Offline perception evaluation. P and R are precision and recall; F1 is their harmonic mean. Material accuracy is conditional on correctly matched detections. The 0.25 and 0.80 rows filter saved predictions without rerunning inference or NMS.

Cutoff	TP/FP/FN	P%	R%	F1%	Material Correct
0.05	190/40/10	82.6	95.0	88.4	178/190 (93.7%)
0.25	181/32/19	85.0	90.5	87.7	170/181 (93.9%)
0.80	160/0/40	100.0	80.0	88.9	150/160 (93.8%)

At cutoff 0.05, material recognition was correct for 178/190 matched detections (93.7%). The evaluation subset contains five annotated object classes—hammer, shoe, glove, cloth and mask—and three material labels: metal, leather and fabric. Material accuracy is conditional on correctly matched detections; it is not end-to-end accuracy over all 200 objects or a ten-class benchmark. No separate occlusion-stratified accuracy is claimed.

Increasing the cutoff to 0.25 changes object recall from 95.0% to 90.5%; at 0.80, precision is 100.0% and recall is 80.0% in this subset. All three cutoffs retain at least one correctly matched acceptable candidate in each of the 40 frames. Candidate coverage and TSA measure different quantities: having an acceptable candidate available does not ensure that the ranking selects it. The base sensitivity setting uses the same R1 candidate caches as the main Ours trials. These saved-frame results do not establish candidate availability for every robot observation, and missed rigid occluders remain a limitation of the contact-based risk estimate.

3.7. Parameter Sensitivity

A one-at-a-time offline sensitivity analysis used the first reference frame from each of the 40 layouts. The material weight w_m, pressed-risk weight w_p and raw-risk threshold τ

were varied individually by ±20%, with other parameters unchanged and no weight renormalization. A no-target output was counted as a TSA miss.The resulting TSA performance, changes in selected targets relative to the base setting, and no-target cases are summarized in Table 6.

Table 6. Offline parameter sensitivity on 40 reference frames. Changed selections are relative to the base setting; no-target cases remain in the denominator.

Setting	TSA/40	TSA %	Changed/40	No Target/40
Base	36	90.0	0	0
$w_m \times 0.8$	37	92.5	5	0
$w_m \times 1.2$	35	87.5	1	0
$w_p \times 0.8$	35	87.5	1	0
$w_p \times 1.2$	36	90.0	2	0
$\tau \times 0.8$	36	90.0	0	0
$\tau \times 1.2$	36	90.0	0	0

TSA ranged from 87.5% to 92.5%, with up to five changed selections and a maximum reduction of 2.5 percentage points from the 90.0% base setting. No tested setting produced a no-target output. Varying the raw-risk threshold by ±20% did not change the selected object in these 40 frames. The base rate of 36/40 matches the R1 subset of the main Ours trials, whereas the overall rate of 104/120 includes all three repeats. This analysis characterizes local target-selection sensitivity, not changes in robot FSR or SDR; it does not establish broad parameter insensitivity or cross-scene robustness.

4. Discussion

This paper addressed first-grasp target selection in mixed rigid–deformable technical-waste mock-ups, where densely cluttered arrangements, partial occlusions, and cross-material stacking make direct execution of locally optimal grasps unreliable. Instead of modifying the grasp generator itself, w_e proposed a lightweight material-aware target preselection framework that operates before 6-DoF grasp execution.

The proposed framework combines candidate perception, depth-guided mask refinement, and an interpretable scoring mechanism that jointly considers visual confidence, depth-based exposure, boundary richness, material priors, overlap suppression, and pressed-object inhibition. In particular, deformable candidates visually inferred to be constrained by upper rigid objects are explicitly penalized or blocked, which discourages selection of constrained deformable targets.

Experimental results on a real robotic platform demonstrate that the proposed method improves first-grasp target selection and achieves lower observed severe-disturbance rates than Highest-Z and Native-AnyGrasp in the evaluated mixed rigid–deformable layouts. The corrected comparisons support higher TSA than both baselines and lower SDR than Native-AnyGrasp, whereas the observed FSR improvements do not reach the corrected significance threshold. These findings support using target preselection to complement local grasp-pose generation in the evaluated mock-up tasks. The present experiments evaluate only the first grasp from predefined layouts and do not represent complete-scene clearing performance.

The comparison with Native-AnyGrasp reflects both target selection and target-specific point-cloud processing, and does not establish superiority over stronger scene-aware or target-driven systems. The evaluation is limited to 40 constructed layouts; fixed method order and manual restoration can introduce order and pose effects. No independent pilot

set was used. Material recognition and visual constraint estimates can be unreliable under severe occlusion. The objects are non-radioactive mock-ups, and operation in radioactive environments was not evaluated. The task-frame material evaluation covers only three label categories. The class-first prior and retained material flags prevent attribution of the full performance gain to material-network predictions alone. Neither end-to-end runtime superiority nor complete collision-free motion in clutter is established here.

Future work will focus on improving candidate perception in scenes containing many highly similar deformable instances, integrating stronger instance separation mechanisms for same-class objects, and combining visual reasoning with tactile or force feedback to further improve robustness during physical manipulation.

5. Conclusions

A material-aware preselection layer improves first-target choice in the evaluated mixed rigid–deformable mock-ups. Across 40 layouts, the method achieved TSA of 86.7%, FSR of 80.8% and SDR of 8.3%. After Holm correction, TSA was significantly higher than for both baselines, and SDR was significantly lower than for Native-AnyGrasp. The observed improvements in FSR over the two baselines and in SDR over Highest-Z did not reach the corrected significance threshold. These conclusions concern the first grasp within the tested setting and do not establish complete-scene clearing performance.

Author Contributions: Conceptualization, Y.L. and H.L.; Methodology, Y.L., J.Z. and H.L.; Software, Y.L.; Validation, Y.L.; Formal analysis, Y.L.; Investigation, Y.L.; Resources, Y.L.; Data curation, Y.L. and H.L.; Writing—original draft, Y.L.; Writing—review & editing, Y.L.; Visualization, Y.L. and Y.M.; Supervision, Y.L. and Y.M.; Project administration, J.Z.; Funding acquisition, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: Nuclear Facility Decommissioning and Radioactive Waste Management Research Project (TCKY-2024-CICDR-029).

Informed Consent Statement: Not applicable.

Data Availability Statement: Trial-level outcomes, layout and repeat identifiers, paired target annotations, and the analysis script used for the reported confidence intervals and tests are available from the corresponding author on reasonable request.

Conflicts of Interest: All authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

References and Note

1. Poozhiyil, M.; Argin, O.F.; Rai, M.; Esfahani, A.G.; Hanheide, M.; King, R.; Saunderson, P.; Moulin-Ramsden, M.; Yang, W.; Palacio García, L.; et al. A framework for real-time autonomous robotic sorting and segregation of nuclear waste: Modelling, identification and control of Dexter Robot. *Machines* **2025**, *13*, 214. [CrossRef]
2. Vitanov, I.; Farkhatdinov, I.; Denoun, B.; Palermo, F.; Otaran, A.; Brown, J.; Omarali, B.; Abrar, T.; Hansard, M.; Oh, C.; et al. A suite of robotic solutions for nuclear waste decommissioning. *Robotics* **2021**, *10*, 112. [CrossRef]
3. Bogue, R. Robots in the nuclear industry: A review of technologies and applications. *Ind. Robot* **2011**, *38*, 113–118. [CrossRef]
4. Zhang, K.; Hutson, C.; Knighton, J.; Herrmann, G.; Scott, T. Radiation tolerance testing methodology of robotic manipulator prior to nuclear waste handling. *Front. Robot. AI* **2020**, *7*, 6. [CrossRef]
5. Lopez Pulgarin, E.J.; Hopper, D.; Montgomerie, J.; Kell, J.; Carrasco, J.; Herrmann, G.; Lanzon, A.; Lennox, B. From traditional robotic deployments towards assisted robotic deployments in nuclear decommissioning. *Front. Robot. AI* **2025**, *12*, 1432845. [CrossRef]
6. Fang, H.-S.; Wang, C.; Fang, H.; Gou, M.; Liu, J.; Yan, H.; Liu, W.; Xie, Y.; Lu, C. AnyGrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Trans. Robot.* **2023**, *39*, 3929–3945. [CrossRef]
7. Mahler, J.; Matl, M.; Satish, V.; Danielczuk, M.; DeRose, B.; McKinley, S.; Goldberg, K. Learning ambidextrous robot grasping policies. *Sci. Robot.* **2019**, *4*, eaau4984. [CrossRef]

8. Newbury, R.; Gu, M.; Chumbley, L.; Mousavian, A.; Eppner, C.; Leitner, J.; Bohg, J.; Morales, A.; Asfour, T.; Kragic, D.; et al. Deep learning approaches to grasp synthesis: A review. *IEEE Trans. Robot.* **2023**, *39*, 3994–4015. [CrossRef]
9. ten Pas, A.; Gualtieri, M.; Saenko, K.; Platt, R. Grasp pose detection in point clouds. *Int. J. Robot. Res.* **2017**, *36*, 1455–1473. [CrossRef]
10. Fang, H.-S.; Wang, C.; Gou, M.; Lu, C. GraspNet-1Billion: A large-scale benchmark for general object grasping. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 14–19 June 2020; pp. 11444–11453. [CrossRef]
11. Sundermeyer, M.; Mousavian, A.; Triebel, R.; Fox, D. Contact-GraspNet: Efficient 6-DoF grasp generation in cluttered scenes. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 13438–13444. [CrossRef]
12. Wang, C.; Fang, H.-S.; Gou, M.; Fang, H.; Gao, J.; Lu, C. Graspness discovery in clutters for fast and accurate grasp detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021; pp. 15944–15953. [CrossRef]
13. Qian, Y.; Zhu, X.; Biza, O.; Jiang, S.; Zhao, L.; Huang, H.; Qi, Y.; Platt, R. ThinkGrasp: A Vision-Language System for Strategic Part Grasping in Clutter. In Proceedings of the 8th Conference on Robot Learning (CoRL), Munich, Germany, 6–9 November 2024; Proceedings of Machine Learning Research. 2025; Volume 270, pp. 3568–3586. Available online: https://proceedings.mlr.press/v270/qian25c.html (accessed on 30 August 2026).
14. Jiao, R.; Fasoli, A.; Giuliari, F.; Bortolon, M.; Povoli, S.; Mei, G.; Wang, Y.; Poiesi, F. Free-form language-based robotic reasoning and grasping. In Proceedings of the 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hangzhou, China, 19–25 October 2025; pp. 17672–17679. [CrossRef]
15. Hou, J.; Xue, X.; Zeng, T. DynTopo: Dynamic Topological Scene Graph for Robotic Autonomy in Human-Centric Environments; ICLR 2026 Conference Submission, OpenReview. 2025. Available online: https://openreview.net/forum?id=OMmgRRh5YL (accessed on 30 August 2026).
16. Murali, A.; Mousavian, A.; Eppner, C.; Paxton, C.; Fox, D. 6-DOF grasping for target-driven object manipulation in clutter. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Virtual, 31 May–31 August 2020; pp. 6232–6238. [CrossRef]
17. 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 2019 IEEE International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 1614–1621. [CrossRef]
18. Zhang, H.; Lu, Y.; Yu, C.; Hsu, D.; Lan, X.; Zheng, N. INVIGORATE: Interactive visual grounding and grasping in clutter. In Proceedings of the Robotics: Science and Systems XVII, Virtual, 12–16 July 2021. [CrossRef]
19. Lou, X.; Yang, Y.; Choi, C. Learning object relations with graph neural networks for target-driven grasping in dense clutter. In Proceedings of the 2022 IEEE International Conference on Robotics and Automation (ICRA), Philadelphia, PA, USA, 23–27 May 2022; pp. 742–748. [CrossRef]
20. Yang, Y.; Liang, H.; Choi, C. A deep learning approach to grasping the invisible. *IEEE Robot. Autom. Lett.* **2020**, *5*, 2232–2239. [CrossRef]
21. Xu, K.; Zhao, S.; Zhou, Z.; Li, Z.; Pi, H.; Zhu, Y.; Wang, Y.; Xiong, R. A joint modeling of vision-language-action for target-oriented grasping in clutter. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023; pp. 11597–11604. [CrossRef]
22. Ultralytics. Ultralytics YOLO26. Available online: https://docs.ultralytics.com/models/yolo26/ (accessed on 21 September 2026).
23. 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. In Proceedings of the Thirteenth International Conference on Learning Representations (ICLR), Singapore, 24–28 April 2025; Available online: https://openreview.net/forum?id=Ha6RTeWMd0 (accessed on 30 August 2026).
24. Liu, Y.; Zhang, J.; Liu, H.; Zhang, Y. DPB-CNN: A Dual-Path Bayesian CNN for Material Recognition with Uncertainty-Aware Adaptive Fusion in Masked Environments. 2026, *under review*.
25. Deen, M.; de Rooij, M. ClusterBootstrap: An R package for the analysis of hierarchical data using generalized linear models with the cluster bootstrap. *Behav. Res. Methods* **2020**, *52*, 572–590. [CrossRef]
26. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.