IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Active-Perceptive Language-Oriented Grasp Policy for Heavily Cluttered Scenes

Yixiang Dai[1*], Siang Chen[1*], Kaiqin Yang[1], Dingchang Hu[1], Pengwei Xie[1], Guosheng Li[1,2], Yuan Shen[1,2], Guijin Wang[1,2,†]

***Abstract*—Language-guided robotic grasping in cluttered environments presents significant challenges due to severe occlusions and complex scene structures, which often hinder accurate target localization. Existing approaches typically suffer from limited observational capabilities, resulting in suboptimal exploration of the target object. In this paper, we propose a novel Active-Perceptive Language-Oriented Grasp Policy (APeG) for heavily cluttered scenes. APeG develops an active perception scheme in the grasp pipeline via an occlusion-aware, semantic-guided viewpoint optimization strategy, enabling efficient exploration of cluttered scenes. In addition, a grasp-wise Reinforcement Learning (RL) policy is proposed to select robust grasp poses. Extensive real-world experiments validate the effectiveness of APeG, demonstrating significant improvements in both task success rate and operational efficiency over existing baselines, highlighting its potential for practical deployment in language-conditioned robotic manipulation.**

***Index Terms*—Active Perception; Language-Oriented Grasping; Reinforcement Learning**

I. INTRODUCTION

GRASPING is a fundamental robotic capability that enables robots to physically interact with and manipulate objects in unstructured and dynamic environments [1], [2]. As a core function in robotic manipulation, it plays a crucial role in a wide range of applications, from industrial automation to household assistance. In recent years, 6-DoF grasping has made significant progress by leveraging deep learning and 3D perception to directly predict grasp poses from point clouds or RGB-D images [3], [4], yielding improved success rates and generalization across object categories.

Within this field, target-oriented grasping, which requires the robot to grasp a specific object according to a high-level goal or language instruction [5], [6], often suffers from severe clutter in realistic scenarios. The obscuring of target objects not only causes immediate perception and grounding failures from a static viewpoint, but also introduces profound ambiguity: the robot must reason about the target's potential hidden location and devise a strategy to access it.

Manuscript received: April, 25, 2025; Revised July, 25, 2025; Accepted August, 13, 2025.

This paper was recommended for publication by Editor Vincze Markus upon evaluation of the Associate Editor and Reviewers' comments.

[1]These authors are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China {daiyx23, csa21, yangkq24, hdc20, xpw18, ligc24}@mails.tsinghua.edu.cn, {shenyuan_ee, wangguijin}@tsinghua.edu.cn

[2]These authors are with Shanghai AI Laboratory, Shanghai 200232, China.

*These authors contributed equally to this work as first authors.

†Corresponding Author: wangguijin@tsinghua.edu.cn.

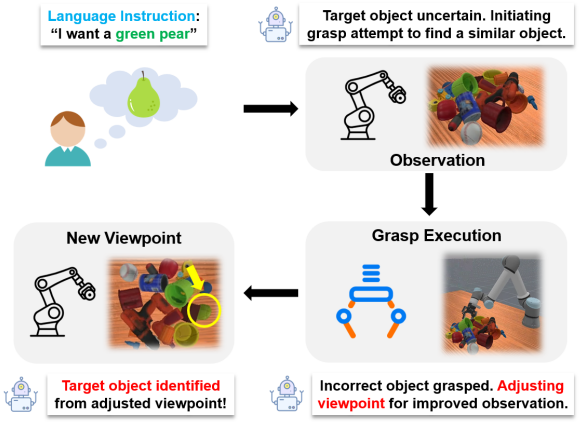


Fig. 1. **Overview**: The proposed language-guided framework leverages active perception to efficiently explore and localize target objects in heavily cluttered environments.

To mitigate the impact of occlusions, recent works have incorporated Next-Best-View (NBV) planning into grasping pipelines [7]–[9]. These methods plan informative viewpoints by modeling grasp uncertainty or occlusion maps, capturing additional observations before executing a grasp. However, most NBV-based approaches adhere to a "scan-then-act" paradigm, where multiple views are collected prior to a single grasp attempt. This model is ill-suited for scenarios requiring physical interaction to resolve occlusions. Retrieving a deeply buried target is fundamentally a sequential decision-making problem, where each grasp should not only attempt to secure an object but also strategically reveal more of the scene. The significant time overhead of pre-scanning and the inability to interleave action and observation make such methods inefficient and often ineffective in these complex, multi-step contexts.

Therefore, we propose a novel closed-loop **A**ctive-**P**erceptive Language-oriented **G**rasping method (**APeG**) in heavily cluttered scenes. APeG leverages active viewpoint optimization into language-oriented grasping tasks for better exploration given a single RGB-D observation for each grasp attempt, as depicted in Fig. 1, and combines Reinforcement Learning (RL) to select robust grasps for the target object, enabling robust execution in cluttered and partially observable environments. In summary, our contributions are as follows:

- We propose APeG, a novel closed-loop active-perceptive framework for language-oriented grasping in heavily cluttered scenes.
- We design an active-perceptive viewpoint optimization

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

scheme based on joint occlusion-semantics gain for target object exploration and localization.
- We present an efficient grasp-wise reinforcement learning policy for grasp selection, facilitating better grasp execution in cluttered scenes.
- We develop a simulated language-oriented grasping benchmark for heavily cluttered scenes based on Maniskill3 [10] and apply it to real-world evaluations.

II. RELATED WORKS

A. Target-oriented 6-Dof Grasp Detection

A critical subtask in robotic manipulation is target-oriented grasping, where a robot retrieves a specific object based on high-level goals such as language instructions. Recent works have explored diverse strategies to address this challenge. For instance, Qian et al. [11] utilized GPT-4o [12] to interpret language prompts and guide object parsing and grasp generation. Xie et al. [13] developed a flexible localization module to support various goal representations and generated grasps over spatially constrained regions. However, a key limitation of these methods is their predominant reliance on a fixed, static viewpoint, which is often insufficient in heavily cluttered scenes where severe occlusions can prevent a successful grasp for the target object.

Reinforcement learning (RL) has also been extensively applied to robotic grasping, often within closed-loop frameworks that can handle cluttered scenes [14]–[17]. Zeng et al. [14] used a Q-learning approach where an agent learns to select between pushing and grasping actions directly from visual input to clear clutter. More recent efforts have focused on effective supervision and reward engineering. For example, Wang et al. [16] utilized expert demonstrations to guide policy learning for target-specific grasps, while Herland et al. [15] and Joshi et al. [17] explored various reward structures to shape agent behavior. Directly addressing the target-oriented challenge, Xu et al. [18] proposed a reinforcement learning framework that jointly encoded vision, language, and grasp features for candidate selection. However, these methods often assume that the target object, even if partially occluded, can be reliably localized from the initial viewpoint, which often breaks down in severe occlusion scenes.

In contrast, our framework incorporates an active perception strategy to resolve the perceptual ambiguity inherent in cluttered scenes. This enables the robot to handle severe occlusions more effectively by dynamically adjusting its viewpoint during the grasping process, leading to more efficient exploration and robust target localization.

B. Active Perception for Grasp Detection

The active perception problem focuses on determining the best perceptual position or viewpoint to improve information gain in vision tasks. Several works have incorporated Next-Best-View (NBV) planning into grasping pipelines to reduce occlusions and improve grasp detection. Breyer et al. [7] estimated occlusion volumes using ray casting to compute viewpoint utility. Zhang et al. [8] predicted grasp affordances across multiple views to guide view selection. Ma et al. [9] rendered depth images from candidate viewpoints and evaluated pseudo-graspness to determine informative perspectives. While effective in improving geometric completeness, these NBV-based approaches typically follow a preplan-then-act paradigm, in which the robot plans a sequence of viewpoints before executing a single grasp. Such strategies can be time-consuming and are often suboptimal in settings that require sequential interactions.

In contrast, our work proposes a closed-loop active perception framework tailored for target-oriented grasping under language guidance. Rather than accumulating multiple views before action, our system dynamically interleaves viewpoint planning and grasp execution in a cycle, enabling the robot to progressively uncover occluded target objects. Furthermore, we integrate task-specific semantic relevance into our NBV utility function, allowing the robot to prioritize views that are both geometrically informative and semantically aligned with the instruction.

III. PROBLEM STATEMENT

This study addresses the problem of object grasping in a cluttered tabletop environment, where the objective of a robotic arm is to grasp a specified target object based on human language instructions. Similar to VLG [18], the objects present on the tabletop are typically arranged in heavily stacked configurations, resulting in significant occlusion and perceptual ambiguity, thereby complicating the precise identification and localization of the target object. The task is considered successful if the robot arm successfully grasps the target object within a limited number of grasp attempts, denoted by T. Different from VLG, the robot is allowed to dynamically adjust its observation viewpoint after each execution step.

At each discrete execution step i, the robotic system acquires an observation O_i, composed of a single-view egocentric RGB-D image $\boldsymbol{I}_i \in \mathbb{R}^{H\times W\times 4}$, a language instruction L, and the robotic arm's pose from the preceding step, denoted by P_{i-1}. Given these inputs, the robot is tasked with learning two complementary policies: a grasping policy $\pi_{grasp} : \mathcal{O} \rightarrow \mathcal{A}_{grasp}$ and a viewing policy $\pi_{view} : \mathcal{O} \rightarrow \mathcal{A}_{view}$. Specifically, π_{grasp} determines the optimal 6-DoF grasp action, identifying both the location and approach for grasping objects at the current step, while π_{view} specifies the subsequent positioning of the robotic arm after the grasp implement to enhance visual perception and facilitate improved object identification in future steps.

IV. METHOD

A. System Pipeline

Our framework develops an active perception scheme to enhance semantic understanding and occlusion exploration, while employing a grasp-wise reinforcement learning (RL) policy to select high-quality grasp actions. It adopts an iterative closed-loop strategy, where the robot executes a grasp attempt followed by an adaptive adjustment of its observation viewpoint in each cycle until language-grounded object retrieval is successfully achieved.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

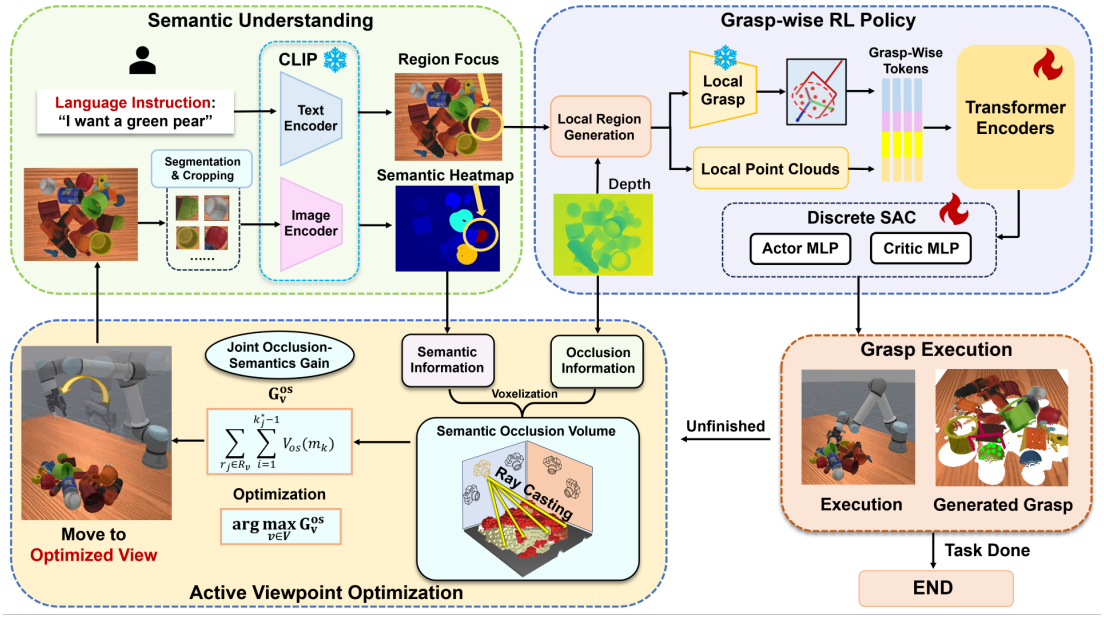


Fig. 2. **Framework**: Given a language instruction, CLIP-based semantic understanding identifies the target object via multimodal feature alignment. Local grasp candidates are then generated from the focused region and refined by an RL-based policy using grasp-wise tokens. Simultaneously, an active viewpoint optimization module estimates occlusion-aware semantic information gain via ray casting, guiding the robot to more informative views. The grasp-observe cycle continues until the target object is successfully retrieved.

At each execution cycle t_i, the robot initiates a semantic understanding module to interpret the language instruction and identify the target region leveraging the CLIP model [19], and generates occlusion and semantic voxel grids via ray casting and heatmap projection. Then, a grasp-wise reinforcement learning (RL) policy is trained to identify robust grasps under occlusion from grasp candidates generated by the Local Grasp (LoG) module [4]. Finally, an active-perceptive viewpoint optimization is planned based on joint occlusion-semantics information gain. This iterative process enables efficient and reliable language-guided grasping in cluttered scenes.

B. Geometric and Semantic Understanding

This module serves as the perception backbone of the framework, responsible for constructing a unified geometric and semantic representation of the scene. It localizes the language-referred target region and reconstructs occlusion and semantic voxel grids, which provide critical inputs for next-best-view planning in cluttered environments.

Given an observed RGB-D image $\boldsymbol{I}_i$, we first segments the scene to obtain a set of N object masks $\{\boldsymbol{M}_j\}_{j=1}^{N}$ and their corresponding image crops $\{\boldsymbol{C}_j\}_{j=1}^{N}$. Similarly to VLG [18], we leverage the CLIP model [19], which includes a visual encoder E_V and a textual encoder E_T. We compute the textual feature embedding $\boldsymbol{f}_L = E_T(L)$ and the visual feature for each object crop $\boldsymbol{f}_{V,j} = E_V(\boldsymbol{C}_j)$. The relevance score s_j for each object j is then calculated as the cosine similarity between the normalized visual and textual features: $s_j = \frac{\boldsymbol{f}_{V,j} \cdot \boldsymbol{f}_L}{\|\boldsymbol{f}_{V,j}\| \|\boldsymbol{f}_L\|}$. The target region is identified by ranking these scores, and the highest-ranked region is used to generate a localized input for the grasp model. Simultaneously, we construct a language-guided heatmap $\mathcal{H}_t$ by assigning the score s_j to all pixels within the corresponding mask $\boldsymbol{M}_j$.

Utilizing the depth channel and intrinsic camera parameters, we convert the observed scene into a 3D point cloud and subsequently voxelize it into a grid structure. However, due to the heavy clutter in the scene, many objects may be partially or fully occluded. To capture this occlusion information, we adopt an occlusion volume construction method inspired by [7]. We observe that computing the truncated signed distance function (TSDF) from a single-view RGB-D image yields low accuracy and is thus inadequate for supporting effective viewpoint optimization. To address this limitation, we define an occlusion volume $V_{\text{occ}} : \mathcal{M} \rightarrow \{0, 1\}$ over a voxel grid $\mathcal{M}$, where each occlusion voxel is assigned a value based on its visibility with respect to rays cast from the camera optical center o_c.

Let $\mathcal{P} = \{p_i\}_{i=1}^{N}$ be the observed 3D point cloud of objects from RGB-D images. Each point p_i is voxelized into a grid cell, forming the surface voxel set $\mathcal{S}$:

$$\mathcal{S} = \{m \in \mathcal{M} \mid m = \text{voxel}(p_i),\ \exists p_i \in \mathcal{P}\}. \tag{1}$$

Next, we define a set of dense rays $\mathcal{R} = \{r_j\}_{j=1}^{|\mathcal{R}|}$ emitted from o_c, one per pixel. Each ray r_j traverses a sequence of voxels $m_1^j, m_2^j, \ldots, m_K^j \subset \mathcal{M}_j$. The index of the first surface voxel encountered along ray r_j is denoted:

$$k_j^* = \min\{k \mid m_k^j \in \mathcal{S}\}. \tag{2}$$

The voxels behind the surface (i.e., $m_{k^*+1}^j, \ldots, m_K^j$) are labeled as occluded:

$$V_{\text{occ}}(m) = 1, \quad \forall m \in \{m_{k_j^*+1}^j, \ldots, m_K^j\}. \tag{3}$$

All remaining voxels are assigned zero:

$$V_{\text{occ}}(m) = 0, \quad \text{otherwise}. \tag{4}$$

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Fig. 3. Examples of simulated cluttered scenes with diverse language instructions in simulated environments.

This process results in an occlusion-aware voxel grid that explicitly distinguishes between observed surfaces, occluded regions, and free space.

In parallel, we project the language-guided heatmap into 3D space using the depth image and camera intrinsics to obtain a point-wise semantic relevance map $\{(p_j, h_j)\}$, where p_j is a 3D point and h_j is its associated relevance score, and voxelize the semantic point cloud to generate the surface semantic voxel grid $\hat{V}_{\text{sem}}$. Then the semantic voxel gird V_{sem} is obtained using ray casting:

$$V_{\text{sem}}(m) = \hat{V}_{\text{sem}}(m_{k_j^*}), \quad \forall m \in \mathcal{M}_j, \tag{5}$$

where M_j denotes the set of voxels that ray r_j traverses.

As a result, the current observation is represented by a pair of volumetric fields $(V_{\text{occ}}, V_{\text{sem}})$, capturing both geometric visibility and semantic alignment. These grids are subsequently fused and utilized for active viewpoint optimization.

C. Active Viewpoint Optimization

This module is proposed to actively select the next observation viewpoint to improve scene understanding and task efficiency. Unlike conventional strategies that rely on fixed poses, we propose a dynamic viewpoint optimization approach that adaptively maximizes joint occlusion-semantics gain, enabling the robot to uncover occluded target objects and reduce unnecessary actions in cluttered environments. Specifically, after each grasp execution, the robot arm must move to an optimized observation pose for better exploration.

We begin by generating a set of candidate viewpoints $\mathbf{V}$ distributed over a hemisphere centered at the centroid of the tabletop. The radius of the hemisphere is chosen to ensure the reachability of the robot arm. To achieve uniform coverage, we employ the Fibonacci sphere sampling method [20], which produces evenly distributed directions on the hemisphere. The latitude and longitude of each sampled point are computed as:

$$\begin{aligned} \text{lat}_i &= \arccos\left(\frac{2i}{2m+1}\right), \\ \text{long}_i &= 2\pi i \Phi^{-1}, \quad i \in [-m, m], \end{aligned} \tag{6}$$

where m controls the sampling density and Φ is the golden ratio.

To evaluate each candidate viewpoint, we estimate the expected information gain by combining the occlusion and semantic voxel grids. Specifically, we construct an occlusion-aware semantic volume V_{os} by fusing the occlusion volume V_{occ} with the semantic volume V_{sem} as follows:

$$V_{\text{os}}(m) = V_{\text{occ}}(m) * V_{\text{sem}}(m). \tag{7}$$

To quantify the potential visibility improvement from each candidate viewpoint $v \in \mathbf{V}$, we simulate a ray-casting process. From each viewpoint, a set of rays R_v is emitted toward the scene. Each ray $r \in R_v$ sequentially traverses a set of voxels $M_r = \{m_1, m_2, \ldots, m_{k^*}\}$, where k^* is the index of the first voxel on the ray that intersects the observed surface (i.e., $m_{k^*} \in \mathcal{S}$). We compute the **joint occlusion-semantics gain** G_v^{os} for each viewpoint v by summing over the traversed voxels before the surface is reached:

$$G_v^{\text{os}} = \sum_{r_j \in R_v} \sum_{i=1}^{k_j^*-1} V_{\text{os}}(m_k), \tag{8}$$

where k_j^* denotes the index at which ray r_j first hits a surface voxel.

Finally, the optimal next-best-view v^* is selected as the one maximizing the estimated information gain:

$$v^* = \arg\max_{v \in \mathbf{V}} \; G_v^{os}. \tag{9}$$

By employing a greedy strategy, our viewpoint optimization module makes a locally optimal decision at each step. Based on the current observation, the robot selects the viewpoint predicted to most effectively reduce occlusions and enhance semantic understanding, thereby dynamically adapting its observation pose for subsequent grasping actions.

D. Grasp-wise RL Policy

In this module, we generate candidate 6-DoF grasp poses from local point clouds centered on the target area using a Local Grasp (LoG) model and train a grasp-wise reinforcement learning (RL) policy to select the most suitable grasp for execution.

We begin by sampling local point clouds around the target region using the Farthest Point Sampling (FPS) algorithm. Given the sampled point set, the LoG model predicts a set of candidate 6-DoF grasp poses $\hat{\mathbf{G}} = \{\hat{\mathbf{g}}_j, j = 1, ..., N\}$. A 6-Dof grasp configuration $\boldsymbol{g} \in \mathbf{G}$ can be defined as:

$$\boldsymbol{g} = (x, y, z, \theta, \gamma, \beta, w), \tag{10}$$

where (x, y, z) represents the 3-DoF position of the grasp and $(\theta, \gamma, \beta) \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ denotes the 3-DoF rotation of the grasp pose in the form of Euler angles within the gripper coordinate system, and w corresponds to the gripper width.

To enhance sim-to-real transferability, we encode each grasp as a Gaussian-distributed point cloud $\hat{\mathbf{P}}_{\mathbf{G}}$, parameterized by its translation and rotation matrices, following GAP-RL [21]. To enable multi-modal feature fusion, we apply PointNet [22] to encode the grasp representations $\hat{\mathbf{P}}_{\mathbf{G}}$ and adopt DGCNN [23] to extract both global and local geometric features from the local point clouds. Each grasp feature is concatenated with the point cloud feature to form grasp-wise tokens, which are subsequently passed through transformer encoders for cross-modal fusion. The fused representations are then used as input to the RL network.

We adopt the Discrete Soft Actor-Critic (DSAC) algorithm to train the RL policy. The observations include K fused grasp-wise features from the grasp representations and local point

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

clouds. The policy and critic MLPs both take K grasp-wise attention features as input, and output the feasible logits and q values of these K grasp poses.

The reward function is designed to guide the agent toward selecting high-quality grasps from the LoG-generated candidates:

$$r(s,a) = \begin{cases} +10, & \text{if } a \text{ leads to successful grasp,} \\ -1, & \text{if } a \text{ fails to grasp the object,} \\ -2, & \text{if } a \text{ results in infeasible motions,} \end{cases} \tag{11}$$

where the final case penalizes poorly positioned grasp configurations that prevent successful motion planning. To improve sample efficiency during early training, we incorporate an ϵ-greedy exploration strategy guided by the grasp confidence scores produced by the LoG model to warm up.

V. EXPERIMENT

In this section, we detail the implementation of the active perception module and the training of the reinforcement learning (RL) policy in simulation. We then present a series of evaluations conducted in both simulated environments and on a real-world robotic platform.

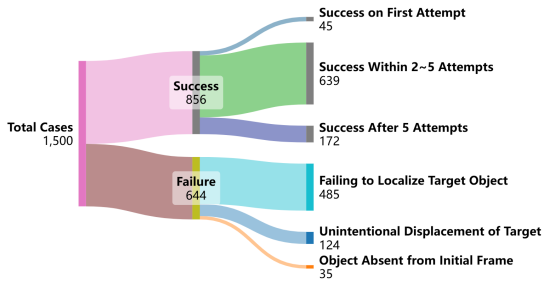
A. Simulation Setup

Following VLG [18], we design language instructions based on five templates: "Give me the {keyword}", "Grasp a {keyword}", "I want a {keyword} object", "I need a {keyword}", and "Get something to {keyword}". In line with baselines, which set a fixed limitation for grasp execution steps, we limit the maximum number of grasp attempts to half the total number of objects to ensure a consistent evaluation across scenes with different object counts.

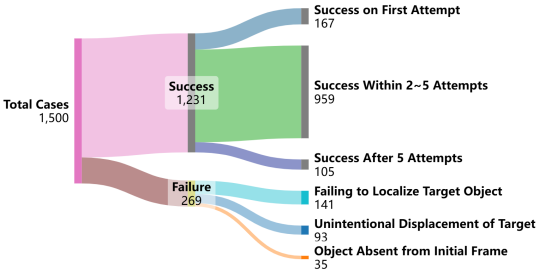
Our simulation environment is built upon the ManiSkill3 platform [10], equipped with a UR5e robotic arm, a Robotiq 2F-85 parallel gripper, and an Intel RealSense D435 depth camera. Cluttered scenes are generated using 57 object models and 28 associated language keywords, covering two levels of complexity: *moderate clutter* and *severe clutter*. Each moderately cluttered scene contains 20 objects randomly placed within a 40 cm $\times$ 40 cm workspace, while severely cluttered scenes include 30 objects within the same area, with increased potential for occlusions and stacking around the target object. Following the settings of VLG [18] and ThinkGrasp [11], ground truth segmentation masks are adopted from the simulator to make fair comparisons.

Evaluation Metrics. To comprehensively evaluate the effectiveness of our proposed method, we adopt the following three metrics:

- **Task Success Rate (TSR)**: The percentage of test scenes in which the target object is successfully grasped within the maximum number of grasp attempts;
- **Average Grasp Motions (AGM)**: The average number of grasp attempts performed by the robot until the target object is successfully grasped or the maximum step limit is reached.



(a) VLG



(b) Ours

Fig. 4. **Case Analysis**: Comparison of task outcome distributions between VLG and our proposed method over 1,500 language-guided grasping trials.

- **Average Completion Time (ACT)**: Average runtime of the whole language-guided grasping task, including active perception, grasp generation, and motion execution.

To evaluate the effectiveness of APeG, we compare its performance with the following methods:

- **VLG** [18]: A vision-language grasping framework that jointly encodes visual, linguistic and grasp features to select grasp candidates using a reinforcement learning policy.
- **OVGrasp** [24]: An open-vocabulary grasping framework that integrates visual-linguistic representations and enhances alignment through dual-modality perception modules.
- **ThinkGrasp** [11]: A language-conditioned grasping framework that utilizes GPT-4o for object parsing and grounding. Grasp candidates are generated based on sub-regional proposals.
- **Breyer's** [7]: A next-best-view planning method that evaluates occlusion information using ray casting to select optimized observation views.

B. Simulation Experiments

To comprehensively evaluate the performance of our proposed method under varying levels of clutter, we conduct experiments under two settings: **moderate clutter** (20 objects) and **severe clutter** (30 objects). The quantitative results are presented in Table I. Among the compared baselines, VLG, OVGrasp, and ThinkGrasp are all single-view-observation methods with fixed observation poses. Our approach outperforms these baselines by over 13.8% in task success rate and reduces the average grasp motions by more than 8.0% across both clutter settings.

To ensure a fair comparison with Breyer's next-best-view method [7], we re-implement it using our language grounding

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE I
SIMULATION RESULTS OF LANGUAGE-GUIDED GRASPING UNDER MODERATE AND SEVERE CLUTTERS. WE REPORT TASK SUCCESS RATE (TSR), AVERAGE GRASP MOTIONS (AGM) AND AVERAGE COMPLETION TIME (ACT) ACROSS ALL METHODS. RESULTS ARE AVERAGED OVER 5 ROUNDS.

	Moderate Clutter (20 Objects)			**Severe Clutter (30 Objects)**		
	TSR ↑	AGM ↓	ACT (s) ↓	TSR ↑	AGM ↓	ACT (s) ↓
VLG [18]†	$57.1\% \pm 2.8\%$	6.07 ± 0.27	59.8	$43.8\% \pm 2.9\%$	7.19 ± 0.28	71.1
OVGrasp [24]†	$44.0\% \pm 9.5\%$	5.16 ± 0.52	52.2	$39.7\% \pm 7.3\%$	6.02 ± 0.48	61.3
ThinkGrasp [11]†	$71.6\% \pm 3.0\%$	3.84 ± 0.15	68.9	$61.8\% \pm 2.2\%$	5.53 ± 0.19	99.7
Breyer's [7]‡	$76.7\% \pm 3.9\%$	3.76 ± 0.17	39.5	$66.4\% \pm 3.7\%$	5.42 ± 0.31	58.4
APeG	**82.1%** $\pm 1.3\%$	**3.37** ± 0.15	**34.8**	**70.3**% $\pm 0.9\%$	**5.09** ± 0.19	**52.9**

†: For a fair comparison, these methods are reproduced with the same grasp module [25].
‡: Breyer's Next-Best-View Planning is re-implemented with our visual grounding and local grasp module.

module. Our approach achieves superior performance, with improvements of over 5.9% in task success rate and 6.1% in average grasp motions. These results validate the effectiveness of our active-perception-enhanced framework in both moderate and heavily cluttered environments.

In terms of time efficiency, Table I shows that APeG achieves the lowest Average Completion Time (ACT) under both clutter settings. With an ACT of 34.8s in moderate clutter and 52.9s in severe clutter, APeG significantly outperforms all baseline methods. It is worth noting that while our active perception strategy introduces a minor computational overhead per grasp cycle, this cost is negligible. Specifically, within each step, the active perception module requires only 0.071s and grasp generation takes 0.029s, while the average physical execution time is 10.16s. This shows that the small computational investment for viewpoint optimization yields a substantial reduction in overall task time by significantly decreasing the total number of required grasp attempts.

Fig. 5 illustrates the task success rate relative to the initial occlusion rate of the targets, including a multi-view baseline that scans the scene from 6 fixed viewpoints and only attempts a single grasp. The results validate that while all methods degrade as occlusion increases, APeG consistently outperforms the baselines and maintains greater stability, showing a significant advantage under severe clutter. This resilience stems from our joint occlusion-semantics gain strategy, which excels even when the target is completely invisible. In such cases, the system intelligently selects new viewpoints that are most likely to reveal the target by uncovering occluded regions with high semantic relevance to the instruction.

We further conduct a case analysis as illustrated in Fig. 4. Compared to VLG, our method exhibits a superior ability to localize the target object. In VLG, the observation viewpoint remains fixed, making it difficult to detect target objects that are heavily occluded in the current view. As a result, more grasping steps are often required to remove surrounding objects before reaching the target. In contrast, our method dynamically adjusts the observation views based on occlusion and semantic cues, enabling earlier discovery of the target object and reducing the number of required grasp attempts.

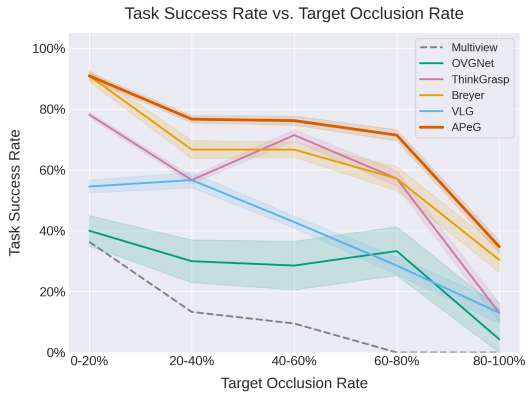


Fig. 5. Task Success Rate vs. Target Occlusion Rate for Different Methods. The results indicate that our proposed APeG consistently achieves superior or competitive success rates across nearly all occlusion levels.

C. Ablation Studies

To validate the contribution of each core component in our framework, we conduct a comprehensive ablation study with results presented in Table II. Occlusion-guided viewpoint optimization provides a notable performance improvement, which confirms the benefit of actively resolving geometric occlusions. Furthermore, integrating semantic-enhanced viewpoint optimization yields an incremental gain in the success rate. This demonstrates the value of adding task-oriented knowledge. Moreover, the reinforcement learning policy for grasp selection further improves performance by effectively filtering out low-quality grasp candidates. These findings demonstrate the critical contributions of both the active viewpoint optimization and the grasp selection policy to the overall system performance.

TABLE II
ABLATION STUDIES: THE RESULTS ARE AVERAGED OVER 5 ROUNDS.

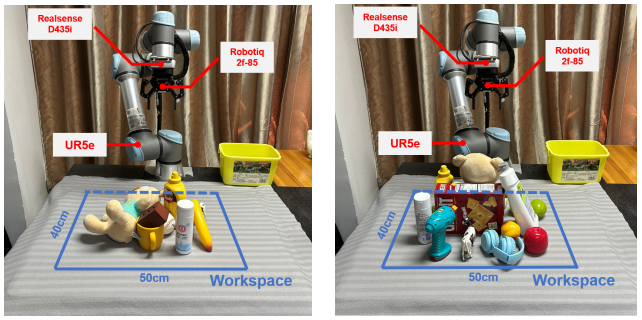
OVO	**SVO**	**G-RL**	TSR ↑	AGM ↓
			$75.2\% \pm 1.9\%$	3.81 ± 0.15
✓			$78.1\% \pm 2.6\%$	3.64 ± 0.16
✓	✓		$79.5\% \pm 2.1\%$	3.56 ± 0.14
✓	✓	✓	**82.1%** $\pm 1.3\%$	**3.37** ± 0.15

OVO: Occlusion-Guided Viewpoint Optimization
SVO: Semantic-Enhanced Viewpoint Optimization
G-RL: Grasp-wise Reinforcement Learning

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

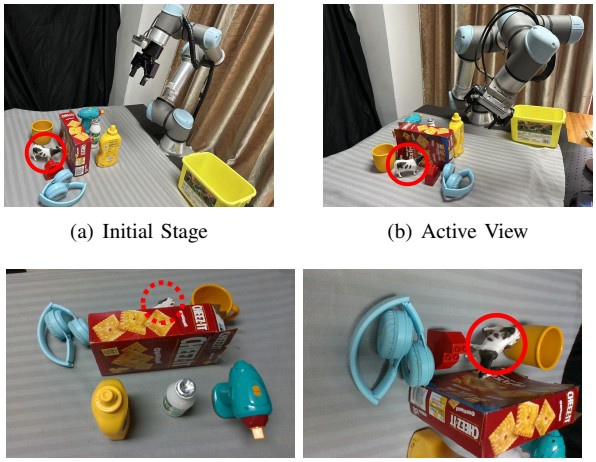
TABLE III
REAL-WORLD EXPERIMENT RESULTS ACROSS TEN CLUTTERED SCENES. **GSR**: AVERAGE GRASP SUCCESS RATE.

Scene	Method	TSR ↑	GSR ↑	AGM ↓
8 Objects	CLIP Grounding	60% (3/5)	65.3%	5.2
	VLG [18]	60% (3/5)	70.0%	6.2
	APeG	**100% (5/5)**	**84.6%**	**2.6**
12-15 Objects	CLIP Grounding	60% (3/5)	56.3%	8.6
	VLG [18]	40% (2/5)	44.7%	11.2
	APeG	**80% (4/5)**	**66.7%**	**5.6**



(a) Moderate Clutter (Realworld) (b) Severe Clutter (Realworld)

Fig. 6. **Realworld Settings**: (a) moderate clutter in realworld with 8 objects; (b) severe clutter in realworld with 12-15 objects.



(a) Initial Stage (b) Active View

(c) Initial Stage (Observation) (d) Active View (Observation)

Fig. 7. **Real-world Robot Experiment**: Language instruction: **"Give me the toy cow"**. (a, c) The initial fixed view fails to reveal the occluded target object. (b, d) After viewpoint adjustment, the target (cat-printed bowl) becomes visible, enabling successful localization and grasping.

D. Real Robot Experiments

To validate the sim-to-real transferability of our framework, we conduct real-world experiments using a UR5e robotic arm equipped with a Robotiq 2F-85 parallel gripper and an Intel RealSense D435 RGB-D camera, as illustrated in Fig. 6. The test environments consist of cluttered tabletop scenes with various occlusions, posing substantial challenges for traditional language-guided grasping methods. We construct two types of cluttered scenes in the real world (different from simulation): with 8 objects and with 12–15 objects in a 40 cm × 50 cm workspace respectively. Each type includes 5 distinct scenes. To evaluate the generalization capabilities of our framework, our real-world scenes featured randomly placed objects, with approximately one-third novel items (not in the simulation training set). For evaluation, the maximum number of allowed grasp motions in the real world is set equal to the number of objects present in the scene. To obtain segmentation masks, we employ the Segment Anything Model (SAM) [26], which achieves a mean Intersection over Union (mIoU) of 87.1% for visible objects in our evaluation.

Table III summarizes the quantitative results across ten representative scenes. The CLIP-Grounding baseline is implemented by combining object segmentation via SAM [26] with CLIP-based language-to-object matching, followed by grasp execution using the Local Grasp Module. Compared to the CLIP-grounding and VLG baselines, our method consistently achieves higher success rates while requiring fewer grasp attempts. Besides, our framework demonstrates effective sim-to-real transferability, which is evidenced by the minimal performance degradation observed between moderate clutter in simulation and equivalent real-world scenes. This highlights the robustness and efficiency of our framework under real-world conditions.

Further qualitative analysis, illustrated in Fig. 7, demonstrates the advantage of our active perception module in resolving severe occlusions. In the initial view (Fig. 7 a, c), the target object—a toy cow—is completely or partially hidden behind other items, making it difficult for the vision-language grounding module to localize the correct region. Traditional language-guided methods, which rely on a fixed observation viewpoint, often fail in such cases or require multiple exploratory grasps. In contrast, our method adaptively adjusts the observation viewpoint (Fig. 7 b, d), revealing the occluded target and enabling successful grasping with fewer attempts. This demonstrates the effectiveness of our dynamic viewpoint optimization in complementing semantic grounding and improving real-world performance.

E. Discussion

Failure Case. We have also observed several failure cases during grasp execution. The most common failure mode was unintended collisions of the robot arm with non-target objects during motion, an issue particularly prominent in severely cluttered scenes. This issue is also due to our current motion planning strategy: the system only examines the gripper's target pose but does not perform full-arm collision checking against all scene objects throughout the entire trajectory.

Sim-to-Real. Our grasp selection policy achieves robust sim-to-real transfer through grasp-as-points representation. By

representing grasps as Gaussian-distributed point clouds and training our reinforcement learning agent on local point cloud data, the policy learns to prioritize object shape and structure. This geometric foundation makes the policy largely invariant to visual domain shifts, such as textures and lighting, that differ between simulation and reality. Besides, our framework demonstrates effective performance handling novel objects as our grasp policy is category-agnostic, which is trained on local geometric features, and the pretrained CLIP is leveraged with strong zero-shot capabilities.

Real-to-Sim. CLIP, which is pre-trained on real-world images, can result in a degradation for visual-text alignment in the simulation environment. Yet, our active perception allows the robot to overcome initial low-confidence or ambiguous groundings by autonomously moving to a more informative viewpoint, demonstrating the effective compensation for initial perceptual uncertainties.

Robustness to Segmentation. Our framework is inherently robust to possible segmentation degradation. First, it does not depend on a single pre-identified target mask, but rather generates candidate masks from all visible objects for visual-text alignment. Second, our active exploration strategy provides a powerful recovery mechanism. If an initial perception is flawed, the system can acquire a potentially clearer viewpoint, allowing it to re-evaluate the scene.

VI. CONCLUSION

In this work, we propose an iterative closed-loop framework for target-oriented grasping in heavily cluttered scenes, integrating a grasp refinement policy with proactive viewpoint optimization. We introduce a proactive viewpoint optimization strategy that jointly considers occlusion and semantic cues, enabling more informative observations and significantly improving task success rates. In addition, a grasp-wise reinforcement learning policy is trained to refine grasp selection in local regions, enhancing grasp quality within target-centric areas. Comprehensive experiments in both simulated and real-world settings validate the effectiveness and robustness of our approach, which outperforms existing baselines in terms of grasp efficiency and accuracy under varying levels of clutter.

In the future, we plan to extend the RL framework to jointly optimize both grasp execution and viewpoint planning, enabling a unified decision-making policy across the grasp-observe cycle.

REFERENCES

[1] S. Chen, W. Tang, P. Xie, W. Yang, and G. Wang, "Efficient heatmap-guided 6-dof grasp detection in cluttered scenes," *IEEE Robot. Autom. Lett.*, vol. 8, no. 8, pp. 4895–4902, 2023.

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

[3] C. Wang, H.-S. Fang, M. Gou, H. Fang, J. Gao, and C. Lu, "Graspness discovery in clutters for fast and accurate grasp detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 964–15 973.

[4] S. Chen, P. Xie, W. Tang, D. Hu, Y. Dai, and G. Wang, "Region-aware grasp framework with normalized grasp space for efficient 6-dof grasping," *arXiv preprint arXiv:2406.01767*, 2024.

[5] A. Murali, A. Mousavian, C. Eppner, C. Paxton, and D. Fox, "6-dof grasping for target-driven object manipulation in clutter," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6232–6238.

[6] Z. Liu, Z. Wang, S. Huang, J. Zhou, and J. Lu, "Ge-grasp: Efficient target-oriented grasping in dense clutter," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 1388–1395.

[7] M. Breyer, L. Ott, R. Siegwart, and J. J. Chung, "Closed-loop next-best-view planning for target-driven grasping," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 1411–1416.

[8] X. Zhang, D. Wang, S. Han, W. Li, B. Zhao, Z. Wang, X. Duan, C. Fang, X. Li, and J. He, "Affordance-driven next-best-view planning for robotic grasping," in *Conference on Robot Learning*. PMLR, 2023, pp. 2849–2862.

[9] H. Ma, M. Shi, B. Gao, and D. Huang, "Active perception for grasp detection via neural graspness field," *Advances in Neural Information Processing Systems*, vol. 37, pp. 38 122–38 141, 2024.

[10] S. Tao, F. Xiang, A. Shukla, Y. Qin, X. Hinrichsen, X. Yuan, C. Bao, X. Lin, Y. Liu, T.-k. Chan *et al.*, "Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai," *arXiv preprint arXiv:2410.00425*, 2024.

[11] Y. Qian, X. Zhu, O. Biza, S. Jiang, L. Zhao, H. Huang, Y. Qi, and R. Platt, "Thinkgrasp: A vision-language system for strategic part grasping in clutter," *arXiv preprint arXiv:2407.11298*, 2024.

[12] OpenAI, "Gpt-4o," 2024, accessed: 2025-4-15. [Online]. Available: https://openai.com

[13] P. Xie, S. Chen, D. Hu, Y. Dai, K. Yang, and G. Wang, "Target-oriented object grasping via multimodal human guidance," *arXiv preprint arXiv:2408.11138*, 2024.

[14] A. Zeng, S. Song, S. Welker, J. Lee, A. Rodriguez, and T. Funkhouser, "Learning synergies between pushing and grasping with self-supervised deep reinforcement learning," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 4238–4245.

[15] S. Herland, K. Bach, and E. Misimi, "6-dof closed-loop grasping with reinforcement learning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 7812–7818.

[16] L. Wang, Y. Xiang, W. Yang, A. Mousavian, and D. Fox, "Goal-auxiliary actor-critic for 6d robotic grasping with point clouds," in *Conference on Robot Learning*. PMLR, 2022, pp. 70–80.

[17] S. Joshi, S. Kumra, and F. Sahin, "Robotic grasping using deep reinforcement learning," in *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2020, pp. 1461–1466.

[18] K. Xu, S. Zhao, Z. Zhou, Z. Li, H. Pi, Y. Zhu, Y. Wang, and R. Xiong, "A joint modeling of vision-language-action for target-oriented grasping in clutter," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 597–11 604.

[19] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[20] J. Ren, F. Wang, J. Zhang, Q. Zheng, M. Ren, and B. Shi, "Diligent102: A photometric stereo benchmark dataset with controlled shape and material variation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 581–12 590.

[21] P. Xie, S. Chen, Q. Chen, W. Tang, D. Hu, Y. Dai, R. Chen, and G. Wang, "Gap-rl: Grasps as points for rl towards dynamic object grasping," *IEEE Robotics and Automation Letters*, 2024.

[22] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 652–660.

[23] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *ACM Transactions on Graphics (tog)*, vol. 38, no. 5, pp. 1–12, 2019.

[24] M. Li, Q. Zhao, S. Lyu, C. Wang, Y. Ma, G. Cheng, and C. Yang, "Ovgnet: A unified visual-linguistic framework for open-vocabulary robotic grasping," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 7507–7513.

[25] W. Tang, S. Chen, P. Xie, D. Hu, W. Yang, and G. Wang, "Rethinking 6-dof grasp detection: A flexible framework for high-quality grasping," *arXiv preprint arXiv:2403.15054*, 2024.

[26] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.