Robotica (2026), **44**, pp. 1830–1842
doi:10.1017/S0263574726103683

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Knowledge-augmented embodied exploration for robotic grasping in constrained environments

Jin Liu[1], Kai Sun[1], Leibing Xiao[2] and Chaoqun Wang[3]

[1]The School of Computer Science and Technology/School of Artificial Intelligence, China University of Mining and Technology, China
[2]Qilu Software College, Shandong University, China
[3]The School of Control Science and Engineering, Shandong University, China
Corresponding author: Jin Liu; Email: sakerliujin@163.com

Received: 21 September 2025 UTC; **Revised:** 5 March 2026 UTC; **Accepted:** 26 May 2026 UTC;
First published online: 20 July 2026 UTC

Keywords: active exploration; knowledge-guided; robotic grasping detection

Abstract

It has always been expected that robots can actively manipulate complex environments to fulfill human requirements. This process typically necessitates that the robot be equipped with the ability for embodied exploration and manipulation. To achieve this goal, in this paper, we propose to incorporate multi-source knowledge to enhance the ability of robotic embodied exploration and manipulation. Specifically, to eliminate the inherent biases in decision-making of large language models (LLMs), we introduce a multi-source knowledge fusion module to generate more reasonable exploration sequences. Notably, grasping detection plays a critical role in the process of robot manipulation. To achieve a better balance between the accuracy and efficiency of the grasping detection network, we design a two-branch feature fusion module with residual blocks to improve network performance. Conditioned on the aforementioned innovations, the robot is capable of actively exploring and manipulating in constrained environments to meet human requirements. Extensive experiments are conducted in both simulation and real-world environments. The results demonstrate the effectiveness and efficiency of our proposed framework.

1. Introduction

With the increasing prevalence of embodied intelligence technology, people have posted a higher requirement on the service capabilities of robots. To better serve daily life and assist humans in managing transactional tasks, the robot needs to incorporate AI models with their inherent mobility and manipulability to cultivate the ability to interact with the environment. For instance, as illustrated in Figure 1, when we ask the robot to retrieve a lemon from the refrigerator, the robot will open the refrigerator and attempt to grasp the target object. Nevertheless, the target object "lemon" may be obstructed or absent, *i.e.*, the object is invisible in the current constrained environment. This situation requires the robot to utilize its current perceptual data and historical knowledge to make exploration decisions and manipulate objects within the constrained environment, thus retrieving the target object.

To achieve better exploration, more and more researchers have attempted to employ large language models (LLMs) in the decision-making process [12, 24]. Based on the perception results obtained from well-trained perception models, such as Faster RCNN [27], LLMs directly generate decision-making sequences to interpret human intentions. Although LLMs are capable of generating the decision-making sequences for object exploration in specific scenarios, these approaches rely solely on the inherent prior knowledge of LLMs. Due to the lack of in-depth analysis of relevant knowledge, such as historical experience and current perception content, the decision-making process of the model may exhibit certain biases [28, 29]. Meanwhile, these decisions are typically accompanied by the robot's manipulation of

© The Author(s), 2026. Published by Cambridge University Press.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 10:24:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0263574726103683

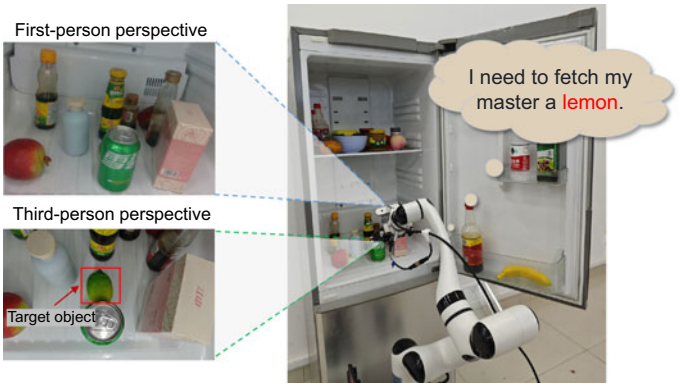


Figure 1. *Overview of the embodied exploration and manipulation task. This task necessitates that the robot should fully utilize its decision-making and manipulation abilities to retrieve the object required by the human.*

various objects, which facilitate the interaction between the robot and the environment. Consequently, the robot can continuously update its historical experience and perception content.

With the advancement of deep learning, current grasp detection techniques have achieved great progress [16, 19, 30]. Existing methods primarily focus on designing model architectures to improve the accuracy of the grasping detection process or reduce the time required for grasping detection. However, they struggle to balance between accuracy and speed. As a result, robots are incapable of grasping target objects in a robust manner. This limitation hinders their practical application in daily scenarios.

In light of the critical considerations outlined above, in this paper, we develop a novel embodied exploration and manipulation framework augmented with knowledge. This framework aims to empower the robot with the ability to explore and grasp the objects required by humans within a constrained environment. It mainly consists of three modules: embodied exploration module, robotic grasping detection module, and grasping execution module. More specifically, the first module enables the robot to actively explore and interact with the complex environments. To achieve this, we introduce a multi-source knowledge guided decision-making module based on the LLMs. On the one hand, the integration of multi-source knowledge can enhance the perception ability of the robot. On the other hand, the inherent decision-making biases of LLMs can be alleviated. Meanwhile, the robot needs to manipulate the objects in the current scene to realize the continuous interaction. Therefore, we integrate a novel robotic grasping detection network with these exploration sequences. Specifically, to improve the accuracy and efficiency of the grasping detection process simultaneously, a two-branch feature fusion module with a residual block is designed. This network is built upon the generative convolutional neural network. Consequently, conditioned on the predicted grasping poses, the robot can transform them into its own coordinate system to achieve the final object manipulation.

The main contributions of this paper are summarized as follows:

- We develop a novel knowledge-augmented embodied exploration and manipulation framework for robots.
- We propose to incorporate the scene-relevant knowledge and object correlation coefficient with the LLMs to generate robust and efficient exploration sequences.
- We propose an efficient and accurate grasping detection network designed to help robots achieve effective operations.
- Extensive experimental results on both simulation and real-world demonstrate the effectiveness and validity of the proposed framework.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 10:24:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0263574726103683

2. Related work

In this section, we categorize and discuss studies closely related to our work into the following groups, *i.e.*, embodied exploration and grasping detection.

2.1. Embodied exploration

Embodied exploration demands that a robot actively engage in environmental exploration and learn to interact with the environment by fully utilizing its own capabilities. This approach has been widely studied within the research domain of robotic navigation [1, 15, 22, 23, 26]. For instance, Ren et al. [23] provide a comprehensive analysis of the critical role of robotic, embodied visual exploration in navigation and other aspects. They propose a standardized framework for classifying and benchmarking current embodied visual exploration algorithms and conduct in-depth evaluations by combining two highly realistic simulated 3D environments.

However, relatively few studies have considered integrating this technology into the robotic operation process [2, 7, 9, 20, 25, 34]. Morrison et al. [20] propose to actively perceive the environment to select informative viewpoints for the subsequent grasping process. Deng et al. [9] introduce a dynamic scene graph to represent the spatial relationships between objects during the embodied exploration process. Meanwhile, they also propose a GRU-based structure to address the sequence-to-sequence task of manipulation planning, and the scene graph updates conditioned on the outcomes. Recently, LLMs have demonstrated significant promise in decision-making for robots [11, 14], showcasing potential advantages in embodied exploration. Despite the fact that these LLMs possess vast knowledge, they still tend to make biased predictions.

2.2. Grasping detection

Grasping detection aims to predict the grasping postures of a given object for subsequent robotic manipulation tasks. With the advancement of deep learning, significant progress has been made in grasping detection [16, 30, 32, 33]. Kumra et al. [16] conceptualize the grasp posture detection process as a generative model and propose a novel generative residual convolutional neural network conditioned on deep convolutional neural modules. Similarly, Morrison et al. [19] introduce a lightweight grasping detection network capable of one-to-one mapping from depth images. Although their approach achieves rapid prediction and can be used in a closed-loop system, it struggles to deliver consistently strong performance. Building upon these approaches, Yu et al. [32] further integrate the UNet and residual attention modules into the grasp detection process.

In contrast, leveraging the advancements in attention mechanisms within Transformer models, Wang et al. [30] propose a transformer-based architecture for robotic grasp detection, featuring two important innovations: window attention and a hierarchical encoder-decoder structure. However, these methods face challenges in balancing the accuracy and efficiency of grasp detection, making them difficult to implement in real-world applications. Recently, there have been several studies on the manipulation of cluttered scenes [4, 17, 31]. However, these studies have focused on improving efficiency in planar manipulation while neglecting manipulation in constrained spaces.

3. Methodology

In this paper, we aim to endow service robots with the ability to actively explore and manipulate within constrained environments, integrating knowledge and LLMs to efficiently tackle a range of tasks required by humans. This task firstly requires the robot to comprehend the human instruction $\mathcal{L}_\pi$ and analyze to identify the target object t_o. Subsequently, based on the environmental perception content $\mathcal{S}$, the robot incorporates the relevant knowledge $\mathcal{K}\pi$ to generate exploration plans $\mathcal{O}\pi$ along with corresponding grasp actions $\mathcal{G}_\pi$. The entire process can be formally represented as $\{\mathcal{O}_\pi, \mathcal{G}_\pi\} = f_\theta(t_o|\mathcal{L}_\pi, \mathcal{K}_\pi, \mathcal{S})$, where f_θ denotes the planner f with parameter θ.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 10:24:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0263574726103683

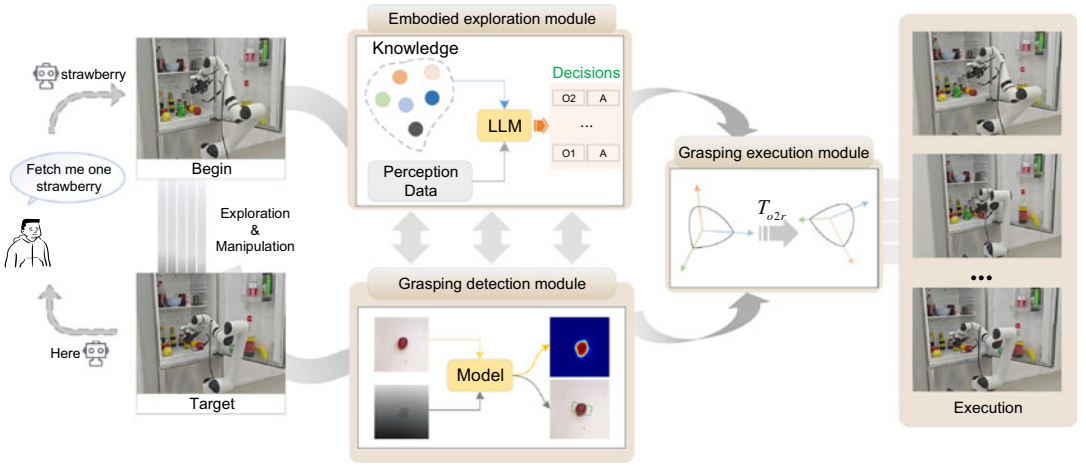


Figure 2. *An overview of the proposed embodied exploration and manipulation framework, mainly consisting of three modules. One module exploits decision-making for the exploration with multi-source knowledge and LLMs. Another module concentrates on improving the accuracy and efficiency of the grasping detection network. The remaining module is responsible for executing the grasping actions.*

An overview of our proposed framework is illustrated in Figure 2, which consists of three main components, *i.e.*, (1) embodied exploration module, which determines the exploration sequences for the objects in the current environment; (2) robotic grasping detection module, which generates the grasping poses for the objects; and (3) grasping execution module, which enables the robot to grasp the objects. The details of each module in our framework will be introduced in the subsequent sections.

3.1. Embodied exploration

In complex operational environments, the sequence in which a robot explores various objects significantly influences the efficiency of task execution. Current decision-making methodologies based on LLMs are constrained by prior training data, which may introduce biases in exploration sequences without comprehensive consideration of the current environmental context. As demonstrated in Figure 2, when instructing a robot to retrieve a strawberry that is not visible in the current robot view, LLM-based approaches may assume its position to be behind a fruit (*e.g.*, a red apple) due to its inherent biases [28, 29], disregarding the discrete spatial distribution characteristic of objects within the refrigerator. Specifically, the strawberry may be obscured by larger objects, not exclusively apples. Furthermore, the dense clustering of apples creates spatial constraints incompatible with strawberry storage requirements. These observations substantiate the weak spatial correlation between strawberries and apples within such configurations. To address these limitations, we propose a multi-source knowledge fusion embodied exploration module designed to mitigate decision-making biases inherent in LLM-based systems. The multi-source knowledge includes domain expert knowledge, historical operation data, and current scene perception information. The architectural implementation is detailed in Figure 3.

Given the current perception RGB image $\mathcal{I} \in \mathcal{S}$, we first utilize one pre-trained Faster RCNN to extract object-level feature F_o as local context for the detected object and global-level feature $F_{\mathcal{I}}$ as global context. Notably, the required object (*i.e.*, strawberry) is not visible. Consequently, we retrieve the features from a pre-built knowledge pool $\{t_o : \{F^k_{t_o}\}, \cdots\}$, comprising object keys t_o and corresponding feature values $F^k_{t_o}$. The knowledge pool is utilized to provide the relevant features extracted from diverse scenes via a pre-trained Faster RCNN for the specific object. Subsequently, specific object-level features can be retrieved from the knowledge pool via a feature fusion component. The process can be formulated as:

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 10:24:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0263574724103683

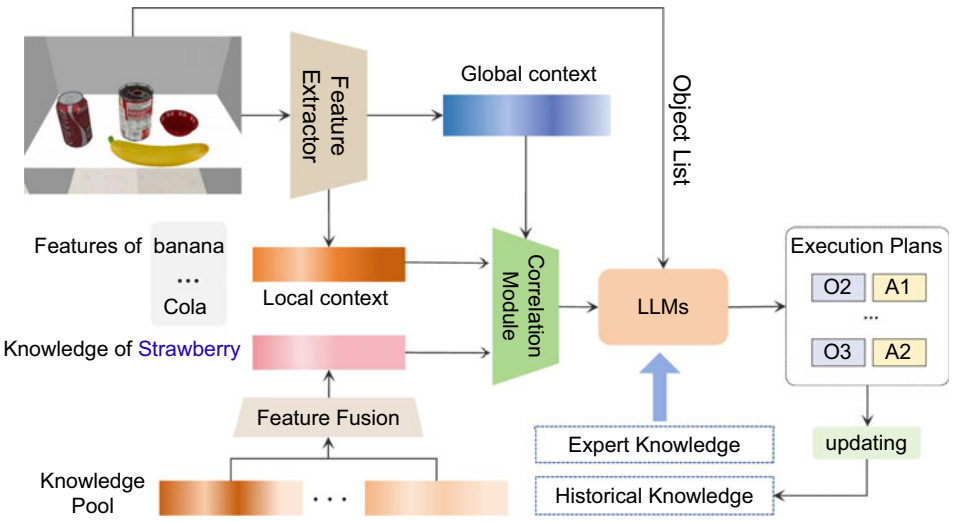


Figure 3. *The architecture of the embodied exploration module. The target object is the strawberry, which is not visible in the current scene.*

$$F^k_{t_o} = \mathrm{MLP}\left(F^k_{t_{o1}}, F^k_{t_{o1}}, \cdots, F^k_{t_{oi}}\right), \quad i \in \mathbb{N} \tag{1}$$

Conditioned on the relevant knowledge features and local and global context, we develop a correlation analysis component to address the issue of bias in the inherent exploration sequence. The correlation module is composed of three fully connected layers. Specifically, five participants were systematically recruited to arrange objects randomly within refrigerators across simulated and real-world environments. These participants were instructed to identify specific target objects for retrieval and their corresponding associated items. Subsequently, perceptual imagery of target objects was integrated with their co-occurring items as training datasets, wherein associated objects served as positive labels $\mathcal{Y} \in \{0, 1\}$ for a binary classification model to evaluate relational proximity among objects. Eq. (1) indicates the target object is behind the object that needs to be manipulated. The training process is formulated as:

$$\min_{\theta} = \mathrm{BCE}(\mathrm{MLP}(\mathrm{MLP}(F_I), \mathrm{MLP}(F_o), F^k_{t_o}), \mathcal{Y}) \tag{2}$$

Upon adequate training, the correlation analysis module can quantitatively assess the degree of association between individual detected objects within the image and the target object. Notably, Eq. (2) is trained on all the objects. Empirically, regions exhibiting higher object density demonstrate increased probability of target occlusion. To leverage this expert knowledge effectively, we employ Intersection over Union (IoU) bounding boxes to assess the local density surrounding each object. Specifically, when the condition $\mathrm{IoU}(t_{o1} \cap t_{o2}) > 0$ is satisfied, the density value associated with t_{o1} is incremented by one.

By integrating the candidate object list, correlation coefficients, and expert knowledge, LLMs are capable of generating unbiased exploration sequences. Additionally, to fully utilize prior exploration knowledge, we integrate historical knowledge into the decision-making process.

3.2. Robotic grasping detection

To improve the effectiveness and efficiency of the grasping detection process, we propose a novel lightweight grasping detection network, as illustrated in Figure 4, consisting of three components, *i.e.*, a visual encoder, two-branch feature fusion, and a feature decoder.

Specifically, following previous methods [16, 18], we employ a similar visual feature extractor composed of 2D-CNN with various sizes of kernels and batch normalization to extract the latent features

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 10:24:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0263574726103683

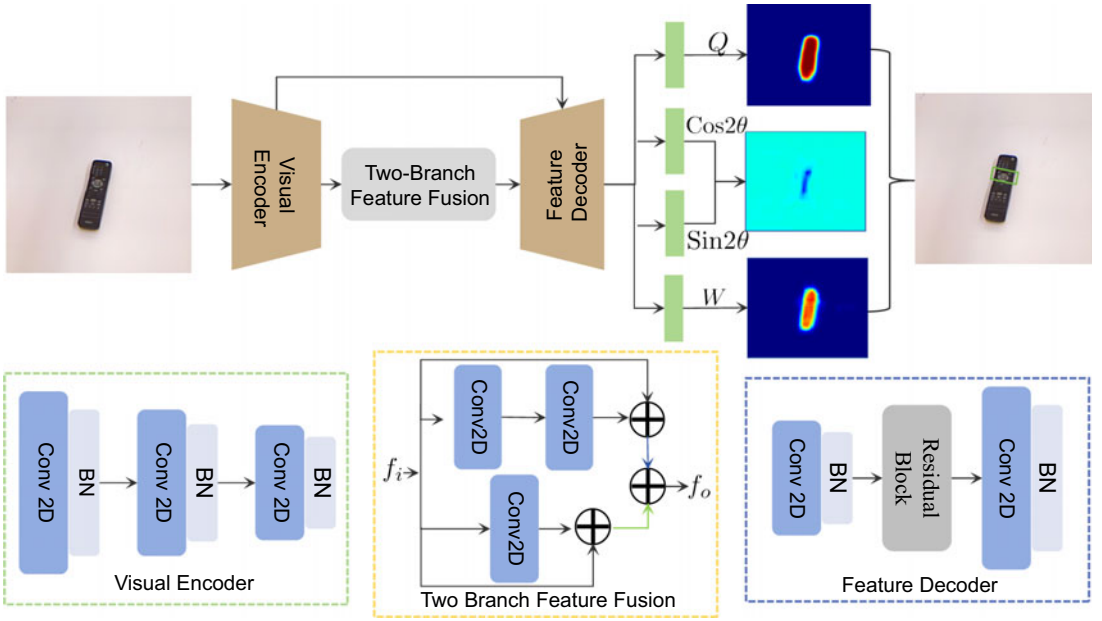


Figure 4. *The architecture of the robotic grasping detection module.*

f_i from the given perception images, including RGB and depth information. To comprehensively capture inherent contextual information from these features, we propose a two-branch feature component. One branch incorporates two sequential 2D-CNN layers to derive fine-grained features f_e from f_i, *i.e.*, $f_e = \text{Conv2D}(\text{Conv2D}(f_i)) + f_i$. The other branch is composed of one 2D-CNN to extract coarse-grained features f_c from f_i, *i.e.*, $f_c = \text{Conv2D}(f_i) + f_i$. Subsequently, these two branches are interconnected through a residual structure, enabling the network to effectively leverage both the newly learned features and the original features. Lastly, the two-branch features are combined to obtain the fused features f_o. The process is formulated as $f_o = f_e + f_c$.

Following the representation update, prior to decoding the fused features, we employ residual blocks to facilitate deep contextual integration by incorporating features from the visual encoder. This approach enables the utilization of more relevant contextual information for the subsequent decoding process. Subsequently, the fused features are fed into the feature decoder, which consists of 2D-CNN and residual blocks, to obtain the pixel-level predictions, including grasping quality Q, grasping angle $\text{Cos}2\theta$ and $\text{Sin}2\theta$, and grasping width W. Additionally, we implement a sigmoid activation function to normalize Q and W values within the bounded range [0,1] to optimize prediction accuracy and computational efficiency.

In order to enable the network to accurately learn grasping detection, we introduce supervised loss constraints to train the proposed network. Specifically, we utilize the binary cross-entropy loss function in the following equation to supervise the training process.

$$\mathcal{L}_{q/w} = -\frac{1}{M}\sum_{m=0}^{M}\left[y^m_{q/w}\cdot\log\left(p^m_{q/w}+\left(1-y^m_{q/w}\right)\right)\cdot\left(1-\log\left(p^m_{q/w}\right)\right)\right], \tag{3}$$

where M denotes the total number of spatial elements (pixels) of the output feature maps. $y^m_{q/w}$ denotes the ground-truth label at position m. $p^m_{q/w}$ denotes the predicted probability at m-th position. The subscripts q and w represent grasping quality and width, respectively. As for the angle, we utilize smooth L1 loss function to supervise the prediction process.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 10:24:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0263574726103683

$$\mathcal{L}_a = \frac{1}{M} \cdot \sum_{m=0}^{M} \begin{cases} 0.5(p^m - y^m)^2, & \text{if } |p^m - y^m| < 1 \\ |p^m - y^m| - 0.5, & \text{otherwise} \end{cases} \tag{4}$$

The whole training loss is formulated as follows:

$$\mathcal{L} = \mathcal{L}_q + \mathcal{L}_w + \mathcal{L}_a \tag{5}$$

3.3. Grasping execution

Based on the grasping quality, grasping angle, and grasping width, the grasping poses for the object can be systematically formulated. Subsequently, these grasping poses will be provided to the robot to execute the grasping actions. The optimal grasp pose $\mathcal{G}_\pi^* = (x^*, y^*, \theta^*, w^*)$ can be calculated by the following:

$$\begin{aligned} (x^*, y^*) &= \arg\max_{(x,y)} (Q), \\ \theta^* &= \frac{1}{2} \arctan \frac{sin(2\Phi)}{cos(2\Phi)}|_{(x^*,y^*)}, \\ w^* &= W|_{(x^*,y^*)}, \end{aligned} \tag{6}$$

Meanwhile, the transform matrices obtained by hand–eye calibration and camera calibration are used to calculate the operation poses g_r in the world frame: Then, the transformation matrices derived from hand–eye calibration and camera calibration are utilized to compute the operational poses for the object $g \in \mathcal{G}_\pi$. The process is computed as follows:

$$g = T_{cr}(T_{ic}(\mathcal{G}_\pi^*)) \tag{7}$$

where T_{ci} and T_{rc} denote transform matrices of the 2D image space to camera space and the camera space to robot space, respectively.

4. Experiments

In this section, we conduct quantitative and qualitative experiments concerning the simulation environment and real-world environment to validate the effectiveness of our proposed method.

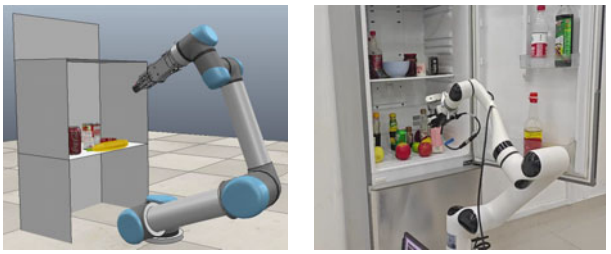
4.1. Experimental settings

The number of objects we utilized in the simulation is set to 15. They are adopted from both YCB [3] and Deng [8]. The target object is varied in each trail. The number of objects in the real-world environment is set to 13. The pre-defined scenes include three types according to the number of objects: mild clutters (≤ 4), moderate clutters($5 \sim 7$), and severe clutters (≥ 8). We select ChatGLM [10] as the basic decision-making LLM. Notably, in this experiment, we assume that there is no stacking of objects, and the objects are arranged randomly by the participants. The simulation and the real-world experimental environment are shown in Figure 5, where the simulation environment is built upon CoppeliaSim.[1] The grasping detection network is trained on Ubuntu 20.04 with Intel(R) Core(TM) i9-10850K CPU and a single RTX 3090.

For the performance of the embodied exploration module, we adopt the Exploration Completion Rate (ECR%) and Completion Steps (CS). ECR represents the success rate of completing the exploration, while CS denotes the optimal number of steps taken to achieve successful completion of the exploration. If the number of explorations exceeds 3, it is considered a failure. As for the robotic grasping detection module, we employ the widely utilized metrics [5, 18] to evaluate the performance of grasping detection, *i.e.*, the angle threshold between the predicted grasp angle and ground truth should be less than 30° and

[1] https://www.coppeliarobotics.com/

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 10:24:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0263574726103683



(a) (b)

Figure 5. *(a) Simulation experiment setup. (b) Real-world experiment setup.*

the Jaccard index should be higher than 25%. As for the whole framework, we adopt the Task Completion Rate (TCR%) to evaluate the performance.

4.2. Evaluation on embodied exploration

To verify the performance of the embodied exploration method, we conducted comparative experiments with a variety of exploration methods in both simulation and real-world environments. Before conducting an in-depth analysis of the results, we first provide a brief introduction to the compared exploration methods.

- Random: Explore objects in a random order.
- Sequential: Explore objects in the field of view in a sequence either from left to right or from right to left.
- Rule-based: Explore based on the size of objects. In the case of objects with the same size, the exploration order is determined randomly.
- Expert-based: Prioritize exploration starting from areas with high density, followed by random exploration.
- LLM-based: Conduct exploration directly based on the decision-making sequences that are generated by LLMs.

The results for the simulation and real-world environment are reported in Tables I and Table II, respectively. As can be observed from the results, we have the following observations: (1) our proposed exploration method demonstrates outstanding performance in both simulated and real-world environments, achieving exploration completion in the minimal number of steps required. (2) As the scenarios gradually become more complex, the success rate of exploration completion declines. Furthermore, the number of steps necessary for completion is higher compared to simpler scenarios. (3) The heuristic-based algorithm exhibits commendable performance in mildly cluttered scenarios. However, as the complexity of the scenarios increases, its performance deteriorates to varying extents. In the simulation environment, when confronted with severely cluttered scenarios, both the random-based and sequential-based methods struggle to achieve successful exploration. (4) In comparison to LLM-based methods, the proposed exploration method demonstrates superior robustness.

4.3. Evaluation on robotic grasping detection

To assess the effectiveness of the grasping detection network proposed in this paper, we conducted a comparison with recent networks using the Cornell dataset [13]. The results reported in Table III demonstrate that the proposed network outperforms others in terms of both accuracy (98.9%) and inference speed (13.2 ms) when utilizing RGB-D modality data. Notably, when compared with algorithms utilizing

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 10:24:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0263574726103683

Table I. *Results of the embodied exploration in the simulation environment.*

	ECR(%)			CS		
Methods	**Mild**	**Moderate**	**Severe**	**Mild**	**Moderate**	**Severe**
Random	40%	20%	0	2	3	4
Sequential	100%	20%	0	2	3	5
Rule-based	60%	20%	40%	2	3	3
Expert-based	60%	40%	20%	2	3	3
LLM-based	60%	100%	40%	3	2	2
Ours	100%	100%	80%	2	1	2

Table II. *Results of the embodied exploration in the real-world environment.*

	ECR(%)			CS		
Methods	**Mild**	**Moderate**	**Severe**	**Mild**	**Moderate**	**Severe**
Random	40%	40%	20%	2	1	3
Sequential	80%	20%	20%	1	3	3
Rule-based	60%	20%	20%	3	3	3
Expert-based	60%	60%	20%	2	3	3
LLM-based	60%	100%	60%	3	2	3
Ours	100%	80%	80%	2	2	2

Table III. *Performance of grasping detection network on Cornell dataset.*

Methods	**Size**	**Input modality**	**Accuracy**	**Inference time**
GR-ConvNet [16]	224 × 224	RGB	94.4	20.0 ms
		RGB-D	95.5	21.0 ms
TF-Grasp [30]	640 × 480	RGB	95.0	41.3 ms
		RGB-D	96.7	41.6 ms
SE-ResUNet [32]	224 × 224	RGB-D	94.3	30.5 ms
		RGB	92.1	29.0 ms
GGCNN [19]	300 × 300	D	76.4	6.3 ms
FANet [33]	300 × 300	RGB-D	97.8	24 ms
VERGNet [21]	224 × 224	RGB-D	97.7	53 ms
Cheng et al. [6]	224 × 224	RGB	96.2	9 ms
Ours	224 × 224	RGB	96.6	12 ms
		RGB-D	98.9	13.2 ms

identical input dimensions (224 × 224 pixels), including GR-ConvNet, SE-ResUNet, and VERGNet, the proposed network maintains significant advantages in both metrics. Furthermore, in scenarios where only RGB data is employed, the proposed network attains an accuracy of 96.6%. In comparison to the algorithm proposed by Cheng et al. [6], the proposed algorithm exhibits a significant advantage in accuracy. Although the inference time is marginally longer, it still satisfies the requirements for practical applications.

To further verify the rationale of using the scene image to predict grasping postures with our proposed network, we present a comparative analysis of grasping detection performance with GR-ConvNet, GGCNN, and SE-ResUNet in Figure 6. The qualitative results are alongside the grasping quality. As can be observed from the results, our grasping detection network generates more precise grasping poses

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 10:24:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0263574726103683

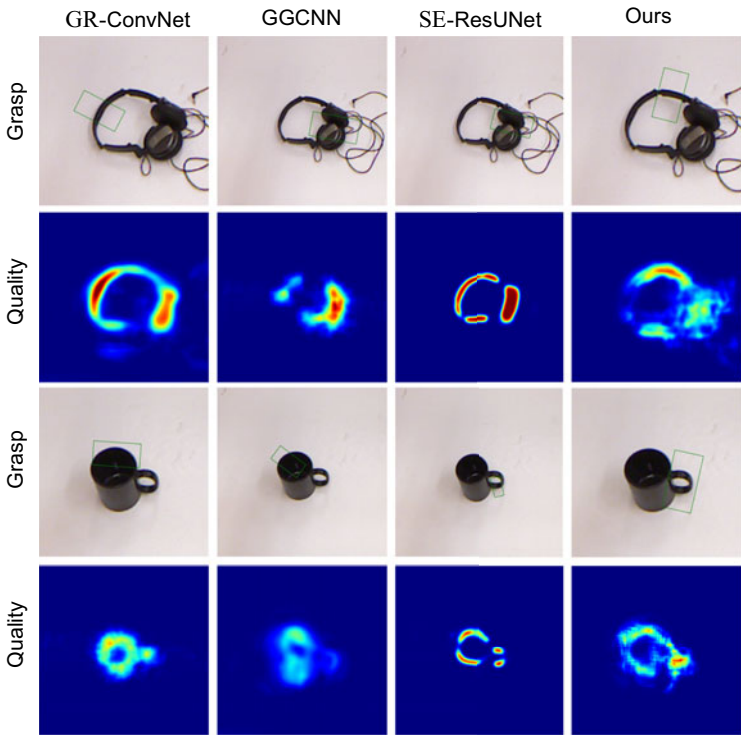


Figure 6. *Comparison of qualitative results on single objects.*

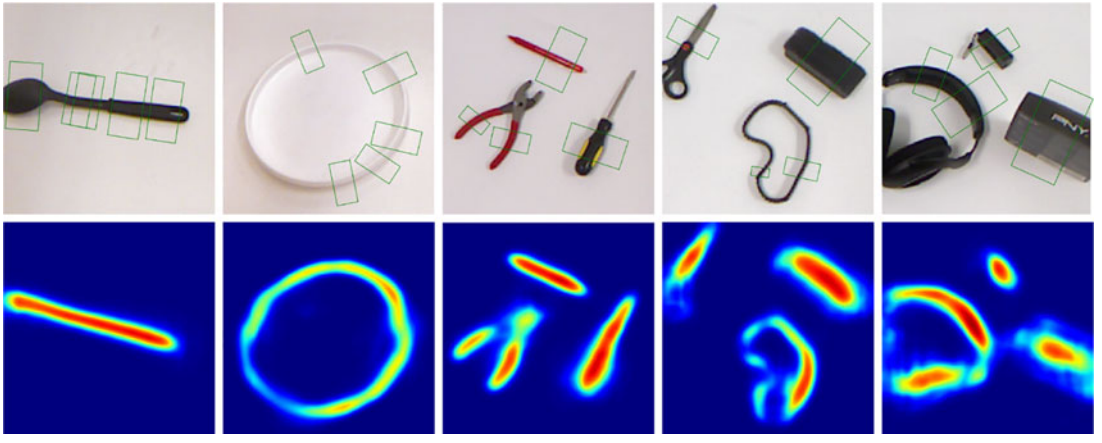


Figure 7. *Qualitative results on multiple objects.*

based on reliable grasping quality and outperforms existing networks. To further verify the rationale of using the scene image to predict grasping postures with our proposed network, we illustrate the grasping results of known objects along with their grasping quality and compare them with those of GR-ConvNet, GGCNN, and SE-ResUNet in Figure 6. As can be observed from the results, compared with other networks, the grasping detection network proposed by us can generate more accurate grasping poses based on reliable grasping quality.

Although the proposed network is exclusively trained on a single-object grasping detection dataset, it is capable of handling multi-object environments by selecting the objects with high grasping quality. As illustrated in Figure 7, comprehensive qualitative evaluations were conducted using a multi-object

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 10:24:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0263574726103683

Table IV. *Performance of the whole framework. EX denotes the embodied exploration.*

Exploration methods	Grasping detection methods	TCR (%)
EX-based	GR-ConvNet [16]	90%(45/50)
EX-based	TF-Grasp [30]	90%(45/50)
EX-based	SE-ResUNet [32]	96%(48/50)
EX-based	Ours	96%(48/50)
LLM-based	Ours	92%(46/50)
Rule-based	Ours	56%(28/50)

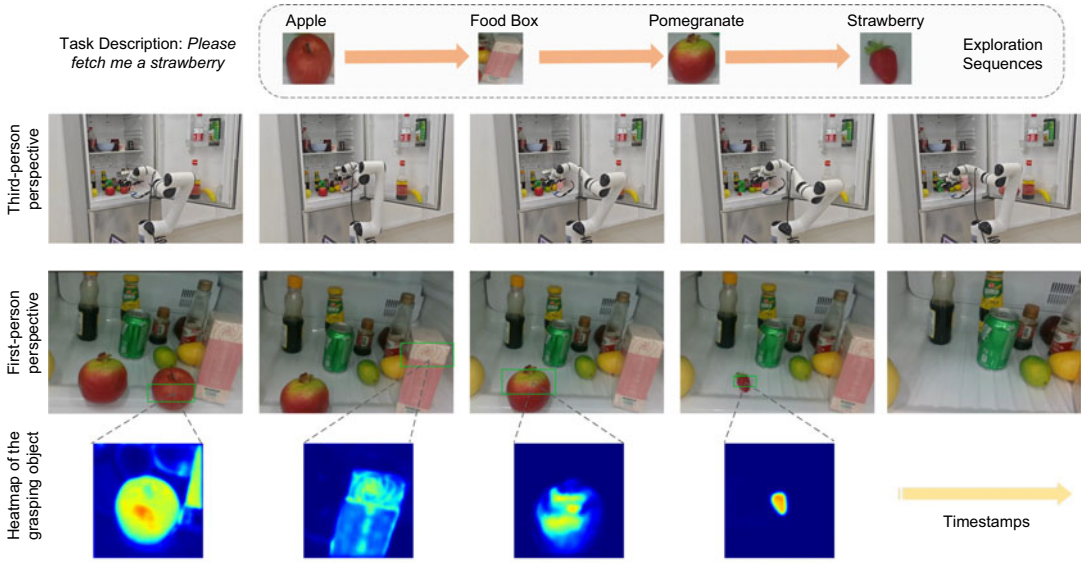


Figure 8. *Screenshots of robotic exploration and manipulation experiments within a refrigerator environment.*

grasping detection dataset. The generated grasping poses and corresponding quality heatmaps reveal that the proposed network can generate precise grasping postures based on calculated quality scores in complex multi-object scenarios.

4.4. Real-world cases

To demonstrate the grasping performance of our proposed knowledge-augmented embodied exploration and manipulation framework in practice, we choose to conduct an overall test of the framework in a moderately cluttered real-world environment. We perform a total of 50 attempts for the whole framework. The results reported in Table IV are evaluated by the Task Completion Rate (TCR%). The experimental results demonstrate that the framework incorporating the embodied exploration algorithm proposed in this study achieved 48 successful grasping manipulations with a 96% success rate. Compared to other methods, our approach demonstrates superior performance. Furthermore, when the grasp detection network introduced in this paper is used to verify the overall success rate across different exploration methods, it becomes evident that utilizing LLMs significantly enhances task completion success rates. Nevertheless, due to the existence of its inherent decision-making bias, its performance remains lower than that of the embodied exploration algorithm proposed in this paper.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 10:24:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0263574726103683

To further reveal the effectiveness of our framework, we present the qualitative results for the task execution process in Figure 8. The robot first comprehends the human instruction and utilizes the embodied exploration module to generate the exploration sequences. During the exploration process, the robot also needs to manipulate the object conditioned on the robotic grasping detection and execution module, realizing the environment interaction. These results further demonstrate that our proposed framework enables robots to actively explore and manipulate within refrigerator environments.

5. Conclusion

In this paper, we propose a knowledge-augmented robotic embodied exploration and manipulation framework. The framework aims to endow service robots with the ability to actively explore and manipulate within constrained environments, integrating knowledge and LLMs to efficiently tackle a range of tasks required by humans. To reveal the effectiveness of our framework, we conducted extensive experiments in both simulation and real-world environments across different scenario settings, demonstrating the effectiveness and efficiency of our proposed framework. In this research, we employed a 2D grasping detection network to provide prediction results for robotic manipulation. However, this approach has certain limitations, causing collisions in the exploration postures at some positions. In future research, we plan to incorporate 6D pose estimation into the network process to achieve more precise grasping.

Author contributions. LJ conducted the research and completed the manuscript. KS and LBX conducted the real-world experiments. CQW provided guidance and reviewed the manuscript.

Financial support. This work is supported by the National Natural Science Foundation of the China under Grant 62473370 and China Postdoctoral Science Foundation under Grant Number 2025M781675.

Competing interests. The authors declare that no conflicts of interest exist.

Ethical approval. Not applicable.

Use of AI. Since our research focuses on LLM knowledge, in our paper, we incorporate ChatGLM with our designed modules as the decision tool for the robot.

References

[1] R. Bigazzi, M. Cornia, S. Cascianelli, L. Baraldi and R. Cucchiara, "Embodied Agents for Efficient Exploration and Smart Scene Description," **In:** *2023 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2023) pp. 6057–6064.

[2] M. Breyer, L. Ott, R. Siegwart and J. J. Chung, "Closed-Loop Next-Best-View Planning for Target-Driven Grasping," **In:** *2022 IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS) 2022* (IEEE, 2022) pp. 1411–1416.

[3] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel and A. M. Dollar, "The YCB Object and Model Set: Towards Common Benchmarks for Manipulation Research," **In:** *2015 International Conference on Advanced Robotics (ICAR)* (IEEE, 2015) pp. 510–517.

[4] Z. Chen, X. Zhang and D. Tu, "Research on robotic grasp detection using improved generative convolution neural network with Gaussian representation," *Robotica* **43**(12), 4191–4211 (2025).

[5] H. Cheng, Y. Wang and M. Q.-H. Meng, "A robot grasping system with single-stage anchor-free deep grasp detector," *IEEE Trans. Instrum. Meas.* **71**(2), 1–12 (2022).

[6] H. Cheng, Y. Wang and M. Q.-H. Meng, "Anchor-based multi-scale deep grasp pose detector with encoded angle regression," *IEEE Trans. Autom. Sci. Eng.* **21**(3), 3130–3142 (2023).

[7] C. De Farias, N. Marturi, R. Stolkin and Y. Bekiroglu, "Simultaneous tactile exploration and grasp refinement for unknown objects," *IEEE Robot. Autom. Lett.* **6**(2), 3349–3356 (2021).

[8] Y. Deng, D. Guo, X. Guo, N. Zhang, H. Liu and F. Sun, "MQA: Answering the question via robotic manipulation," arXiv preprint arXiv:2003.04641. **1**(1), 1–10 (2020).

[9] Y. Deng, Q. Sima and H. Liu, "Scene graph for active exploration in cluttered scenario," academia.edu. **1**(1), 1–9 (2022).

[10] Team GLM, A. Zeng and B. Xu, "ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools," arXiv preprint arXiv:2406.12793. **1**(1), 1–19 (2024).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 10:24:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0263574726103683

[11] Y. Hu, F. Lin, T. Zhang, L. Yi and Y. Gao, "Look before you leap: Unveiling the power of GPT-4V in robotic vision-language planning," arXiv preprint arXiv:2311.17842. **1**(1), 1–20 (2023).
[12] Z. Hu, F. Lucchetti, C. Schlesinger, Y. Saxena, A. Freeman, S. Modak, A. Guha and J. Biswas, "Deploying and evaluating LLMs to program service mobile robots," *IEEE Robot. Autom. Lett.* **9**(3), 2853–2860 (2024).
[13] Y. Jiang, S. Moseson and A. Saxena, "Efficient Grasping From RGBD Images: Learning Using a New Rectangle Representation," **In:** *2011 IEEE International Conference on Robotics and Automation 2011* (IEEE, 2011) pp. 3304–3311.
[14] Y. Jin, D. Li, A. Yong, J. Shi, P. Hao, F. Sun, J. Zhang and B. Fang, "RobotGPT: Robot manipulation learning from ChatGPT," *IEEE Robot. Autom. Lett.* **9**(3), 2543–2550 (2024).
[15] Y. Jing and T. Kong, "Learning to Explore Informative Trajectories and Samples for Embodied Perception," **In:** *2023 IEEE International Conference on Robotics and Automation (ICRA) 2023* (IEEE, 2023) pp. 6050–6056.
[16] S. Kumra, S. Joshi and F. Sahin, "Antipodal Robotic Grasping Using Generative Residual Convolutional Neural Network," **In:** *2020 IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS) 2020* (IEEE, 2020) pp. 9626–9633.
[17] O. Kundu, S. Dutta and S. Kumar, "A novel method for finding grasping handles in a clutter using RGBD Gaussian mixture models," *Robotica* **40**(3), 447–463 (2022).
[18] J. Liu, J. Xie, S. Huang, C. Wang and F. Zhou, "Continual learning for robotic grasping detection with knowledge transferring," *IEEE Trans. Ind. Electron.* **7**(4), 11019–11027 (2023).
[19] D. Morrison, P. Corke and J. Leitner, "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach," *Robotics: Science and Systems* **1**(1), 1–10 (2018).
[20] D. Morrison, P. Corke and J. Leitner, "Multi-View Picking: Next-Best-View Reaching for Improved Grasping in Clutter," **In:** *2019 International Conference on Robotics and Automation (ICRA) 2019* (IEEE, 2019) pp. 8762–8768.
[21] M. Niu, Z. Lu, L. Chen, J. Yang and C. Yang, "VERGNet: Visual enhancement guided robotic grasp detection under low-light condition," *IEEE Robot. Autom. Lett.* **12**(8), 8541–8548 (2023).
[22] S. K. Ramakrishnan, D. Jayaraman and K. Grauman, "An exploration of embodied visual exploration," *Int. J. Comput. Vis.* **129**(5), 1616–1649 (2021).
[23] A. Z. Ren, J. Clark, A. Dixit, M. Itkina, A. Majumdar and D. Sadigh, "Explore Until Confident: Efficient Exploration for Embodied Question Answering," **In:** *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024* (IEEE, 2024) pp. 1–5.
[24] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(6), 1137–1149 (2016).
[25] N. Saito, T. Ogata, S. Funabashi, H. Mori and S. Sugano, "How to select and use tools?: Active perception of target objects using multimodal deep learning," *IEEE Robot. Autom. Lett.* **6**(2), 2517–2524 (2021).
[26] Q. Sima, S. Tan, H. Liu, F. Sun, W. Xu and L. Fu, "Embodied Referring Expression for Manipulation Question Answering in Interactive Environment," **In:** *2023 IEEE International Conference on Robotics and Automation (ICRA) 2023* (IEEE, 2023) pp. 7635–7641.
[27] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason and A. Garg, "ProgPrompt: Generating Situated Robot Task Plans using Large Language Models," **In:** *2023 IEEE International Conference on Robotics and Automation (ICRA) 2023* (IEEE, 2023) pp. 11523–11530.
[28] Z. Sun, L. Du, X. Ding, Y. Ma, Y. Zhao, K. Qiu, T. Liu and B. Qin, "Causal-Guided Active Learning for Debiasing Large Language Models," **In:** *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 2024* (ACL, 2024) pp. 14455–14469.
[29] W. Tan, W. Zhang, S. Liu, L. Zheng, X. Wang and B. An, "True knowledge comes from practice: Aligning LLMs with embodied environments via reinforcement learning," arXiv preprint arXiv:2401.14151. **1**(1), 1–48 (2024).
[30] S. Wang, Z. Zhou and Z. Kan, "When transformer meets robotic grasping: Exploits context for efficient grasp detection," *IEEE Robot. Autom. Lett.* **7**(3), 8170–8177 (2022).
[31] H. Xi, S. Li and X. Liu, "A pixel-level grasp detection method based on Efficient Grasp Aware Network," *Robotica* **42**(9), 3190–3210 (2024).
[32] S. Yu, D.-H. Zhai, Y. Xia, H. Wu and J. Liao, "SE-ResUNet: A Novel robotic grasp detection method," *IEEE Robot. Autom. Lett.* **7**(2), 5238–5245 (2022).
[33] D.-H. Zhai, S. Yu and Y. Xia, "FANet: Fast and accurate robotic grasp detection based on keypoints," *IEEE Trans. Autom. Sci. Eng.* **21**(3), 2974–2986 (2023).
[34] 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:** *2023 Conference on Robot Learning 2023* (PMLR, 2023) pp. 2849–2862.

Cite this article: J. Liu, K. Sun, L. Xiao and C. Wang (2026). "Knowledge-augmented embodied exploration for robotic grasping in constrained environments", Robotica **44**, 1830–1842. https://doi.org/10.1017/S0263574726103683

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 12 Nov 2026 at 10:24:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0263574726103683